From Technological Disruption to Adaptation: Optimal Policies for Taxation and Retraining

Xilin Yuan *

May 1, 2026

Abstract

Technological progress raises aggregate productivity but generates uneven gains across occupations, making redistribution and occupational transitions central policy concerns. This paper studies the joint optimal design of income taxation and retraining subsidies in response to technological disruptions in a two-sector Mirrlees framework with risky retraining outcomes and unobserved productivity. A key determinant of optimal policy is productivity–training complementarity, which shapes both the efficiency gains from and the incentive costs of retraining. I estimate the retraining technology using German administrative data (SIAB), documenting strong complementarity at the margin from no training to short-term training. Quantitatively, under an automation shock, the constrained optimum prescribes universal participation in short-term training across all manual productivity types. In contrast, laissez-faire underinvests in training due to uninsured outcome risk, and the prevailing policy regime can misallocate workers across types and distort training choices due to misaligned incentives. In the calibrated economy, training expenditures amount to 0.6 percent of GDP, and moving from the actual regime to the optimal joint policy yields welfare gains equivalent to 0.38 percent of GDP.

*Department of Economics, Yale University. Contact: xilin.yuan@yale.edu. I thank Giuseppe Moscarini, Zhen Huo, Joel Flynn, Ilse Lindenlaub, and other participants at Yale Macro Lunch for their valuable guidance and helpful suggestions. All errors are my own.

1 Introduction

Technological progress fosters aggregate economic growth, but it rarely does so evenly. New technologies raise productivity and expand demand for some tasks while eroding the value of others, leading to declining wages, displacement, or even obsolescence for the latter. In principle, disrupted workers can move toward thriving occupations, yet occupational transitions are costly: acquiring new skills takes time and money, and success is uncertain. Traditional policy responses, such as unemployment benefits and income tax adjustments, provide immediate financial relief. But they often do little to improve long-term labor market outcomes. Subsidized retraining programs are attracting renewed attention as an alternative policy instrument. They are widely used in developed economies like Germany and the United States to help disrupted workers adapt to shifting economic conditions.[1]

However, retraining itself is not a panacea. Concerns persist about its actual impact.[2] Participants may lack incentives to succeed for two reasons. First, generous cash transfers may weaken the private return to successful adaptation. Second, retraining is a risky human capital investment with uncertain returns, and insurance against poor training outcomes is virtually nonexistent. In addition, even when programs improve labor-market outcomes, their fiscal costs can be large.

This paper studies the joint optimal design of income taxation (including transfers) and subsidized retraining in response to technological shocks. The central premise is that the two instruments are intrinsically linked. Redistribution policies change the private return to retraining—by altering disposable-income gains from moving to higher-paying occupations—and therefore shape participation incentives. Conversely, retraining changes the future distribution of incomes and tax bases, and thus affects the fiscal cost of providing insurance. Overly generous transfers may discourage socially desirable retraining, while poorly targeted training subsidies may induce socially costly investments.

[1] For example, see Hyman (2018) who studies Trade Adjustment Assistance for Workers (TAA) which provides retraining opportunities to displaced workers hit by trade shocks, and Schulze (2024) who studies the effect of training programs in the context of the Great Recession.

[2] Public skepticism toward subsidized retraining programs never seems to end. See reports from New York Times:
https://www.nytimes.com/2017/02/23/magazine/retraining-jobs-unemployment.html
and from The Atlantic:
https://www.theatlantic.com/education/archive/2018/01/the-false-promises-of-worker-retraining/549398/

I formalize these interactions in a two-sector Mirrlees framework with risky retraining outcomes. Workers in the disrupted sector privately observe their incumbent productivity and can choose a training intensity that stochastically shifts their productivity in the undisrupted sector. Because post-training outcomes are risky and cannot be perfectly insured in decentralized markets, laissez-faire may feature inefficiently low training participation. At the same time, because productivity is private information, an optimal policy must manage information rents: training that disproportionately benefits high-productivity workers can improve efficiency but also raises the incentive cost of truthful revelation. A key determinant of how the planner should sort workers with different levels of productivity into different levels of training intensity is productivity–training complementarity, which governs how the marginal return to training varies with a worker's productivity in the disrupted sector. High complementarity creates an efficiency force toward positive sorting: the planner wants to assign more training to high-productivity workers because they gain more from it and raise total output. But the same complementarity also raises their information rents. From an incentive perspective, this pushes the planner in the opposite direction: higher complementarity can lead the planner to reduce training subsidies for high-productivity workers.

To discipline the model, I estimate the retraining technology using the Sample of Integrated Labour Market Biographies (SIAB), an administrative dataset in Germany, which records workers' labor-market histories and training participation. I first use a matching event study to show that retraining programs in Germany, including short-term and long-term training, indeed improve post-displacement labor market outcomes. I then estimate the retraining technology in the structural model, taking the German policy regime in the sample period as given. The estimates show that higher-ability manual workers benefit more from training than their lower-ability counterparts for short-term training programs, indicating productivity–training complementarity there.

Feeding the estimated technology into the model, I quantify optimal policy responses to a large automation shock, modeled as a 40 percent wage shock to the manual sector. The constrained optimum, which can be decentralized with a history-independent tax-and-transfer system combined with history-dependent training subsidies, prescribes universal participation in short-term training for all manual-origin workers. But it limits the use of long-term training programs, primarily due to their high costs. In contrast, the laissez-faire economy underinvests in training because workers face uninsured risk, and the actual policy regime in Germany during the period of study can misallocate resources by over-insuring low-productivity workers—reducing their incentive to retrain—while implicitly subsidizing socially costly long-term training for high-productivity workers.

The constrained optimum allocation features substantial sectoral redistribution and higher retraining spending compared to the actual policy after the 40 percent wage shock to the manual sector. Total redistribution from the non-manual sector to the manual sector under optimal policy accounts for 2.24 percent of GDP. By comparison, without the shock, the manual sector would redistribute 6.33 percent of GDP to the non-manual sector, because the manual workers have higher productivity on average. In terms of training expenditure, 0.6 percent of GDP is spent on training programs. To put this into perspective, Germany spent 0.22 percent of its GDP on training programs in 2000. Moving from the actual regime to the optimal one yields a welfare gain equivalent to 0.38% of GDP.

Literature This paper relates to several strands of the literature. First, there is a large body of research that discusses policy designs that support disrupted workers. Most of them focus on production-side taxation (Guerreiro et al., 2022, Thuemmel, 2023, Costinot and Werning, 2023, Lehr and Restrepo, 2022). By taxing productive factors like robots to discourage their use, labor demand for disrupted workers will partly be restored, and the tax revenue can be transferred to them. This paper complements that approach by studying policies that operate directly on workers: redistribution, human-capital accumulation, and occupational switching.

Second, this paper also relates to the Mirrlees optimal taxation literature (see Stantcheva (2020)). Retraining in my model is one type of risky human capital accumulation. A few papers study optimal policies that encourage risky human capital accumulation in the Mirrlees framework (Stantcheva, 2015, 2017, Findeisen and Sachs, 2016). Their skills are one-dimensional, and there is no notion of occupation. Equivalently, their productivity is perfectly transferable across different jobs. I follow Findeisen and Sachs (2016) to model retraining as a one-shot risky human capital investment, but I introduce two-dimensional skills. Two-dimensional skills make it possible to study occupational switching, an important margin of worker reallocation in the labor market. Some papers (Rothschild and Scheuer, 2013, Gomes et al., 2018) study optimal taxation in a Roy model, but their productivity is fixed.

Introducing two-dimensional productivity not only allows me to talk about occupational reallocation, but also bears some implications for the planner's allocation. If there is only one-dimensional productivity, and one models the sectoral shock as productivity depreciation for a subset of workers, then supposedly, the planner would like them to engage in the same level of training. With two-dimensional skills, the optimal amount of retraining depends on a worker's current productivity and, crucially, on the training technology that maps current skills into productivity in the other sector. If training is particularly beneficial for low-skilled workers, the planner would train them more and ask them to relocate. Conversely, if training is highly complementary to their current productivity, enabling high-skilled workers in the disrupted sector to effectively acquire the productivity relevant to the other sector, the planner would encourage high-skilled workers to train more and relocate. Simultaneously, greater monetary support would be provided to low-skilled workers who remain in the disrupted sector.

Third, this paper complements a large empirical literature that evaluates training programs and other active labor market policies (Crépon and Van Den Berg, 2016, Lechner et al., 2011, Osikominu, 2013). I study the normative aspect of retraining programs in an equilibrium model, highlighting its role in facilitating redistribution and promoting occupational switching, as well as risk-sharing for different training outcomes. I show what the optimal training policy should look like under idiosyncratic unemployment shocks, automation shocks, and AI shocks.

Roadmap This paper is organized as follows. I introduce the model and the planner's problem in Section 2. Then, I study the optimal allocation in the first-best benchmark and under information frictions in Section 3. Next, in Section 4, I introduce the data, provide suggestive empirical evidence of training effectiveness, and estimate the training technology. Section 5 presents the quantitative exercise. Section 6 concludes.

2 Model

2.1 Environment and the Decentralized Problem

Environment There are two occupations, M and N, in the economy.[3] Occupation $K \in \{M, N\}$ has measure $\mathbb{M}_K$ of workers, where $\mathbb{M}_M + \mathbb{M}_N = 1$. Workers in each occupation have a sector-specific productivity denoted by lowercase m and n, respectively.[4] In sector K, productivity k has a cumulative distribution function (CDF) $\tilde{F}_k(k)$, with density $\tilde{f}_k(k)$ and support $[\underline{k}, \bar{k}]$. It is possible that $\bar{k} = \infty$. Here, $\tilde{F}_k$ is unnormalized so that $\tilde{F}_k(\bar{k}) = \mathbb{M}_K$, and $\tilde{F}_k$ is exogenously given.[5] With l units of physical labor supply, a worker's efficient labor supply is $kl, k = m, n$, if he is in sector K, and occupation K offers a constant wage rate w_k per unit of efficient labor supply.

There are two stages: the training stage and the working stage. In the training stage, workers in occupation M have access to a training program and can choose their retraining level $z \geq 0$, which incurs a money cost z and disutility $\phi(z)$, with $\phi'' > 0$.[6] $z = 0$ means the worker does not get training. At the end of the training stage, a worker with productivity m and training z draws his new productivity n from a training technology characterized by a distribution with CDF $G(n|z, m)$.[7] When the worker chooses the training level, he perfectly observes his productivity m but only knows the distribution of the latent productivity n. For simplicity, assume workers in occupation N have no access to this training technology.

In the second stage, after observing their new productivity draw, workers in sector M can choose whether to switch to sector N or not. Finally, workers in both sectors choose their labor supply l and consumption c which yields utility $u(c) - \psi(l)$.[8]

[3] I use occupation and sector interchangeably in this paper. One can think of M as standing for the manual sector, and N as standing for the non-manual sector.

[4] This sector-specific productivity may include both sector-specific skills and general human capital. I do not take a stance on the nature of it.

[5] That means we do not model the initial sorting into the two sectors but take worker measure and productivity distribution as given.

[6] Assuming constant marginal monetary cost is only a normalization.

[7] Throughout the paper, I refer to $G(n|z, m)$ as the training technology. This is a slight abuse of terminology, because the object also includes $z = 0$. In that case,the training technology refers to the latent joint distribution of the two productivities among M workers.

[8] I use separable utility here only for expositional simplicity.

The decentralized problem Workers have no access to financial markets, so they consume what they earn net of the training cost.[9] For a worker in sector M with productivity m, denote by $v^K(k,z)$ the ex post (i.e., after occupational switch decision) indirect utility function if the worker is in sector K and received training z, $(K,k) \in \{(M,m),(N,n)\}$. Denote by $v(n,m,z)$ the interim utility after training z but before occupation is chosen. The expected value from training z at stage 1 is therefore $\mathbb{E}^n[v(n,m,z)]$ net of training disutility. Hence, we can formulate the worker's problem as

$$\max_z -\phi(z) + \int_{\underline{n}}^{\bar{n}} v(n,m,z) g(n|m,z) dn$$
$$s.t. v(n,m,z) = \max\left\{v^N(n,z), v^M(m,z)\right\}$$
$$v^K(k,z) = \max_y u(y-z) - \psi\left(\frac{y}{w_k k}\right), (K,k) \in \{(M,m),(N,n)\}$$

An incumbent worker in sector N with productivity n only chooses his labor supply in the second stage. His problem is simply

$$v(n) = \max_y u(y) - \psi\left(\frac{y}{w_n n}\right)$$

2.2 The Planning Problem

This section describes the problem of a benevolent social planner. Wage rates in the two sectors and the training technology are public information, and the planner can observe and hence, contract on, a worker's training level,[10] occupation, total income, and consumption. However, the planner cannot observe a worker's productivity level nor labor supply. This is a standard assumption in the optimal taxation literature (Mirrlees, 1971). Given this information friction, I use a direct mechanism design approach to characterize the constrained optimal allocation and then discuss the implementation of it. In particular, in the first stage, the planner asks workers in occupation M to report their productivity m. Given their report, the planner chooses their training level. Then, the planner asks them to report their new productivity draw n. Based on their report,

[9] This amounts to saying that workers cannot insure against their training outcomes.

[10] The training level here refers to the prescribed training level. Though I do not model unobserved training effort here, one point this paper tries to make is that if the income tax schedule is properly designed, taking the prescribed training (and therefore exerting effort) is incentive-compatible for a worker.

the planner assigns them to one of the two sectors. Finally, along with the report of productivity n from workers in sector N, the planner decides labor supply and consumption level for each worker. The problem is described backwards below.

Second-stage IC condition for incumbent sector N workers The problem for sector N workers is fairly standard. Suppose a worker in sector N with productivity n makes a report r_n for his productivity. To induce truthful reporting, the recommended consumption $c^N(r_n)$ and income $y^N(r_n)$ must satisfy

$$U^N(n) := \tilde{U}^N(n;n) = \max_{r_n \in [\underline{n},\bar{n}]} u(c^N(r_n)) - \psi\left(\frac{y^N(r_n)}{w_n n}\right) \tag{1}$$

where $\tilde{U}^N(r_n;n)$ is the utility of a worker with productivity n and reports r_n, and $U^N(n)$ denotes the utility from truthful reporting.

Second-stage IC condition for sector M workers At the beginning of the second stage, workers from sector M have made a report r_m of their productivity m in the first stage. Let $c^M(r_n, r_m)$ and $y^M(r_n, r_m)$ denote the recommended consumption and income for a M-origin worker who makes report (r_n, r_m) and stays in M, and $c^{MN}(r_n, r_m)$ and $y^{MN}(r_n, r_m)$ if he switches to sector N. Let $S(r_m)$ be the switching set such that if $r_n \in S(r_m)$, the planner will recommend the worker to switch sector. Otherwise, the worker stays. With the above notation, the second-stage IC condition reads

$$\begin{aligned} v(r_m, n, m) := \tilde{v}(n; r_m, n, m) = \max_{r_n \in [\underline{n},\bar{n}]} &\left[u(c^M(r_n, r_m)) - \psi\left(\frac{y^M(r_n, r_m)}{w_m m}\right)\right] \mathbb{I}(r_n \notin S(r_m)) \\ &+ \left[u(c^{MN}(r_n, r_m)) - \psi\left(\frac{y^{MN}(r_n, r_m)}{w_n n}\right)\right] \mathbb{I}(r_n \in S(r_m)) \end{aligned} \tag{2}$$

where $\tilde{v}(r_n; r_m, n, m)$ denotes the second-stage value if the worker reports r_n when he has true productivity bundle (n, m) and reported r_m in the first stage, $v(r_m, n, m)$ denotes the value of truthful reporting in the second stage, and $\mathbb{I}(.)$ is an indicator function.

First-stage IC condition for sector M workers Suppose a worker in sector M with true productivity m makes a report r_m in the first stage. After this report, the planner recommends a training level $z(r_m)$ and promises a value $v^p(r_m)$ to the worker. Then, truthful reporting requires

$$U^M(m) := \tilde{U}^M(m;m) = \max_{r_m \in [\underline{m}, \bar{m}]} -\phi(z(r_m)) + v^p(r_m) \tag{3}$$

where $\tilde{U}^M(r_m;m)$ is the expected value of a worker from sector M with productivity m and reports r_m in the first stage, $U^M(m)$ is the expected value of truthful reporting in the first stage, and $v^p(r_m)$ is the continuation value in the second stage given by

$$v^p(r_m) = \int_n v(r_m, n, m) dG(n|z(r_m), m) \tag{4}$$

The planner's problem Assume the planner assigns Pareto weights $F_i(i), i = m, n$ to workers in order to aggregate social welfare. The weights sum to unity with density $f_i(i)$. To ease notation, I sometimes suppress the arguments in some functions as long as there should be no confusion. Then, the planner's problem can be written as

$$\max_{c^N, c^{MN}, c^N, y^N, y^{MN}, y^M, z, v^p, S} \int_n U^N(n) dF_n(n) + \int_m U^M(m) dF_m(m)$$

$$s.t. \quad E \geq \text{training expenditure} + \text{net subsidy to } M\text{-stayers}$$

$$+ \, MN \text{ switchers} + \text{incumbent } N \text{ workers}$$

$$\text{The allocation satisfies IC conditions in (1)-(4)}$$

The first constraint is the planner's resource constraint, where E is some unmodeled exogenous government endowment. For brevity, I describe each term in words. The complete expression can be found in Appendix B.1.

Define Wedges Given the complicated problem, to aid the analysis, I follow the optimal taxation literature by characterizing the optimal wedges that the planner would like to implement (Stantcheva, 2017). They capture the optimal distortion that the planner wants to impose on the decentralized economy so that the privately optimal choices align with the socially optimal ones. In particular, I define two sets of wedges: labor wedge, $\tau_y^h, h = M, MN, N$, for different sectoral histories and retraining wedge, τ_z.[11] With these distortions, the effective local marginal wage rate for a worker with sectoral history h is $(1-\tau_y^h)w$, and the effective local marginal monetary cost of training is $(1-\tau_z)$.[12] The formal definition can be found in Appendix A. They follow from workers' optimality conditions in the decentralized problem.

3 Model Characterization

Conceptually, retraining allocation in the planner's problem specified in the last section hinges on both the training technology and the incentive issues caused by information frictions. To study the economic tradeoff as clearly as possible, in Section 3.1, I first analyze the first-best benchmark where each worker's productivity level is observable. As I will show, productivity-training complementarity plays an important role in training allocation due to efficiency concerns. Next, in Section 3.2, I bring information frictions back and study the constrained problem. There, incentive concerns make productivity-training complementarity push training allocation in the opposite direction. Finally, in Section 3.3, I discuss briefly how such an allocation can be implemented. I also discuss briefly one consequence of having independent tax and retraining systems.

3.1 The First-Best Allocation

When productivity is observable, the planner has no incentive concerns. This effectively removes IC conditions (1)-(4) in the planner's problem. Hence, the planner can maximize social welfare in two steps. First, he maximizes total output by assigning training, occupation, and labor supply properly. Then, given the pie, he chooses the preferred consumption allocation. I first describe occupation, labor, and consumption, then turn to training assignment. Given the productivity bundle after the training stage, occupation allocation is simple: it coincides with a worker's comparative advantage. That is, an M-origin worker will move to N if and only if $nw_n > mw_m$. Labor supply is not distorted, as redistribution is achieved through lump-sum transfer. As for consumption, the planner will allocate consumption such that a worker's marginal utility is proportional

[11]There is actually another distortion on the occupational choice margin. Because it involves a discrete choice, I cannot define a corresponding wedge based on first-order conditions. I characterize the distortion in some special cases in the Appendix and also discuss it in the quantitative part.

[12]They are not constants but functions of worker productivities and training levels.

to his Pareto weight. In particular, if the planner is utilitarian, consumption will be equalized regardless of a worker's productivity level. The proof of those statements can be found in Appendix B.1.

Training allocation is chosen to maximize the output gain. The planner allocates training to expand the productivity set as much as possible. Then, how should workers with different levels of productivity sort into different levels of training intensity? Should a worker with higher manual productivity get more or less training than a worker with lower manual productivity? We analyze the costs and benefits of an extra amount of training for low- and high-manual productivity workers. The cost contains two parts: a direct cost in utility and money terms, which is assumed to be the same for different workers; and an opportunity cost of leaving the manual sector. The latter is higher for a worker with higher manual productivity. The benefit is a higher chance of drawing a better non-manual productivity n.[13] The size of the benefit depends on the training technology, and in particular on the productivity-training complementarity captured by G_{zm}. If $G_{zm} < 0$, manual productivity and training complement each other, meaning high-productivity workers benefit relatively more from the same amount of training. If $G_{zm} > 0$, they are substitutes. In other words, training has diminishing returns in a worker's manual productivity. However, it is hard to draw any definite prediction in the general case. To make things more transparent, I impose some additional parametric assumptions and characterize training allocation in Proposition 1.

Proposition 1. *Assume 1) the planner is utilitarian; 2) worker utility is* $\psi(l) = \frac{l^{1+\sigma}}{1+\sigma}$ *with* $\sigma > 1$*; and 3)* $1 - G(n|z,m) = \frac{zm^p}{n^\alpha}$*. Denote by* λ *the multiplier on the planner's budget constraint. Then*

1) $z(m) = 0$ *for all* m *if* $\lambda \geq (1+\sigma)^{\sigma^2-1}$ *where* λ *is the multiplier on the resource constraint;*

2) for any interior solution z*,* $z(m)$ *increases in* m *iff* $p > \alpha - \frac{1+\sigma}{\sigma}$*;*

3) for any interior solution z*, retraining wedge* $\tau_z^{FB}(m) > 0$*, and* $\tau_z^{FB}(m)$ *increases in* m *iff* $p > \alpha - \frac{1+\sigma}{\sigma}$.

Proof. See Appendix B.2. □

Notice λ measures the marginal value of one unit of additional social resources. When λ is

[13] As one will see, data implies workers benefit from retraining. In other words, $G_z < 0$ so that their productivity draw distribution improves in FOSD sense as they take more training.

high, social resources are very scarce. In that case, the planner would never use retraining because in that case, retraining is inferior to direct monetary transfer due to its intrinsic risky nature. Only when social resources are abundant, i.e., λ is small, may the planner consider training as a policy tool.

The second part of Proposition 1 tells us how the planner should optimally sort different workers into different levels of training intensity when the training technology follows a Pareto distribution.[14] In this case, the power p captures the strength of productivity-training complementarity. As m goes up, the opportunity cost of leaving M increases. Therefore, if p is small, the benefit from retraining is not enough to cover the cost, and workers with higher m get less training. However, if p is large so that high-m workers gain much more from training, the planner will prioritize them to increase total output.

The last part characterizes the retraining wedge in the first-best allocation relative to the decentralized economy. First, as long as training is used, the planner wants to distort training upward compared to laissez-faire. The source of market failure here is the incomplete insurance markets against risky training outcomes. Workers are discouraged from choosing the socially optimal training level, because a bad draw would leave them with low returns, despite the monetary and effort costs. Second, the retraining wedge in the first-best increases with productivity m if and only if the complementarity between m and z is strong enough in the training technology. This follows the same intuition as the training assignment $z(m)$. Without information frictions, $\tau_z^{FB}(m)$ reflects only efficiency concern. It provides a useful benchmark for the constrained problem, where the efficiency force interacts with incentive constraints.

3.2 The Constrained Optimal Allocation

Having described the first-best benchmark, I now add back information frictions and characterize the constrained optimal allocation. I follow the standard practice of using the first-order approach, i.e., replacing global IC conditions with their local counterparts by applying the envelope theorem. The local conditions show how a worker's value changes with productivity. The plan-

[14]The general tradeoff does not depend on the parametric assumption. If we assume G is log-normal, similar pattern will appear quantitatively.

ner must compensate workers for these marginal gains to preserve truthful reporting. I leave the envelope conditions and conditions that guarantee their sufficiency to Appendix C.2 and C.3. The set of sufficient conditions depends on exogenous objects like the training technology G as well as endogenous objects like income and switching threshold, which need to be verified ex post. I next characterize optimal wedges using the first-order approach.

3.2.1 Optimal Labor Wedge

I first study the planner's desired labor supply by characterizing the labor wedge in the constrained problem. For incumbent N workers and M-origin workers who stay in M, their labor wedges resemble the standard Mirrlees formula (Mirrlees, 1971, Saez, 2001), which are left to Appendix C.4. The labor wedge for M-to-N switchers in the next proposition is a more novel result here, as shown next.

Proposition 2. *(Optimal labor wedge) Denote by $\tau_y^{MN}(n,m)$ the optimal labor wedge for workers who originally had productivity m, drew productivity n in the training stage, and switched from M to N. Then,*

$$\frac{\tau_y^{MN}(n,m)}{1-\tau_y^{MN}(n,m)} = \frac{[\mathcal{L}_1(n,m)+\mathcal{L}_2(n,m)]u'(c^N(n,m))}{\lambda g(n|z(m),m)\tilde{f}(m)n}\frac{1+\varepsilon_u}{\varepsilon_c} \tag{5}$$

where ε_u (ε_c) is the uncompensated (compensated) labor supply elasticity, $\mu(m)$ is the multiplier on worker m's first-stage IC condition, and

$$\mathcal{L}_1(n,m) = \lambda\tilde{f}(m)G(n|z(m),m)\left(\mathbb{E}^{\tilde{n}}\left[\frac{1}{u'(c(\tilde{n},m))}\middle|\tilde{n}>n\right]-\mathbb{E}^{\tilde{n}}\left[\frac{1}{u'(c(\tilde{n},m))}\right]\right)$$
$$\mathcal{L}_2(n,m) = -\mu(m)\frac{\partial G(n|z(m),m)}{\partial m}$$

Proof. See Appendix C.4. □

Equation (5) resembles the form of the standard Mirrlees formula, but the interpretation is a little different. $\mathcal{L}_1(n,m)$ captures the risk-sharing motive. Consider a variation for the group of manual workers with initial productivity m: it reduces the utility of workers who drew $\tilde{n}$ higher than n by one unit and rebates them to all m-workers. This variation increases overall welfare for m-workers if $\mathcal{L}_1(n,m) > 0$. In that case, the planner tends to increase the labor wedge at (n,m).

Intuitively, the planner wants workers who luckily draw a good n to compensate for their ex ante identical peers who are not so lucky. This term collapses to zero if workers are risk-neutral.

$\mathcal{L}_2(n, m)$ reflects the incentive concern. $\mu(m)$ captures the shadow cost of compensating for truthful reporting in the first stage locally at m. Consider two workers with productivity m and $m + \Delta m$ for some $\Delta m > 0$ and sufficiently small. If $G_m < 0$, then the worker with productivity $m + \Delta m$ is more likely to draw a good n from the same training. Therefore, a higher m raises information rents through two channels: the worker supplies labor more efficiently, and he gains more from the same retraining assignment. To prevent workers with productivity $m + \Delta m$ from mimicking workers with productivity m, the planner increases the distortion at m even more compared to the standard Mirrlees formula.

3.2.2 Optimal Retraining Wedge

Without information frictions, sorting between worker productivity and training intensity is governed only by efficiency. Whether workers with higher m should receive more training, and whether the training wedge is positive or negative, depends on productivity-training complementarity. With information frictions, one may expect that sorting also depends on the incentive concern. For example, if high-m workers benefit disproportionately more from the same extra amount of training compared to low-m workers, high-m workers can enjoy even more information rent if they are trained more. Therefore, the planner may reduce their training level in the first place to save incentive cost. The next proposition relates the optimal retraining wedge in the constrained problem to that in the first-best case.

Proposition 3. *(Optimal retraining wedge) Denote by $\tau^{SB}(m)$ the optimal retraining wedge in the constrained problem, and it satisfies*

$$1 - \tau_z^{SB}(m) = \mathcal{C}(m)(1 - \tau_z^{FB}(m)) + \mathcal{I}_1(m) + \mathcal{I}_2(m)$$

where $\tau_z^{FB}(m)$ is the retraining wedge in FB, and $\mathcal{C}(m) = f_m(m) \left(\lambda \tilde{f}_m(m) \mathbb{E}^{\tilde{n}} \left[\frac{1}{u'(c(\tilde{n}, m))} \right] \right)^{-1}$. Moreover, for all m such that $\mu(m) > 0$, $\mathcal{I}_1(m) \leq 0$ and $\mathcal{I}_2(m) > 0$ if $G_{zm} < 0$.

Proof. See Appendix C.5. □

First, $1 - \tau_z^{FB}(m)$ on the RHS implies that the inefficiency originating from the incomplete insurance markets for training outcomes carries over, but with an adjustment term $\mathcal{C}(m)$. This adjustment appears only when both information frictions and a redistribution motive are present. When there are no information frictions, $\mathcal{C}(m)$ reduces to one, and we recover the first-best case. Even if there are information frictions, if the planner is utilitarian, and the preference is quasilinear in c so that there is no redistribution motive, $\mathcal{C}(m)$ still reduces to one.

The other two terms, $\mathcal{I}_1(m)$ and $\mathcal{I}_2(m)$, reflect additional incentive concerns that shape the optimal retraining wedge. Recall $\mu(m)$ is the shadow incentive cost. $\mathcal{I}_1(m) + \mathcal{I}_2(m)$ measures, for previous m-workers, how marginally more training, dz, changes the information rent to induce truthful reporting of m. After the training stage, m-workers either stay in M or move to N. When training increases marginally, the group that stays in M shrinks, lowering the total incentive cost for them mechanically. This corresponds to $\mathcal{I}_1(m)$. For those who move to N, if manual productivity and training are complementary, i.e., $G_{zm} < 0$, marginally more training actually raises the incentive cost above m.[15] Hence, to reduce incentive cost in the first place, the planner would like to distort training downward at m. The term $\mathcal{I}_2(m)$ captures this incentive to distort training downward.

Recall our earlier discussion about productivity-training complementarity. In the first-best case, stronger complementarity tends to induce positive sorting between worker productivity and training intensity, as workers with higher m benefit relatively more and can contribute more to total output. That force is absorbed in $\tau^{FB}(m)$. With information frictions, the same complementarity has an additional counteracting force that tends to induce negative sorting between productivity and training intensity, because more training makes distinguishing types more costly.

In Appendix C.1, I also discuss the occupational assignment and redistribution between manual and non-manual sectors in the constrained optimum. Occupational assignment still follows a threshold rule, but different from the first-best case where the threshold coincides with a worker's comparative advantage, here the planner may have an incentive to distort occupational choice. This is again due to incentive concern, as it is possible the saved incentive costs outweigh the out-

[15]Compare a worker with productivity m and a worker with $m + \Delta m$. Marginally more training subsidy at m benefits $m + \Delta m$ relatively more, raising the incentive cost for his truthful reporting. Hence, the planner has an incentive to distort training subsidy at m downwards.

put losses from distorted occupational choice.[16] Finally, at optimum, the value of one marginal dollar must be the same in the two sectors, which is equal to the social marginal value of resources, λ.

3.3 Implementation

Having characterized the optimal allocation in the constrained problem, I discuss how to implement it in a decentralized economy. Consider a retraining schedule $\mathcal{Z} : \mathbb{R}_+ \to \mathbb{R}_+, w_- \mapsto z$ and a history-dependent nonlinear tax schedule $\mathcal{T} : \mathbb{R}_+ \times \mathbb{R}_+ \times \{N, M, MN\} \to \mathbb{R}$, $(w_-, y, h) \mapsto T$, where w_- is the worker's past wage, y is a worker's income, h is his occupational history, and z and T are his retraining level and tax payment, respectively. Without loss of generality, assume T also includes retraining payment. The implementation contains two steps. First, manual workers report their past wage, based on which a training level z is assigned. Then, workers choose their preferred occupation and labor supply, taking into account that their reporting of past wage, employment history, and current income affect the tax schedule $\mathcal{T}$. I show in Appendix C.8 that an incentive-compatible allocation is implementable through a combination of $(\mathcal{Z}, \mathcal{T})$.

Though such a history-dependent income tax schedule may sound impractical in the real world, one can decompose it into different components. For example, it can be split as a history-independent income tax and a sector-training-income-contingent retraining subsidy. Details are left to Appendix C.8.

Finally, why should training subsidies and income taxes be jointly designed? What may go wrong if income taxes depend only on current income, and training subsidies depend only on training intensity? In that case, changes in $T(y)$ may distort the training incentive. I show in an example in Appendix C.9, higher income subsidy can reduce a worker's desired training level. As I will discuss in the next two sections, these two independent instruments resemble German practices during the period of study. In this case, a more redistributive tax schedule distorts retraining incentives downward. The reason is that the gains from training become less attractive relative to training disutility. The distortion is especially strong for low-productivity workers: a given transfer lowers their marginal utility more quickly, which weakens their incentive to retrain.

[16] A general characterization is difficult. I study an example of distorted occupational assignment in Appendix C.6.

4 Data and Calibration

The theoretical model offers some general and qualitative guidance for redistribution across disrupted and undisrupted sectors through income tax and training subsidy, where the feature of the training technology plays an important role. In this section, I first introduce the data and document basic patterns of retraining participation. Then, I provide suggestive evidence on the effectiveness of training programs using a matching event-study design. After that, I move on to estimate the training technology and some other parameters within the structural model, which are the key inputs to the quantitative exercises in the next section.

4.1 German Labor Market Data and Institutional Background

Retraining is an important active labor market policy in many developed economies that helps workers to gain skills and prepare for new jobs. I focus on Germany in this paper because it has a well-developed retraining system, and worker participation is well-documented. The dataset that I use to estimate the model is called the Sample of Integrated Labour Market Biographies (SIAB), which is a 2% sample of the German workforce. It contains detailed labor market histories and importantly, information on retraining participation. I use data from 2000-2021 since extensive retraining records appeared from 2000 on.

SIAB documents three types of retraining programs: activation and vocational integration, career choice and vocational training, and vocational retraining and further education. Activation and vocational integration aims at helping workers to enter or reenter the labor market. These programs provide job-search assistance, language training, and other short courses. Career choice and vocational training provides, for example, career advising and government-sponsored internships, which help workers to discover career possibilities and accumulate new skills. Vocational retraining and further education is a more formal way to learn a new skill. The curriculum ranges from marketing and health-care assistance to skills used in the solar energy industry (Conwell et al., 2024).

Most of the training programs, especially vocational retraining and further education, are administered through a voucher system. Caseworkers in the employment agency award vouchers to workers based on their needs as well as their chances of labor market success. Workers can redeem the voucher at available training providers. Such vouchers cover all direct costs of the available training programs and basic living costs, and sometimes they even cover transportation and childcare costs. However, it is often not easy for caseworkers to make a detailed assessment about a worker's chance of success. In addition, whether or how to award a voucher also depends on other factors like the availability of courses at the time. Those issues add some randomness to the voucher awarding process (Conwell et al., 2024, Osikominu, 2013).

I model training as a continuous variable in the theoretical part, which yields analytical convenience and can be viewed as choosing the time spent in training. Though the dataset contains the starting and ending dates of each training episode, they can be misleading and introduce measurement errors.[17] Therefore, to map the model to the data,[18] I follow the literature and label the first two as short-term training (ST), as they are relatively short, lasting from a few weeks to several months, and the last one as long-term training (LT), since it is longer and can last for as long as 2-3 years (Osikominu, 2013).[19]

I construct a worker-year panel, where each observation contains a worker identifier and the observation year, employment status, establishment identifier, daily wage, days of working, occupation, job tenure, education, state of residence, training program participation, and other demographic information. The data construction process follows standard practices in the literature, see, for example, Stüber et al. (2023).

4.2 Basic Pattern of Retraining Participation

Retraining programs target primarily unemployed workers (Conwell et al., 2024). I restrict the sample to unemployed workers who had a job before. That is, I exclude new entrants or trainees who newly join the labor force. Table 1 shows some basic patterns for retraining participation in Germany.

[17] For example, for the first two types, the registered time span can be several months, but the worker only gets trained for a few days.

[18] When one solves for the constrained optimal allocation numerically, one has to discretize the training intensity anyway. So what I am doing here matters up to the number of discretization.

[19] Different training programs may offer very different curricula. Here, I use duration as a univariate sufficient statistic to measure how much training a worker obtains and make different programs comparable. This is a standard way of modeling human capital accumulation. In addition, for long-term training, I only label an instance as LT if the training episode is longer than three months to avoid dropouts as much as possible. Otherwise, I only mark the observation as no training.

Table 1: Retraining participation patterns in Germany

	NT (%)	ST (%)	LT (%)
Overall fraction	86.29	7.55	6.16
Gender			
male	41.33	4.31	3.18
female	44.96	3.23	2.98
Education level			
no vocational training	24.26	2.56	1.31
vocational training	54.98	4.50	4.43
university and above	6.93	0.47	0.56
Age group			
$<= 25$	28.20	2.40	1.16
$26 - 35$	28.55	2.00	2.10
$36 - 50$	21.45	2.20	2.39
$>= 51$	8.09	0.95	0.50
Pre-U occupation group			
unskilled manual	17.78	1.74	1.61
skilled manual	19.85	1.77	1.24
technicians and engineers	5.25	0.30	0.40
unskilled service	22.07	2.75	1.69
skilled service	7.44	0.49	0.36
semiprofessions	8.82	0.45	0.42
professions	2.86	0.10	0.09
managers	2.25	0.12	0.15

Notes: NT = no training; ST = short-term training; LT = long-term training.

Of all the unemployment spells, about 14 percent ever participated in some form of training, while slightly more workers took short-term training than long-term training.[20] Male workers are slightly more likely to participate in training programs than female workers. In terms of education, workers at all levels of education participate in training programs, though those with lower educational attainment are more likely to participate. Similarly, there are non-trivial participation rates for all age groups. In terms of the pre-unemployment occupational groups, manual workers, either skilled or unskilled, as well as unskilled service workers are most likely to participate. Workers from other occupational groups also participate, with a slightly lower rate.

4.3 Suggestive Evidence about the Effectiveness of Retraining in Germany

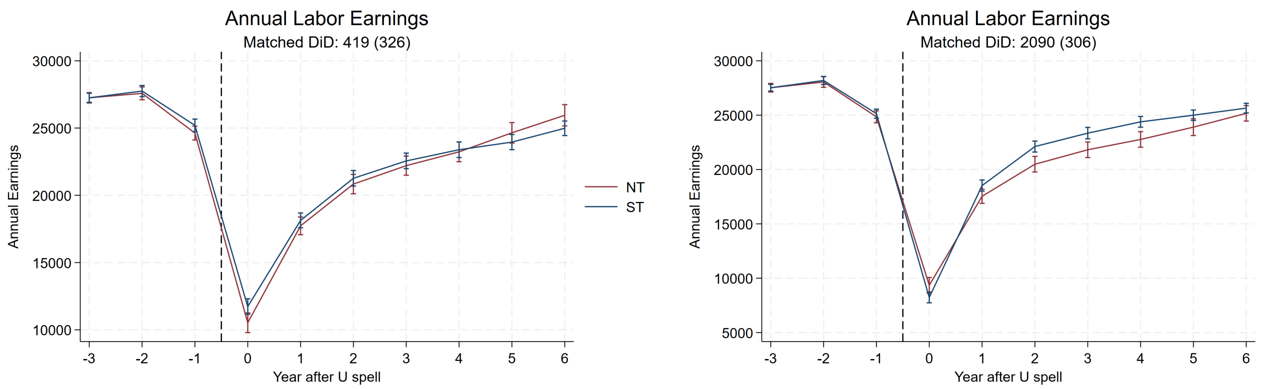
One may be concerned that what workers learn in those programs is actually useless, or workers only participate to make sure they are eligible for unemployment insurance or other subsidies while never exerting effort. In this section, I show some suggestive evidence about the overall effectiveness of training programs in Germany using a matching method, while in the next section, I formally estimate the training technology using the structural model.

The idea is to match workers who take training to otherwise similar workers who never do, and then compare their labor-market outcomes over time. Specifically, for each worker who may have multiple EUE spells, I only keep their first EUE spell to avoid the scarring effect documented in the literature (Jarosch, 2023). In their first EUE spell, I label training participation as NT, ST, or LT. I implement the matching separately for short-term and long-term training. For each exercise, I require workers to have positive earnings in all three pre-unemployment years. Then, I match treated workers (i.e., workers who participate in training programs), and control workers (i.e., workers who never participate) based on their 1) gender, 2) education level, 3) occupation group two years before unemployment, 4) age group two years before unemployment, and 5) earnings in the three years prior to unemployment. For the first four characteristics, I use exact matching, while for the last one, I use propensity score matching with logit within each exactly matched group. The matching is one-to-one so that every treated worker in the final sample has a control

[20] It is possible that a worker may take both short-term and long-term training at the same time. Workers may have short-term training at first, and then switch to long-term programs (Osikominu, 2013). In that case, I label such observation as long-term training.

counterpart.[21] Figure 1 shows the time path of annual earnings for treated and control groups in the matched sample, separately for short-term and long-term training programs. The matched difference-in-difference estimate is also included in the figure, with standard errors in parentheses.

Figure 1: Annual labor earnings for different worker groups in the matched sample



Notes: This figure shows the trajectory of annual labor earnings before and after unemployment for the no training (NT), short-term training (ST), and long-term training (LT) groups. In each panel, different worker groups are matched one-to-one based on 1) gender, 2) education level, 3) occupation group two years before unemployment, 4) age group two years before unemployment, and 5) earnings in the three years prior to unemployment. For the first four characteristics, I use exact matching, while for the last one, I use propensity score matching with logit within each exactly matched group. The matched diff-in-diff coefficients are also included.

Figure 1 indicates that both short-term and long-term training have some positive effects on earnings. Short-term training raises earnings growth slightly relative to no training during the first four years after unemployment, but the effect vanishes after that. Hence, the overall effect, as summarized by the diff-in-diff estimator, is positive but not statistically significant. In contrast, the effect from long-term training on earnings is positive and significant throughout.

Similar figures are plotted in terms of days of employment, which I report in Appendix D.1. Overall, the matching event study exercise suggests that, setting the cost aside, training programs, especially the long-term ones, are indeed effective for future labor market outcomes for unemployed workers. In the next subsection, I will estimate the training technology within the structural model. As one will see, different approaches lead to similar results.

[21]The control groups for the two matching exercises are not necessarily the same.

4.4 Calibration Strategy

4.4.1 External Calibration

I choose worker utility function $u(c,l) = \frac{c^{1-\varphi}}{1-\varphi} - \frac{l^{1+\sigma}}{1+\sigma}$, and set $\varphi = 1$ and $\sigma = 2$, i.e., workers have log utility from consumption and a constant Frisch elasticity equal to 0.5. I use a quadratic approximation to model the income tax schedule in Germany, as shown in panel A of Table 2.

To map a two-sector model to the data, I need to define sector M and N in the data first. In this paper, I focus on policies that help automation-disrupted workers.[22] Therefore, the natural definition of M is the manual sector, and N stands for the non-manual sector. I label all workers in the low- and high-skilled manual as in sector M, while all other workers are in sector N.[23]

For the reason that I have explained in the institutional background, I discretize training level $z = NT, ST, LT$ in the model to map it to the data. They stand for no training, short-term training, and long-term training, respectively. The monetary cost of training comes from Osikominu (2013). The cost for short-term training is 560 in 2000 dollars, while the cost for long-term training is 5,850 in 2000 dollars. I express these costs as shares of annual GDP per capita to make them comparable across quantities in the model. Hence, short-term training costs 2.57% of GDP per capita, and long-term training costs 26.84% of GDP per capita. Finally, I set government endowment E to zero. All the externally calibrated parameters are shown in panel A of Table 2.

4.4.2 Internal Estimation

Next, I internally estimate the initial productivity distribution, the training technology, and training disutilities. I proceed in three steps.

Step 1: Measure productivity m and n

I first measure productivity m or n within an employment spell for each worker. According to my model, worker i's wage per unit of labor supply, ω_i, is the product of his productivity, k_i, and

[22]The model is versatile for other policy studies as well. For example, by defining M as the AI-disrupted sector, one can study the optimal income taxation and retraining policy with the fast advent of AI. By defining M as workers who are trapped by the unemployment scar, one can study policies that help those workers.

[23]The classification follows Schimpl-Neimanns (2003) which groups the three-digit KldB88 occupation coding scheme used by SIAB into 12 broad categories. Groups 2 and 3 include unskilled and skilled manual occupations.

occupation-specific wage rate, w_k, i.e.,

$$\omega_i = k_i w_k, k = m, n$$

I normalize the occupation-specific wage w_k to one in both sectors,[24] then the observed ω_i is exactly worker $i's$ productivity level. In the data, other factors would also affect a worker's wage rate. To get rid of them, I first run the following regression separately for workers in sector N and in sector M:

$$\omega_{it} = \mathcal{E}_{iq(it)} + \eta_t + X_{it}\beta + \varepsilon_{it}$$

where ω_{it} is the (log of) wage rate for worker i in observation year t, $\mathcal{E}_{iq(it)}$ is a worker-by-employment-spell fixed effect, which I discuss in more detail below, η_t is a time fixed effect, and X_{it} contains time-varying variables like experience, job tenure, and an indicator of full-time employment.

To better understand the term $\mathcal{E}_{iq(it)}$, consider a worker i who has spell history $E_1U_1E_2U_2E_3$. Then, $\mathcal{E}_{i1}$ can be viewed as the average wage after controlling for other fixed effects and covariates during employment spell E_1. So are $\mathcal{E}_{i2}$ and $\mathcal{E}_{i3}$. I then use $\mathcal{E}_{iq(it)}$ as a measure of the worker's (log) productivity m or n within his q-th E spell, depending on which sector the worker is in and up to the normalization.[25]

Two things need discussing about this regression. First, because I add worker-by-employment-spell effects, other worker-level time-invariant variables like worker demographics cannot be controlled, but they may affect training choice and training outcome. In terms of training choice, as I have shown, training participation rates across different demographic groups do not vary by a huge amount. As for training outcomes, the implicit (and maybe strong) assumption here is that the training technology is the same for different demographic groups. One can view the estimated

[24]Normalizing wage rates in both sectors to one is without loss of generality. In the model, all worker-level decisions are made based on the effective wage rate $wk, k = m, n$, not productivity k alone, while its effect on the training technology is absorbed by the coefficients. One caveat is that this normalization does matter when we interpret the magnitude of the wage shock in the quantitative exercise. I demonstrate the results under different shock magnitudes there for robustness.

[25]I rescale the estimated fixed effects to match the median wage in Germany.

training technology as applying to an average worker. In principle, one can estimate everything within finer demographic groups. However, even a sample as large as SIAB is still not large enough to allow separate estimation of the training technology for different demographic groups, especially for high-wage workers. Finer estimation of the training technology for different demographic groups (age, education, gender) is left for future research.

Second, one may think of m and n as manual and non-manual skill, respectively. However, it is better to view m and n as the combination of sector-specific human capital and general human capital, and the latter is transferable. This explains why, as one will see soon, m and n are positively correlated. What matters here is not their innate skill, e.g., manual, cognitive, interpersonal, but their effective productivity in each sector.

Step 2: Estimate initial productivity distribution I assume the initial productivity distribution $\tilde{f}_k(k), k = m, n$ is log-normally distributed with log mean μ_0^k and log standard deviation σ_0^k. Given the estimated productivity m and n for workers in the M and N sector respectively, I pool all worker-year observations together. Then, I fit the productivity distribution with a log normal distribution and read the resulting μ_0^k and σ_0^k directly. The measure of sector M workers $\mathbb{M}_M$ is the share of sector M workers, while $\mathbb{M}_N = 1 - \mathbb{M}_M$.

Step 3: Estimate the training technology

The theory highlights the training technology $G(n|z, m)$ as a key determinant of optimal training assignment.

In the empirical implementation, I replace the Pareto specification with a log-normal distribution for $G(n|z, m)$, as a log-normal distribution can fit the income distribution quite well over most of its support. In particular, the distribution has log-mean $\mu(m, z)$ and log-std $\sigma(m, z)$. Then, I further parametrize $\mu(m, z)$ and $\sigma(m, z)$ as flexible functions of m and z. Specifically,

$$\begin{aligned} \mu(m, z) &= \mu_1^z + \mu_2^z \log m + \mu_3^z (\log m)^2 \\ \sigma(m, z) &= \sigma_1^z + \sigma_2^z \log m + \sigma_3^z (\log m)^2 \end{aligned} \tag{6}$$

This general functional form nests many special cases. For instance, one may simply model the new productivity production technology as Cobb-Douglas with a multiplicative log-normal shock, i.e., $n = \zeta_z m^{\alpha_z}$, where $\zeta_z \sim$ log-$N(\mu_z, \sigma_z)$, and α_z is a constant. This is equivalent to setting $\mu_1^z = \mu_z$, $\sigma_1^z = \sigma_z$, $\mu_2^z = \alpha_z$, and all other coefficients to 0 in the above functions.

For empirical purposes, I assume when workers choose their retraining level, there is an idiosyncratic observation noise in their expected value under each retraining choice, which I call "caseworker shock". This is consistent with the discussion in the institutional background.[26] With this shock, workers with the same productivity level m may select into different levels of retraining. Assume the caseworker shock is drawn from a type-one extreme value distribution with location parameter 0 and scale parameter ν, and the shock is independent of worker characteristics and training programs.[27]

The training technology, training disutilities, and the scale parameter of the caseworker shock are jointly estimated by targeting moments from the data. Conditional on a specific productivity m, three sets of moments are chosen: the probability of choosing a training level z, $\mathbb{P}[z^* = z|m]$; the probability of switching sectors given training level z, $\mathbb{P}[S = 1|z, m]$; and the mean of productivity n conditional on training level z and switching, $\mathbb{E}[n|z, m, S = 1]$. Detailed information about the construction of these moments can be found in Appendix D.3. The next proposition shows that those moments can recover the structural parameters through the lens of the model.

Proposition 4. *(Identification) Assume that a worker's selection of the training intensity depends only on the expected value conditional on m and the caseworker shock, and the income tax schedule depends on income alone. Then, the μ's and σ's in the training technology (6), the scale parameter of the caseworker shock, ν, and training disutility* ϕ^{ST}, ϕ^{LT} *(normalizing* $\phi^{NT} = 0$*) can be identified from* $\mathbb{P}[z^* = z|m]$, $\mathbb{P}[S = 1|z, m]$, *and* $\mathbb{E}[n|z, m, S = 1]$ *at each z and at least three different m values.*

Proof. See Appendix D.2. □

To understand this result, let's first look at the μ's and σ's that govern the training technology in (6). Because for each training level z, the mean and the standard deviation of the training

[26]This is a measurement error shock coming from the caseworker, not a preference shock from workers themselves. In particular, the shock will not change the expected value workers get under each training choice.

[27]This does not mean retraining choice is totally random. This model does have self-selection based on the expected value from different training programs. Loosely speaking, the caseworker shock helps to smooth the discrete training choice.

technology are quadratic functions of m, knowing their values at three different m points is sufficient to back out the coefficients of the quadratic function. Two moments help us to pin down $(\mu(m,z), \sigma(m,z))$ at a given (m,z) pair: the switching probability, $\mathbb{P}[S=1|z,m]$, and the expected n conditional on switching, $\mathbb{E}[n|z,m,S=1]$. They are actually truncated probability and truncated mean, and they jointly and uniquely determine the shape of the log-normal distribution. These two sets of moments are the primary source of identification for $g(n|z,m)$.

As for the scale parameter of the caseworker shock, ν, and training disutility, $\phi^{ST}$and ϕ^{LT}, they are identified from the distribution of the participation rate over training levels for different levels of m. To see why, notice that given the productivity production function, we can compute a worker's value if he chooses training level z. Without the extreme value shock, workers with the same productivity m will choose the same training level. Therefore, within-m dispersion in different training levels informs the scale parameter ν. After accounting for the within-m dispersion, the across-m dispersion over different training levels informs $\phi^{ST}$and ϕ^{LT} up to a normalization.

In practice, I choose m to be the 5th, 15th, ..., 95th percentiles of the manual productivity distribution and $z = NT, ST, LT$ to calculate the above moments. I also calculate $var[n|z,m,S=1]$, the variance of non-manual productivity for the switched group, which is an untargeted moment. I solve the model under the actual policy and generate the corresponding moments. I choose the set of parameters that minimize the distance between model-generated and data moments.

4.4.3 Estimation Results and Model Fit

All the internally estimated parameters are shown in panel B of Table 2.

Figure 2 plots the targeted moments computed from the data (solid line) and from the model (dashed line) for long-term training programs. Similar figures for short-term training and no training are reported in Appendix D.4. The participation rate in long-term training is hump-shaped with respect to worker productivity level, though the variation is limited. The rate of successful sectoral switching declines quickly with productivity level, while conditional on successful switching, the average new productivity n increases very fast with the old productivity m, suggesting strong selection in sectoral switching. The same selection pattern also appears in the short-term training and no-training groups.

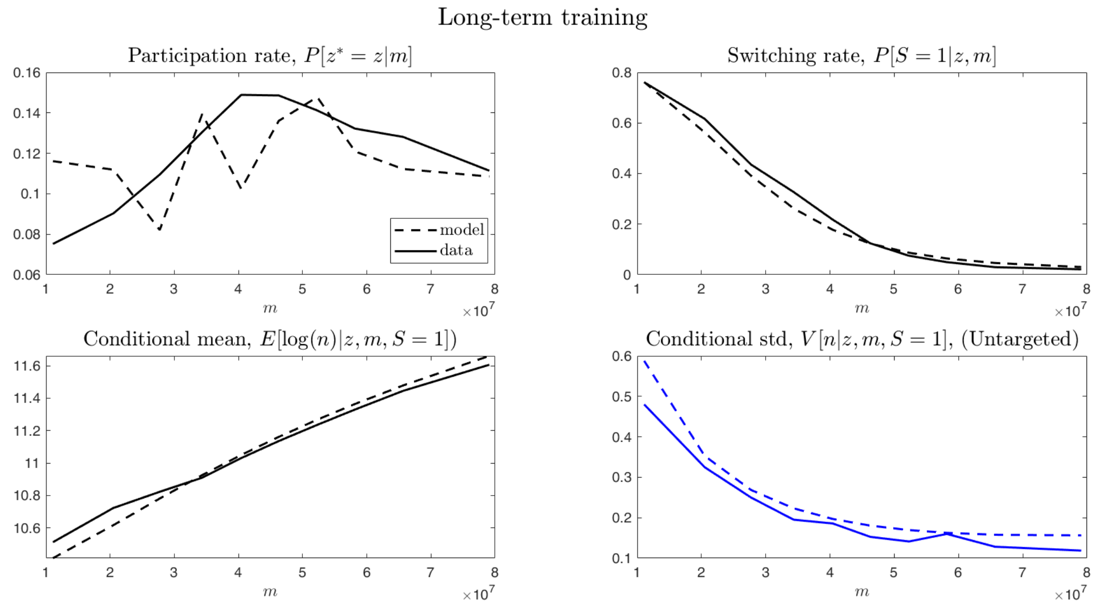
Table 2: Parameters Used in the Model

Parameter	Value
Panel A: Externally Calibrated Parameters	
Relative risk aversion: φ	1
Frisch labor supply elasticity: σ	0.5
Actual income tax function: $T(y)$	$-9630 + 0.35y + 1.45 \times 10^{-7}y^2$
Short-term training cost: c_{ST}	2.57% of GDP per capita
Long-term training cost: c_{LT}	26.84% of GDP per capita
Government endowment: E	0
Panel B: Internally Estimated Parameters	
Initial measure of sector M workers: $\mathcal{M}_m$	0.2347
Initial log mean of sector M workers: μ_0^m	11.0506
Initial log std of sector M workers: σ_0^m	0.4301
Initial log mean of sector N workers: μ_0^n	10.7445
Initial log std of sector N workers: σ_0^n	0.6298
$\mu^{NT}(m)$	$2.94 - 0.13\log(m) + 0.063\log(m)^2$
$\mu^{ST}(m)$	$2.87 - 0.067\log(m) + 0.06\log(m)^2$
$\mu^{LT}(m)$	$2.92 - 0.016\log(m) + 0.046\log(m)^2$
$\sigma^{NT}(m)$	$2.25 - 0.64\log(m) + 0.044\log(m)^2$
$\sigma^{ST}(m)$	$2.3 - 0.78\log(m) + 0.071\log(m)^2$
$\sigma^{LT}(m)$	$2.72 - 1.2\log(m) + 0.16\log(m)^2$
Caseworker shock scale parameter: ν	0.1529
Short-term training disutility: ϕ^{ST}	0.0216
Long-term training disutility: ϕ^{LT}	0.0334

The model matches these moments closely. The last panel with blue lines in Figure 2 illustrates the conditional variance of $\log(n)$, i.e., $var(\log(n)|z, m, S = 1)$, in both the model and the data, which is untargeted. The model-implied conditional variance matches the data very well.

Figure 3 shows the estimated log mean $\mu(z, m)$ and log standard deviation $\sigma(z, m)$ of the training technology as a function of training z and former productivity m. In terms of log mean, there is a clear ranking of different training levels: at each productivity m, long-term training (dotted line) yields the highest log mean, while no training (solid line) yields the lowest log mean. In addition, long-term training is especially helpful for workers with low m, but its effectiveness slows down and finally converges to short-term training as m increases. As for log standard deviation, long-term training generates the lowest variation, while no training generates the highest variation. Only for workers with high m, the outcome from long-term training fans out. Another point worth mentioning is that the right panel also shows substantial risk in switching sectors, especially for low-m workers. Training programs can reduce such risks to some extent but not

Figure 2: Targeted and untargeted moments



Notes: This figure shows the data and model-generated moments for long-term training, including the participation rate, the occupational switching rate, the average and the standard deviation of the non-manual productivity conditional on long-term training and switch, for different manual productivity levels. The conditional standard deviation is not a targeted moment.

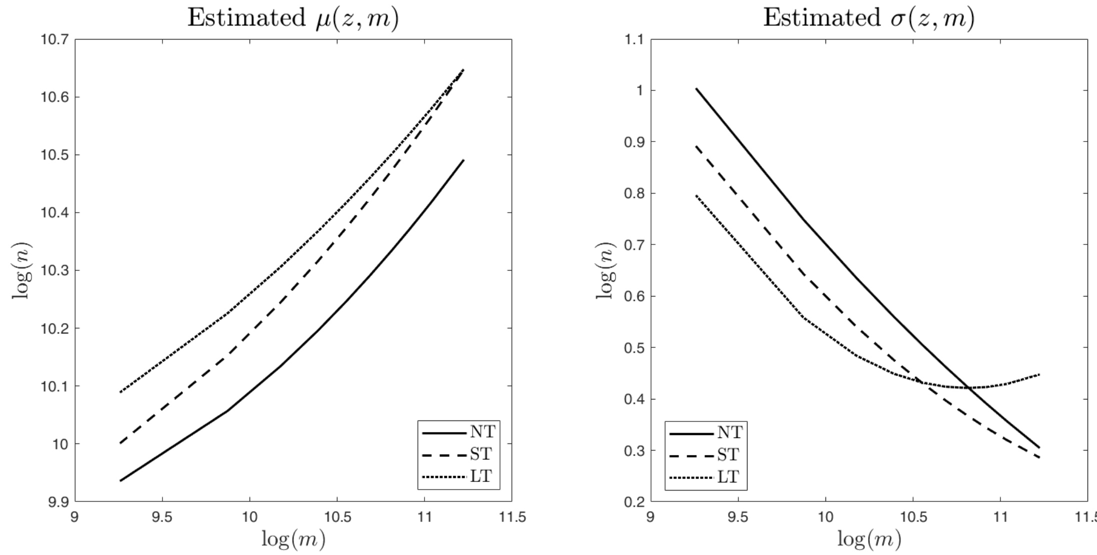
all. There is still a significant amount of risk in the training outcomes. Overall, the shape of the training technology is broadly consistent with the suggestive evidence about the effectiveness of training programs in Section 4.3.

Finally, it is helpful to have a sense of training-productivity complementarity in the estimated training technology. One difficulty is that, in the theory part, this object is captured by the cross partial derivative G_{zm}. But now training level is discrete. Nevertheless, we can replace the partial derivative with respect to z with the partial difference in z. In particular, the left panel of Figure 4 plots $G_m(n|ST, m) - G_m(n|NT, m)$, i.e., from no training to short-term training; while the right panel shows $G_m(n|LT, m) - G_m(n|ST, m)$, that is, from short-term to long-term training. Each panel shows the object for m at the 10th, 50th, and 80th percentiles.

The main takeaway is that complementarity is strongest when workers move from no training to short-term training. For most of the relevant mass of n, $\Delta_z G_m$ is negative, especially for higher-m workers.[28] By contrast, the move from short-term to long-term training shows little complemen-

[28]For each of the three lines on the left panel, the mass of the draws concentrates at the trough area.

Figure 3: Estimated $\mu(z, m)$ and $\sigma(z, m)$



Notes: This figure plots the log mean and log standard deviation of the training outcome distribution, which is assumed to be log normal, as a function of previous manual productivity m, and training level $z = NT, ST, LT$.

tarity across manual productivity levels. Economically, these patterns imply that workers with a higher productivity m gain relatively more from short-term training compared to those with a lower m. But such advantages disappear if they further pursue long-term training.

5 Quantitative Analysis

In this section, I apply the calibrated model to study the optimal income taxation and retraining policy after an unexpected automation shock, modeled as a drop in the wage rate w_m in the manual sector M. Because the model takes initial sectoral sorting as given, the policy experiment applies to workers already allocated across the two sectors.[29]

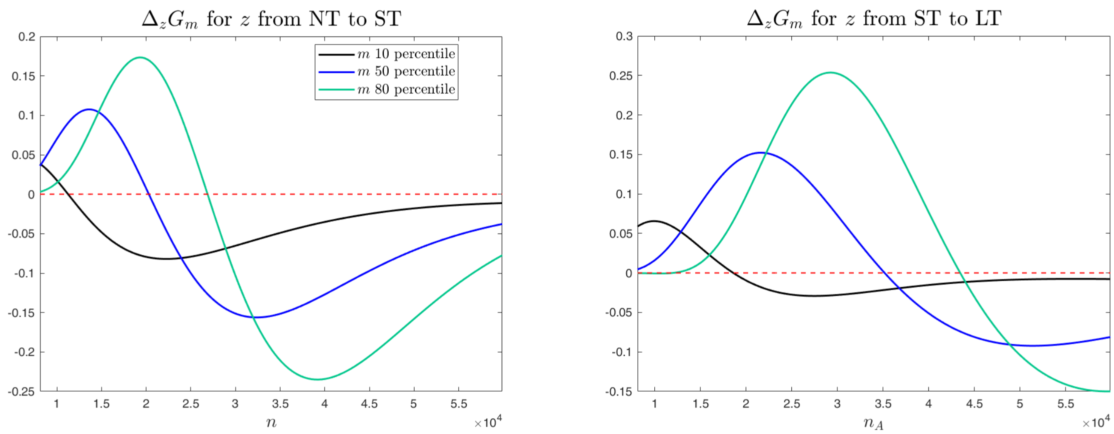
In the baseline case, I assume a 40 percent decline in w_m, namely, $w_m = 0.6$, while w_n remains at 1.[30]

Retraining may raise human capital over several years, whereas workers live for only one period in the model. I therefore annualize training costs by dividing them by t. Based on the

[29] Alternatively, one can view the experiment as a policy response to an unanticipated shock.

[30] A 40 percent wage drop may seem large, but considering that there is no unemployment in this model, the 40 percent wage drop also reflects losses from being displaced.

Figure 4: Empirical training-productivity complementarity



Notes: This figure demonstrates the empirical complementarity between training and manual productivity. Since training level is discrete, the complementarity is defined by $\Delta_z G_m$, where differentiation is taken with respect to manual productivity m, and the difference is taken with respect to different training levels. See the main texts for the formal definition. The left panel shows the complementarity from no training to short-term training at different manual productivity levels, while the right panel shows that from short-term to long-term training. The value being negative means there is local training-productivity complementarity, and being positive means there is local substitutability.

suggestive evidence in the matching event study, I set $t^{ST} = 4$ for short-term training while $t^{LT} = 6$ for long-term training. Finally, when solving the constrained problem, I assume the planner can overcome the caseworker observation shock.[31] To make different policy regimes comparable, I also calculate worker allocation and social welfare under the actual policy and under laissez-faire, assuming there are no caseworker observation shocks in any case.

5.1 The Constrained Optimal Allocation

In this section, I show the constrained optimal allocation after the automation shock, including the retraining allocation, the labor wedge, and the occupational switching threshold. The tax schedule that implements the allocation is left to Appendix E.

Figure 5 plots the expected value of M-origin workers from different training levels under a particular implementation.[32] The shaded area is the density of initial productivity distribution in M. As we can see, the expected value from short-term training always dominates those from

[31] Adding that shock will change the nature of the constrained problem, and one can think of the resulting policy as an upper bound on what the constrained planner can achieve.

[32] The value from LT or NT is not unique because they are off-equilibrium path, and the planner can manipulate their values as long as they are not chosen in equilibrium. The optimal training allocation, on the other hand, is unique.

long-term training or no training over the support of m, implying that the constrained planner would like every M-origin worker to go through short-term training before making the new occupational choice decision. One may be surprised that both the empirical evidence in Section 4.3 and estimation results in Section 4.4.3 show that long-term training is the most effective program, but the planner leaves it in the toolbox and never uses it. The reason is cost-effectiveness: long-term training costs more than ten times as much as short-term training, while its additional benefit is not that huge. This result is in accordance with Osikominu (2013): though long-term training is the most effective program, it may not be cost-effective. In contrast, short-term training programs are cheaper, yet they can yield considerable gains.

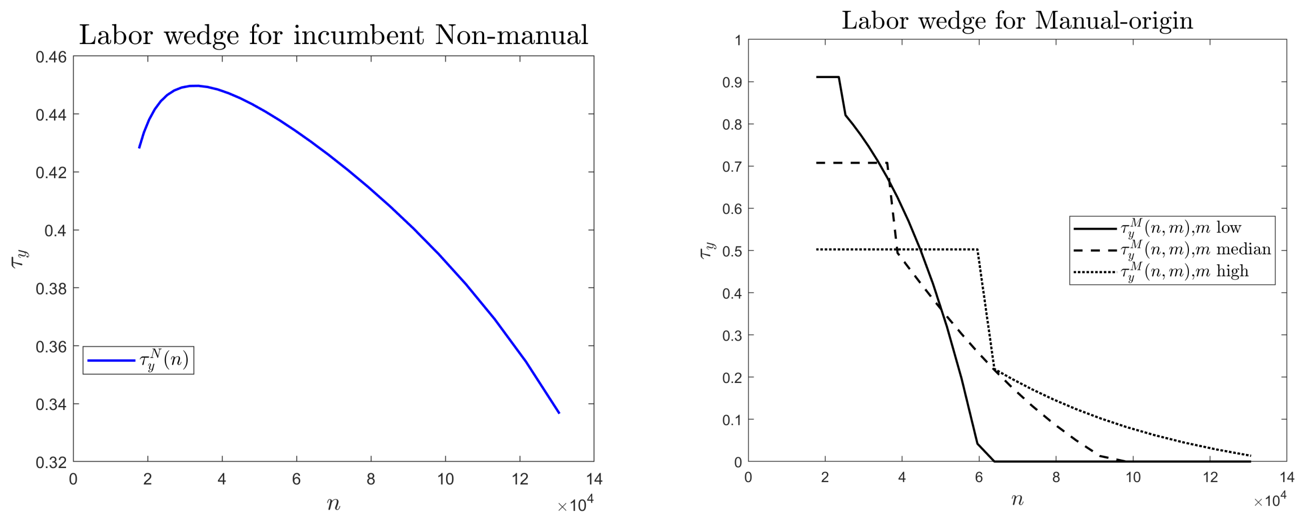
The optimal training allocation corroborates the discussion of productivity-training complementarity in Proposition 3. Figure 4 shows that there is productivity-training complementarity from no training to short-term training. The fact that the constrained planner wants all workers to take short-term training, i.e., there is no strong sorting pattern in worker productivity and training intensity, implies the two forces almost offset each other.

Figure 5: Worker's expected value from different training levels

Notes: This figure plots the expected value of an M-origin worker with manual productivity m from different training levels under optimal policy. The shaded area is the density of M-origin workers over their manual productivity.

Figure 6 plots the labor wedges for incumbent N workers (left) and M-origin workers with different levels of productivity m (right). The labor wedge for incumbent N workers is standard and exhibits the tendency of no distortion at the top. The labor wedge for M-origin workers has richer properties. First, the sudden drop in each curve represents the occupational switching threshold. For a marginal worker at the switching threshold, the labor distortion is lower if he switches to N compared to staying at M. The gap is larger for workers with high m. Second, if a worker stays in the disrupted M sector, the labor wedge is quite high, especially for workers with low m. Intuitively, because the wage rate in M is low after the shock, the planner finds it more effective to reduce the labor supply of M-stayers to save labor disutility. Third, after switching sectors, the labor wedge starts to drop with n for each m. It declines faster for workers with lower m because they face the normal wage rate, which weakens the need to distort their labor supply.

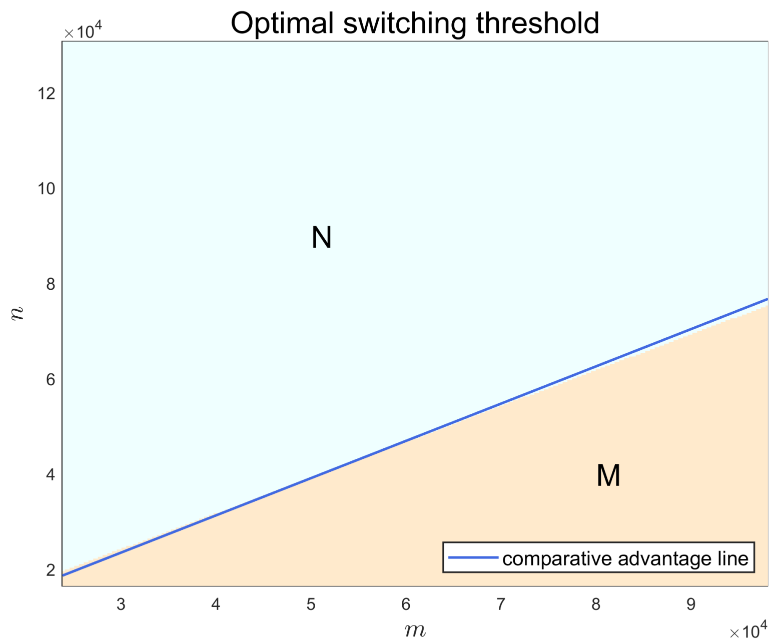
Figure 6: Optimal labor wedges



Notes: This figure plots the optimal labor wedge for incumbent N workers (left) and M-origin workers with different non-manual productivity draws at selected manual productivity levels (right). "m low" means m at the 25th percentile, "median" at the 50th percentile, and "high" at the 75th percentile. The flat area means they stay in M so that the non-manual productivity draw is irrelevant, and the jump means occupational switching occurs there.

Figure 7 illustrates the optimal occupational switching threshold, where the colored area indicates to which sector an M-origin worker with productivity (m, n) should go, and the blue solid line is the boundary of comparative advantage on which $mw_m = nw_n$. The planner's sectoral allocation roughly follows comparative advantage, with small deviations at the bottom and top of the m distribution. Relative to the comparative-advantage rule, the planner keeps slightly more low-m workers in M and moves slightly more high-m workers to N. The reason is, the cost of inducing truthful reporting is cumulative in the sense that if the planner increases the incentive cost to induce truthful reporting at a given productivity m or n, he has to increase the payment to all workers with productivity above that. By keeping more low-productivity workers at M, the planner reduces the starting level of the incentive cost in sector N. Then, the planner can push more high-m workers to N with lower incentive costs.

Figure 7: Optimal occupational switching threshold



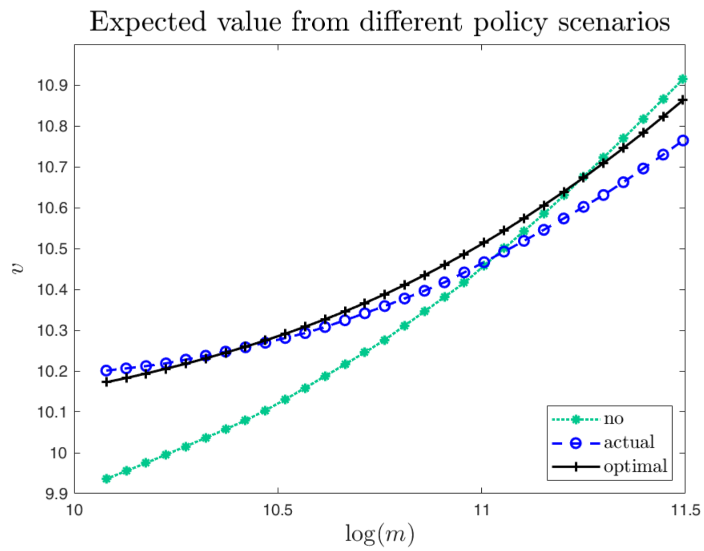
Notes: This figure plots the optimal occupational choice for an M-origin worker with productivity pair (m, n). The light brown area implies the worker would stay at M, while the light blue area implies the worker would go to N. The blue line is the comparative advantage line on which $mw_m = nw_n$.

5.2 Comparison Between Different Policy Scenarios

Section 5.1 shows the optimal allocation after the automation shock. To put that into perspective, I now compare the results with those from alternative policy scenarios, including laissez-faire (no policy intervention) and the actual German policy. In particular, I compare the expected value of M-origin workers and the retraining allocation for different levels of m. To make the results comparable, I assume in all cases, there are no caseworker observation shocks, and tax revenues are rebated equally to all workers.

Figure 8 shows the expected value for disrupted workers with different levels of productivity m after the automation shock under different policy scenarios. Not surprisingly, the span of the expected value (i.e., the difference between the maximum and the minimum) is the largest under laissez-faire, as there is no redistribution at all. In contrast, the actual system reduces that gap through a progressive income tax schedule and essentially free training programs. Under optimal policy, the gap is slightly larger compared to that of the actual policy, but the expected value is higher for most workers. This results from the sectoral redistribution that does not exist in the actual policy, and from the better targeted retraining allocation. I explain the training allocation in more detail next.

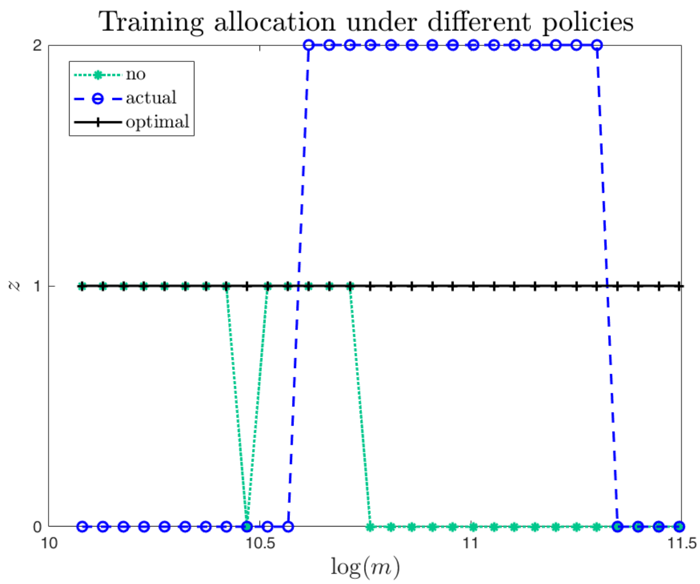
Figure 8: Worker expected value under different policy regimes



Notes: This figure plots the expected value of M-origin workers with manual productivity m under different policy scenarios. Here, "no" means no policy intervention (laissez-faire); "actual" means policy in Germany, where the income tax schedule is uniform for all workers, and retraining is almost free if a worker is awarded a voucher; and "optimal" means the optimal constrained allocation.

Figure 9 shows the training allocation under the three policy scenarios. The black solid line corresponds to the optimal allocation in which all workers go through short-term training. The green dotted line corresponds to the allocation under laissez-faire, where only workers with low-to-median m want to participate in short-term training programs, while workers with median-to-high m do not take any training. The gap arises because laissez-faire workers bear training risk without insurance. This risk discourages efficient training participation, especially among workers with median-to-high m. They already enjoy a relatively high expected value without training, so the effort cost and outcome risk deter participation. For workers with low-to-median m, short-term training can significantly improve their productivity draw, which is captured by a higher mean and a lower dispersion (see Figure 3). However, due to the lack of redistribution and insurance, their expected values are still quite low (see Figure 8).

Figure 9: Training participation under different policy regimes



Notes: This figure plots the training allocation for M-origin workers with manual productivity m under different policy scenarios. Here, "no" means no policy intervention, that is, there is no income tax nor training subsidies; "actual" means policy in Germany, where the income tax schedule is the same for all workers, and retraining is almost free if a worker is approved; and "optimal" means the optimal non-linear tax schedule. To make training allocation available across different policy scenarios, I assume there are no caseworker observation shocks in all cases, and hence, the training level that gives the maximum expected value will be chosen.

The blue dashed line plots training participation under the actual German policy, assuming there are no caseworker observation shocks. Workers with low-to-median m and very high m do not take any training, while workers with median-to-high m take long-term training. For workers with low-to-median m, they already receive substantial transfers under the actual policy. Hence, there is no need for them to spend time in training. As Appendix C.9 formally shows: when the income tax schedule and the retraining system are independent, more transfers reduce the incentive to take training. Workers with median-to-high m gain substantially from long-term training. Because the program is essentially free from their private perspective, they are willing to take it. For top m workers, their value from no training is already high, and they have no motivation to take any training.

Therefore, compared to the optimal policy, the actual policy yields substantial misallocation in training participation. On the one hand, the redistribution system is excessively generous to work-

Table 3: Aggregate statistics under different shock magnitudes

	$\mathbf{w_m}$ **=0.4**	**0.6**	**0.8**	**1**
Redistribution from N to M, % GDP	5.71	2.24	-0.88	-6.33
Total Training Expenditure				
% GDP	2.04	0.6	0	0
% total subsidies to M	9.01	6.6	0	0
Welfare gain, money metric % GDP	0.65	0.38	0.8	1.43

ers with low-to-median levels of productivity m, thus reducing their retraining incentives. This may provide one reason some people have the impression that training programs are not useful, and why shirking and dropouts are a concern for those training programs: given the generous transfers, e.g., UI and income subsidy, not taking training is their optimal choice. On the other hand, the actual regime offers free long-term training. Although this benefits individual workers, it is not cost-effective from a social point of view.

5.3 Training Expenditure and Welfare Gain from the Optimal Policy

Finally, I show some aggregate statistics about the extent of sectoral redistribution, the form of redistribution in terms of cash transfers and tax deductions on one hand, and training provision on the other, as well as welfare gains from changing the actual policy to the optimal one. I calculate these statistics for different magnitudes of the wage shock in sector M. The no-shock case corresponds to $w_m = 1$. Table 3 reports the selected statistics.

The first row is the amount of redistribution from the non-manual sector to the manual sector as a share of total GDP. It increases with the size of the automation shock. In the baseline case ($w_m = 0.6$), total redistribution from the non-manual sector to the manual sector accounts for 2.24 percent of GDP. Notice when the shock is small or in the absence of the shock, the planner actually wants the manual sector to subsidize the non-manual sector (negative number). Without the shock, manual workers earn higher wages than non-manual workers on average. Hence, if we look at the change from -6.33 percent ($w_m = 1$) to 2.24 percent ($w_m = 0.6$), the optimal sectoral redistribution is actually quite large after the shock.

The middle two rows show the total training expenditure as a share of GDP (row 2) and as a share of total subsidies to the manual sector (row 3).[33] When $w_m = 0.6$, the planner will spend 0.6 percent of GDP on training programs. As a reference point, Germany spent about 0.22 percent of its total GDP on various training programs in 2000 (Osikominu, 2013). Of all the subsidies to the manual sector, 6.6 percent is spent on training programs when $w_m = 0.6$. With smaller shocks, the planner stops using retraining as a policy instrument because sectoral reallocation is less urgent, and retraining is not cost-effective. Disrupted workers either switch sectors without training or accept a lower wage while receiving transfers or tax deductions. However, when the shock is very large, e.g., $w_m = 0.4$, total retraining expenditure surges to 2 percent of GDP, and the planner actually wants workers with median-to-high m to participate in long-term training. When the manual wage is very low, the opportunity cost of retraining and occupational switching falls, making LT attractive despite its high direct cost.

The last row reports the welfare gains from adopting the optimal policy instead of the actual policy. To make them comparable, I convert the gains in social welfare to dollar terms,[34] expressed as a share of GDP. In the baseline case when $w_m = 0.6$, the welfare gain is equivalent to 0.38 percent of GDP, which amounts to almost two thirds of total training expenditure.[35]

6 Concluding Remarks

Both income taxation and retraining programs are useful policy tools that help disrupted workers after large sectoral shocks that affect different groups of workers unequally. Those shocks are no longer hypothetical and have become more realistic and tangible threats with the fast advancement of new technologies like automation and AI. While cash transfers and tax deductions through income taxation provide workers with a basic cushion, retraining programs help them learn new skills and leave the affected occupations. These two policy tools should not be used in isolation. A proper tax system is important to incentivize training participation and insure against its risk, and retraining, as one form of human capital investment, increases worker productivity and output, and ultimately tax revenue.

[33] Total subsidies to the manual sector include both within-sector and between-sector redistribution.

[34] In particular, I divide them by λ, the social value of marginal resources.

[35] Interestingly, the welfare gain is not monotonic when the shock becomes larger. One reason is that the optimal allocation involves sector-based income tax, which has large social values even without the shock.

This paper offers some practical policy implications. First, the actual policy might be overly generous to low-wage workers so that they may not have the right incentive to take socially optimal training programs. Second, the government should be more cautious when offering training programs, especially long-term ones, as they may not be cost-effective from a social point of view, especially when the shock is not extremely dramatic. However, those training programs do have considerable effects, as both the empirical evidence and the structural estimation show. Therefore, reducing the cost of training would make it a more valuable policy tool. Online resources and AI-based tools may help lower these costs. It can also be beneficial to offer more training programs with flexible lengths. Lastly, sector-specific taxes or subsidies can substantially improve welfare in response to sectoral shocks.

As a first step toward studying this question, the paper leaves several directions for future work. For example, how should the policy change with demographics like age? How should we formally deal with moral hazard issues where workers may not put enough effort during training? To answer those questions, more detailed data and richer quantitative models are needed.

References

Lucas Conwell, John Finlay, and Trevor Williams. Schooled by trade? retraining and import competition. *Retraining and Import Competition (October 21, 2024)*, 2024.

Arnaud Costinot and Iván Werning. Robots, trade, and luddism: A sufficient statistic approach to optimal technology regulation. *The Review of Economic Studies*, 90(5):2261–2291, 2023.

Bruno Crépon and Gerard J Van Den Berg. Active labor market policies. *Annual Review of Economics*, 8:521–546, 2016.

Sebastian Findeisen and Dominik Sachs. Education and optimal dynamic taxation: The role of income-contingent student loans. *Journal of Public Economics*, 138:1–21, 2016.

Renato Gomes, Jean-Marie Lozachmeur, and Alessandro Pavan. Differential taxation and occupational choice. *The Review of Economic Studies*, 85(1):511–557, 2018.

Joao Guerreiro, Sergio Rebelo, and Pedro Teles. Should robots be taxed? *The Review of Economic Studies*, 89(1):279–311, 2022.

Benjamin G Hyman. Can displaced labor be retrained? evidence from quasi-random assignment to trade adjustment assistance. In *Proceedings. annual conference on taxation and minutes of the annual meeting of the national tax association*, volume 111, pages 1–70. JSTOR, 2018.

Gregor Jarosch. Searching for job security and the consequences of job loss. *Econometrica*, 91(3): 903–942, 2023. doi: 10.3982/ECTA14008.

Michael Lechner, Ruth Miquel, and Conny Wunsch. Long-run effects of public sector sponsored training in west germany. *Journal of the European Economic Association*, 9(4):742–784, 2011. doi: 10.1111/j.1542-4774.2011.01029.x.

Nils Haakon Lehr and Pascual Restrepo. Optimal gradualism. Technical report, National Bureau of Economic Research, 2022.

J. A. Mirrlees. An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, 38(2):175–208, 1971. doi: 10.2307/2296779.

Aderonke Osikominu. Quick job entry or long-term human capital development? the dynamic effects of alternative training schemes. *Review of Economic Studies*, 80(1):313–342, 2013.

Casey Rothschild and Florian Scheuer. Redistributive taxation in the roy model. *The Quarterly Journal of Economics*, 128(2):623–668, 2013.

Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229, 2001. doi: 10.1111/1467-937X.00166.

Bernhard Schimpl-Neimanns. Microdata tools: Implementation of blossfeld's occupational classification to the microcensuses 1973–1998. 2003.

Karl Schulze. Re-skilling, up-skilling, and the role of education in the adjustment to economic shocks. Technical report, Technical Report, Mimeo, 2024.

Stefanie Stantcheva. Learning and (or) doing: Human capital investments and optimal taxation. Technical report, National Bureau of Economic Research, 2015.

Stefanie Stantcheva. Optimal taxation and human capital policies over the life cycle. *Journal of Political Economy*, 125(6):1931–1990, 2017.

Stefanie Stantcheva. Dynamic taxation. *Annual Review of Economics*, 12:801–831, 2020.

Heiko Stüber, Wolfgang Dauth, and Johann Eppelsheimer. A guide to preparing the sample of integrated labour market biographies (siab, version 7519 v1) for scientific analysis. *Journal for Labour Market Research*, 57(1):7, 2023. doi: 10.1186/s12651-023-00335-w.

Uwe Thuemmel. Optimal taxation of robots. *Journal of the European Economic Association*, 21(3): 1154–1190, 2023.

A Define Wedges

The local labor wedge $\tau_y^h, h = M, MN, N$, for different sectoral histories and retraining wedge τ_z are formally defined as follows.

Definition 1. The labor wedge for a M-origin worker with sectoral history $h \in \{M, MN\}$ in sector $I \in \{M, N\}$, satisfies

$$1 - \tau_y^h(n,m) = \frac{1}{u'(c(n,m))}\psi'\left(\frac{y(n,m)}{w_i i}\right)\frac{1}{w_i i}$$

while the labor wedge for an incumbent N worker is

$$1 - \tau_y^N(n) = \frac{1}{u'(c(n))}\psi'\left(\frac{y(n)}{w_n n}\right)\frac{1}{w_n n}$$

The retraining wedge satisfies

$$(1 - \tau_z(m)) \int u'(c(n,m))g(n|z(m),m)dn = -\phi'(z(m)) + \int v(n,m)\frac{\partial g(n|z(m),m)}{\partial z}dn$$

They follow from workers' optimality conditions in the decentralized problem with corresponding wedges. With these distortions, the effective local marginal wage rate for a worker with sectoral history h is $(1 - \tau_y^h)w$, and the effective local marginal monetary cost of training is $(1 - \tau_z)$.

B Planner's Problem under First-best

B.1 First-best Allocation in the General Case

Proposition 5. *(First-best allocation) Under Assumption 1-2,*

(1) Consumption is equalized: $c^N(n) = c^M(n,m) = c^{MN}(n,m) = c$ such that

$$u'(c) = \lambda$$

where λ is the multiplier on the government budget constraint.

(2) Workers who switched from M to N get the same income as incumbent N workers:

$$y^N(n) = y^{NM}(n,m)$$

In addition, labor wedge is zero for every worker.

(3) Occupational choice coincides with individual comparative advantage, i.e., it follows a threshold rule where the threshold $\theta(m)$ satisfies

$$\theta(m) = \frac{w_m m}{w_n}$$

so that a worker with skill m switches if and only if he draws a skill higher than $\theta(m)$.

(4) When retraining z is interior, it satisfies

$$-\phi'(z) - \lambda + (\lambda y^M(m) - \psi(l^M(m)))G_z(\theta(m)) + \int_{\theta(m)}^{\bar{n}} (\lambda y^N(n) - \psi(l^N(n)))g_z(n)dn = 0$$

Proof. Without information friction, the planner can directly choose the retraining intensity, sector, labor supply (or income), and consumption for each worker. The problem is similar to the constrained problem, but the IC constraints are all removed. Concretely, the problem can be formulated as

$$\begin{aligned}
\max_{c,y,,z,S} & \int_m \left[-\phi(z(m)) + \int_n v(n,m)dG(n;z(m),m) \right] dF_M(m) + \int_n \left[u(c^N(n)) - \psi\left(\frac{y^N(n)}{w_n n}\right) \right] dF_N(n) \\
s.t \;\; & v(n,m) = \int_{n\in S} v^N(n)dG(n|z(m),m) + \int_{n\notin S} v^M(m)dG(n|z(m),m) \\
& E \geq \int_m \Bigg[z(m) + \int_{n_A} \Bigg[\Big(c^M(n,m) - y^M(n,m) \Big) \mathbb{I}(n \notin S(m)) \\
& + \Big(c^{MN}(n,m) - y^{MN}(n,m) \Big) \mathbb{I}(n \in S(m)) \Bigg] dG(n;z(m),m) \Bigg] d\tilde{F_M}(m) \\
& + \int_n \left[c^N(n) - y^N(n) \right] d\tilde{F}_N(n)
\end{aligned}$$

We immediately have that occupational choice follows a threshold rule that is consistent with a worker's comparative advantage, namely, for a worker with skill bundle (n,m), the worker stays in the manual sector if and only if

$$n \leq \frac{w_m m}{w_n} := \theta(m)$$

Otherwise, the planner can assign the worker the other sector and supply the same unit of labor. By doing so, total output increases while social cost does not. This also implies that when we know a worker's skill bundle (n,m), we know in which sector the worker will be. Hence, there is no need to add a superscript to indicate a worker's occupational history. With a little abuse of notation, we use $x(n)$ to denote variables related to incumbent non-manual workers, and $x(n,m)$ manual-origin workers when there is no confusion, where x can be consumption c and income y, etc.

Let λ be the multiplier of the resource constraint. Taking First-order conditions, we have

$$
\begin{aligned}
u'(c(n))f_n(n) - \lambda \tilde{f}_n(n) &= 0 \\
u'(c(n,m))f_m(m) - \lambda \tilde{f}_m(m) &= 0 \\
-\psi'\left(\frac{y(n)}{w_n n}\right)\frac{1}{w_n n} f_n(n) + \lambda \tilde{f}_n(n) &= 0 \\
-\psi'\left(\frac{y^M(n,m)}{w_m m}\right)\frac{1}{w_m m} f_m(m) + \lambda \tilde{f}_m(m) &= 0 \\
-\psi'\left(\frac{y^{MN}(n,m)}{w_n n}\right)\frac{1}{w_n n} f_m(m) + \lambda \tilde{f}_m(m) &= 0 \\
\left[-\phi'(z(m)) + \int_{\underline{n}}^{\bar{n}} v(n,m) dG_z(n; z(m), m)\right] f_m(m) - \lambda \left[1 - \int_{\underline{n}}^{\bar{n}} T(n,m) dG_z(n; z(m), m)\right] \tilde{f}_m(m) &= 0
\end{aligned}
$$

where $T(n,m) := y(n,m) - c(n,m)$. Under Assumption 2, we have

$$
u'(c) = \lambda
$$

and

$$
y^N(n) = y^{NM}(n,m)
$$

As for labor wedge, for incumbent non-manual workers,

$$
\begin{aligned}
1 - \tau_y^{FB}(n) &= \frac{\psi'\left(\frac{y(n)}{w_n n}\right)\frac{1}{w_m m}}{u'(c)} = 1 \\
\implies \tau_y^{FB}(n) &= 0
\end{aligned}
$$

Similarly for manual-origin workers.

To prove the last equation, notice that

$$\begin{aligned}\int_{\underline{n}}^{\bar{n}} v(n,m)dG_z(n;z(m),m) &= \int_{\underline{n}}^{\bar{n}} \left[u(c) - \psi(l(n,m)\right] dG_z(n;z(m),m) \\ &= u(c)\int_{\underline{n}}^{\bar{n}} dG_z(n;z(m),m) - \int_{\underline{n}}^{\bar{n}} \psi(l(n,m)dG_z(n;z(m),m) \\ &= -\int_{\underline{n}}^{\bar{n}} \psi(l(n,m)dG_z(n;z(m),m)\end{aligned}$$

where the second line uses the fact that c is independent of n, and the third line holds because $\int_{\underline{n}}^{\bar{n}} dG_z(n;z(m),m) - \frac{\partial}{\partial z}\int_{\underline{n}}^{\bar{n}} dG(n;z(m),m) = 0$. Similarly, we have

$$\begin{aligned}\int_{\underline{n}}^{\bar{n}} T(n,m)dG_z(n;z(m),m) &= \int_{\underline{n}}^{\bar{n}} \left[y(n,m) - c(n,m)\right] dG_z(n;z(m),m) \\ &= -\int_{\underline{n}}^{\bar{n}} y(n,m)dG_z(n;z(m),m)\end{aligned}$$

Plug them in the FOC of z, we have that part (4) of the proposition holds as long as z is interior. □

B.2 Optimal Training Allocation

Proposition 6. *If we additionally assume $\psi(l) = \frac{l^{1+\sigma}}{1+\sigma}$ with $\sigma > 1$, and training technology follows a Pareto distribution*

$$1 - G(n|z,m) = \frac{zm^p}{n^\alpha}$$

where $\alpha > \frac{1+\sigma}{\sigma}$. Then for any interior solution z

1) $z(m) = 0$ for all m if $\lambda \geq (1+\sigma)^{\sigma^2-1}$ where λ is the multiplier on the resource constraint;

2) there exists some $\bar{p} > 0$ s.t. $z(m)$ increases in m iff $p > \bar{p}$;

3) retraining wedge $\tau_z^{FB}(m) > 0$, and $\tau_z^{FB}(m)$ increases in m iff $p > \bar{p}$.

Proof. We first proceed with a more general Pareto distribution of skills so that

$$1 - G(n|z,m) = \left(\frac{x(z,m)}{n}\right)^\alpha$$

where $x(z,m)$ is the location parameter of the distribution. Rewrite the FOC for z from Proposition 5:

$$-\phi'(z) - \lambda + (\lambda y^M - \psi^M)G_z(\theta(m)) + \int (\lambda y^N - \psi^N)dG_z(n) = 0$$

where, to ease notation, I suppress the argument as long as there is no confusion. Under power labor disutility, for a worker who uses skill n (skill m similarly)

$$\begin{aligned} y(n) &= \lambda^{\sigma}(wn)^{\frac{1+\sigma}{\sigma}} \\ l(n) &= \lambda^{\sigma}(wn)^{\frac{1}{\sigma}} \\ \psi(l(n)) &= \frac{\lambda^{\sigma(1+\sigma)}(wn)^{\frac{1+\sigma}{\sigma}}}{1+\sigma} \end{aligned}$$

Therefore,

$$\begin{aligned} \lambda y(n) - \psi(l(n)) &= \lambda^{\sigma+1}(wn)^{\frac{1+\sigma}{\sigma}} - \frac{\lambda^{\sigma(1+\sigma)}(wn)^{\frac{1+\sigma}{\sigma}}}{1+\sigma} \\ &= \left[\lambda^{\sigma+1} - \frac{\lambda^{\sigma(1+\sigma)}}{1+\sigma}\right](wn)^{\frac{1+\sigma}{\sigma}} \\ &:= \Lambda(wn)^{\frac{1+\sigma}{\sigma}} \end{aligned}$$

where I define $\Lambda = \lambda^{\sigma+1} - \frac{\lambda^{\sigma(1+\sigma)}}{1+\sigma}$. Using this result and the expression of the distribution function G, we have

$$(\lambda y^M - \psi^M)G_z(\theta(m)) = \Lambda(w_m m)^{\frac{1+\sigma}{\sigma}}\left(-\frac{\alpha x(m,z)^{\alpha-1}\frac{\partial x}{\partial z}}{\theta(m)^{\alpha}}\right)$$

and

$$\begin{aligned} \int(\lambda y^N - \psi^N)dG_z(n) &= \Lambda\int_{\theta(m)}^{\infty}(w_n n)^{\frac{1+\sigma}{\sigma}}dG_z(n) \\ &= \Lambda w_n^{\frac{1+\sigma}{\sigma}}\frac{\alpha^2 x(m,z)^{\alpha-1}\frac{\partial x}{\partial z}}{\frac{1+\sigma}{\sigma}-\alpha}\theta(m)^{\frac{1+\sigma}{\sigma}-\alpha} \\ &= \Lambda(w_n\theta(m))^{\frac{1+\sigma}{\sigma}}\frac{\alpha}{\alpha-\frac{1+\sigma}{\sigma}}\frac{\alpha x(m,z)^{\alpha-1}\frac{\partial x}{\partial z}}{\theta(m)^{\alpha}} \end{aligned}$$

where the condition

$$\alpha > \frac{1+\sigma}{\sigma}$$

is imposed to guarantee integrability. Combining the two pieces, the FOC is reduced to

$$\begin{aligned}
\phi'(z) + \lambda &= \Lambda(w_m m)^{\frac{1+\sigma}{\sigma}} \left(-\frac{\alpha x(m,z)^{\alpha-1} \frac{\partial x}{\partial z}}{\theta(m)^\alpha} \right) + \Lambda(w_n \theta(m))^{\frac{1+\sigma}{\sigma}} \frac{\alpha}{\alpha - \frac{1+\sigma}{\sigma}} \frac{\alpha x(m,z)^{\alpha-1} \frac{\partial x}{\partial z}}{\theta(m)^\alpha} \\
&= \frac{\Lambda \alpha}{\theta(m)^\alpha} \frac{\frac{1+\sigma}{\sigma}}{\alpha - \frac{1+\sigma}{\sigma}} (w_m m)^{\frac{1+\sigma}{\sigma}} x(m,z)^{\alpha-1} \frac{\partial x}{\partial z} \\
&= \frac{\frac{1+\sigma}{\sigma}}{\alpha - \frac{1+\sigma}{\sigma}} \Lambda \alpha w_n^\alpha (w_m m)^{\frac{1+\sigma}{\sigma} - \alpha} x(m,z)^{\alpha-1} \frac{\partial x}{\partial z}
\end{aligned}$$

Finally, set $x(m,z) = (zm^p)^{1/\alpha}$, the FOC becomes

$$\phi'(z) + \lambda = \frac{\frac{1+\sigma}{\sigma}}{\alpha - \frac{1+\sigma}{\sigma}} \Lambda w_n^\alpha (w_m m)^{\frac{1+\sigma}{\sigma} - \alpha} m^p$$

Since the LHS must be positive, this equation has an interior solution only if $\Lambda > 0$, which means

$$\lambda < (1+\sigma)^{\sigma^2 - 1}$$

When $\lambda \geq (1+\sigma)^{\sigma^2-1}$, there is only a corner solution where $z = 0$. Now assume that there is an interior solution. Because $\phi'' > 0$, there exists $\bar{p} = \alpha - \frac{1+\sigma}{\sigma}$ such that $z(m)$ is increasing in m if and only if $p > \bar{p}$.

We next turn to the retraining wedge in FB. Combining the definition of retraining wedge and the FOC for z, we get the retraining wedge in FB in the general case:

$$\begin{aligned}
(1 - \tau_z^{FB}(m)) \int_{\underline{n}}^{\bar{n}} u'(c(n,m)) g(n|z(m),m) dn f_m(m) &= \lambda \left[1 - \int_{\underline{n}}^{\bar{n}} T(m,n) dG_z(n|z(m),m) \right] \tilde{f}_m(m) \\
\implies 1 - \tau_z^{FB}(m) &= \frac{\lambda \left[1 - \int_{\underline{n}}^{\bar{n}} T(m,n) dG_z(n|z(m),m) \right] \tilde{f}_m(m)}{\int u'(c(n,m)) g(n|z(m),m) dn f_m(m)}
\end{aligned}$$

With the additional functional form assumptions, we can derive

$$\tau_z^{FB}(m) = \frac{\frac{1+\sigma}{\sigma}}{\alpha - \frac{1+\sigma}{\sigma}} \alpha w_n^\alpha (w_m m)^{\frac{1+\sigma}{\sigma} - \alpha} m^p$$

Therefore, the retraining wedge in FB is always positive, and it's increasing in m if and only if $p > \bar{p} = \alpha - \frac{1+\sigma}{\sigma}$, which completes the proof. □

C Constrained Optimum

C.1 Switching Threshold

Lemma 1. *(Occupation threshold rule) The planner uses a threshold rule to assign workers to the two sectors at the beginning of the second stage, given truthful report. Namely, for any $m \in [\underline{m}, \bar{m}]$, there exists a $\theta(m) \in [\underline{n}, \bar{n}]$ such that the switching set $S(m) = [\theta(m), \bar{n}]$.*

Proof. Suppose the threshold rule does not hold. Then, there exist workers 1 (n^1, m) and 2 (n^2, m) with $n^2 > n^1$ such that worker 1 is assigned to non-manual sector N, while worker 2 is assigned to the manual sector M. Note that worker 1 can mimic worker 2 and be assigned to sector M. In that case, they would have the same consumption and labor supply since they have the same m. IC condition then requires that worker 1 obtains a higher value in sector N than in M. Now, note that worker 2 can also mimic worker 1 and be assigned to sector N. Again in that case, they would have the same consumption and labor supply. But notice that worker 2 would have a strictly higher value than worker 1 since he has a higher n. This implies worker 2 could obtain a higher value by mimicking worker 1, which contradicts worker 2's IC condition. □

C.2 Envelope Conditions

Lemma 2. *(Envelope conditions) Applying envelope theorem, the IC condition for sector M workers in stage 1 is replaced with*

$$\frac{\partial U(m)}{\partial m} = \int_{n<\theta(m)} \psi' \left(\frac{y^M(n,m)}{w_m m} \right) \frac{y^M(n,m)}{w_m m^2} dG(n|z(m),m) + \int_n v(n,m) dG_m(n|z(m),m)$$

while that in stage 2 is replaced with

$$\frac{\partial v(n,m)}{\partial n} = \psi' \left(\frac{y^{MN}(n,m)}{w_n n} \right) \frac{y^{MN}(n,m)}{w_n n^2} \mathbb{I}(n > \theta(m))$$

and the IC condition for sector N workers is replaced with

$$\frac{\partial U^N(n)}{\partial n} = \psi' \left(\frac{y^N(n)}{w_n n} \right) \frac{y^N(n_n)}{w_n n^2}$$

Proof. Immediately from the original IC constraints. □

C.3 Sufficient Conditions

Lemma 3. *(Sufficiency for first-order approach). Given Assumption 1 and further assuming (i)* $\frac{\partial G(n|m,z)}{\partial m} < 0$*; (ii)* $\frac{\partial y^N(n)}{\partial n} > 0$, $\frac{\partial y^{MN}(n,m)}{\partial n} > 0$*, and* $\frac{\partial y^M(n,m)}{\partial m} > 0$*; (iii)* $\frac{\partial z(m)}{\partial m}$ *has the opposite sign as* $\frac{\partial^2 G(n|m,z)}{\partial z \partial m}$*; and (iv) switching threshold* $\theta(m)$ *is continuous almost everywhere, we have that an allocation that satisfies the envelope conditions in Lemma 2 is incentive compatible.*

Proof. I first prove the incentive compatibility for sector N workers. Consider a worker with skill n who reports $r_n = n'$. Given a deviation from his report, we have

$$\frac{\partial U^N(r_n; n)}{\partial r_n} = u'(c^N(r_n))\frac{dc^N(r_n)}{dr_n} - \psi'\left(\frac{y^N(r_n)}{w_n n}\right)\frac{1}{w_n n}\frac{dy^n(r_n)}{dr_n}$$

For a worker with skill n' who also reports $r_n = n'$, the envelope condition implies

$$0 = u'(c^N(r_n))\frac{dc^N(r_n)}{dr_n} - \psi'\left(\frac{y^N(r_n)}{w_n n'}\right)\frac{1}{w_n n'}\frac{dy^n(r_n)}{dr_n}$$

Subtracting the two, we have

$$\frac{\partial U^N(r_n; n)}{\partial r_n} = -\frac{dy^n(r_n)}{dr_n}\frac{1}{w_n}\left[\frac{1}{n}\psi'\left(\frac{y^N(r_n)}{w_n n}\right) - \frac{1}{n'}\psi'\left(\frac{y^N(r_n)}{w_n r_n}\right)\right]$$

It is sufficient to show that $\frac{\partial U^N(r_n;n)}{\partial r_n}$ has the same sign as $n - r_n$. This is immediate given (ii) and that $\psi'' > 0$.

Given the report in the first stage, the incentive compatibility for sector a M worker in the second stage is similar if he draws a skill $n > \theta(m)$. Note he is indifferent of reporting any $n < \theta(m)$ if he draws one smaller than $\theta(m)$. In that case, I assume he truthfully reports his skill draw. The knife edge case of $n = \theta(m)$ has probability zero, and therefore, it doesn't affect the discussion here.

Finally, I discuss incentive compatibility for sector M workers in the first stage. Consider a worker with skill m who reports $r_m = m'$. Given a deviation from his report, we have

$$\begin{aligned}\frac{\partial U^M(r_m;m)}{\partial r_m} =& -\phi'(z(r_m))\frac{dz(r_m)}{dr_m} \\ &+ \int_n \left[\frac{\partial v(r_m;n,m)}{\partial r_m} g(n|z(r_m),m) + v(n;z(r_m),m)\frac{\partial g(n|z(r_m),m)}{\partial z(r_m)}\frac{dz(r_m)}{dr_m}\right] dn\end{aligned}$$

While for a worker with skill m' who also reports $r_m = m'$, the envelope condition is

$$\begin{aligned}0 =& -\phi'(z(r_m))\frac{dz(r_m)}{dr_m} \\ &+ \int_n \left[\frac{\partial v(r_m;n,m')}{\partial r_m} g(n|z(r_m),m') + v(r_m;n,m')\frac{\partial g(n|z(r_m),m')}{\partial z(r_m)}\frac{dz(r_m)}{dr_m}\right] dn\end{aligned}$$

Subtracting, we have

$$\begin{aligned}\frac{\partial U^M(r_m;m)}{\partial r_m} =& \int_n \left[\frac{\partial v(r_m;n,m)}{\partial r_m} g(n|z(r_m),m) - \frac{\partial v(r_m;n,m')}{\partial r_m} g(n|z(r_m),m')\right] dn \\ &+ \frac{dz(r_m)}{dr_m}\int_n \left[v(r_m;n,m)\frac{\partial g(n|z(r_m),m)}{\partial z(r_m)} - v(r_m;n,m')\frac{\partial g(n|z(r_m),m')}{\partial z(r_m)}\right] dn \\ =& \int_n \left[\frac{\partial v(r_m;n,m)}{\partial r_m} g(n|z(r_m),m) - \frac{\partial v(r_m;n,m')}{\partial r_m} g(n|z(r_m),m')\right] dn_A \\ &- \frac{dz(r_m)}{dr_m}\int_{n_A} \left[\frac{\partial v(r_m;n,m)}{\partial n}\frac{\partial G(n|z(r_m),m)}{\partial z(r_m)} - \frac{\partial v(r_m;n,m')}{\partial n}\frac{\partial G(n|z(r_m),m')}{\partial z(r_m)}\right] dn_A\end{aligned}$$

where the second equality uses integration by parts. It is sufficient to show that $\frac{\partial U^M(r_m;m)}{\partial r_m}$ has the same sign as $m - r_m$. Consider $m > m'$, then we want to show $\frac{\partial U^M(r_m;m)}{\partial r_m} > 0$. The first integral on the RHS is positive if $\frac{\partial v(r_m;n,m)}{\partial r_m}$ is increasing in n as $g(n|z(r_m),m)$ dominates $g(n|z(r_m),m')$ in FOSD sense by assumption (i). Applying envelope theorem to a worker's reporting problem in the second stage, one can see that $\frac{\partial^2 v(r_m;n,m)}{\partial n \partial r_m}$ has the same sign as $\frac{\partial y^M(n,m)}{\partial m}$, which is positive according to (ii). Here, I also use assumption (iv) so that a marginal change in the reporting strategy has only a marginal effect on the threshold, which cannot lead to any abrupt change to the value except on a set with measure zero, whose effect can be ignored. Finally, for the second integral on the RHS, by applying envelope theorem again to the worker's reporting problem in the second stage, we have

$$\frac{\partial v(r_m;n,m)}{\partial n} = \frac{\partial v(r_m;n,m')}{\partial n}$$

Intuitively, given the same report r_m and the same skill draw n, the difference in the underlying m and m' won't change the value except possibly at the threshold. But the probability of such scenarios is zero. Hence, (iii) guarantees it is positive. □

The set of sufficiency conditions contains some requirements on the endogenous objects. Therefore, one needs to check them ex post after solving the model.

C.4 Labor Wedge

Proposition 7. *Denoted by $\tau^{MN}(n,m)$ the optimal labor wedge for workers who switch from M to N, $\tau^{M}(n,m)$ the workers who stay in M, and $\tau^{N}(n,m)$ original N workers. Then,*

$$\frac{\tau_y^{MN}(n,m)}{1-\tau_y^{MN}(n,m)} = \frac{\eta(n,m)u'(c^N(n,m))}{\lambda g(n|z(m),m)\tilde{f}(m)n}\frac{1+\varepsilon_u(n,m)}{\varepsilon_c(n,m)}$$
$$\frac{\tau_y^{M}(n,m)}{1-\tau_y(n,m)} = \frac{\mu(m)u'(c(n,m))}{\lambda g(n|z(m),m)\tilde{f}_m(m)m}\frac{1+\varepsilon^u(n,m)}{\varepsilon^c(n,m)}$$
$$\frac{\tau_y^{N}}{1-\tau_y^{N}} = \frac{\xi(n)u'(c(n))}{\lambda\tilde{f}_n(n)n}\frac{1+\varepsilon^u(n)}{\varepsilon^c(n)}$$

where ε_u (ε_c) is the uncompensated (compensated) labor supply elasticity, λ is the multiplier on the government budget constraint, and

$$\eta(n,m) = \lambda\tilde{f}(m)G(n|z(m),m)\left(\mathbb{E}^{\tilde{n}}\left[\frac{1}{u'(c(\tilde{n},m))}\middle|\tilde{n}>n\right] - \mathbb{E}^{\tilde{n}}\left[\frac{1}{u'(c(\tilde{n},m))}\right]\right) - \mu(m)\frac{\partial G(n|z(m),m)}{\partial m}$$
$$\mu(m) = \int_m^{\bar{m}}\left(\lambda\int_n\frac{1}{u'(c(n,m))}g(n|z(m),m)dn - \frac{f_m(\tilde{m})}{\tilde{f}_m(\tilde{m})}\right)d\tilde{F}_m(\tilde{m})$$
$$\xi(n) = \int_n^{\bar{n}}\left(\frac{\lambda}{u'(c^A(\tilde{n}_a))} - \frac{f_n(\tilde{n})}{\tilde{f}_n(\tilde{n})}\right)d\tilde{F}_n(\tilde{n})$$

Proof. Let $\mu(m)$ and $\eta(n,m)$ denote the multiplier on the first- and second-period local IC constraints for manual-origin workers, respectively, and $\xi(n)$ denote the multiplier on the local IC constraint for incumbent non-manual workers. To ease notation, I suppress arguments of some functions as long as this does not cause confusion. I first further simplify the local IC constraints by integration by parts. In particular,

$$\int_m \mu(m)\left[\frac{\partial U(m)}{\partial m} - \int_{\underline{n}}^{\theta(m)} \psi'\left(\frac{y^M}{w_m m}\right)\frac{y^M}{w_m m^2} dG(n|z(m),m) + \int_n v(n,m)\frac{\partial g(n|z(m),m)}{\partial m} dn\right] dm$$
$$= -\int_m \mu'(m)\left[-\phi(z(m)) + \int_n v(n,m) dG(n|z(m),m)\right] dm$$
$$-\int_m \mu(m)\left[\int_{\underline{n}}^{\theta(m)} \psi'\left(\frac{y^M}{w_m m}\right)\frac{y^M}{w_m m^2} dG(n|z(m),m) + \int_n v(n,m)\frac{\partial g(n|z(m),m)}{\partial m} dn\right] dm$$

where we use the transversality condition that $\mu(\underline{m}) = \mu(\bar{m}) = 0$. Similarly,

$$\int_m \int_n \eta(n,m)\left[\frac{\partial v(n,m)}{\partial n} - \psi'\left(\frac{y^{MN}}{w_n n}\right)\frac{y^{MN}}{w_n n^2}\mathbb{I}(n > \theta(m))\right] dndm$$
$$= -\int_m \int_n \left[\frac{\partial \eta(n,m)}{\partial n} v(n,m) + \eta(n,m)\psi'\left(\frac{y^{MN}}{w_n n}\right)\frac{y^{MN}}{w_n n^2}\mathbb{I}(n > \theta(m))\right] dndm$$

where we use $\eta(\underline{n}, m) = \eta(\bar{n}, m) = 0$. For incumbent N workers,

$$\int_n \xi(n)\left[\frac{\partial U^N(n)}{\partial n} - \psi'\left(\frac{y^N(n)}{w_n n}\right)\frac{y^N(n)}{w_n n^2}\right] dn = -\int_n \left[\xi'(n)U^N(n) + \xi(n)\psi'\left(\frac{y^N(n)}{w_n n}\right)\frac{y^N(n)}{w_n n^2}\right] dn$$

where we use $\xi(\underline{n}) = \xi(\bar{n}) = 0$. In addition, follow the tradition of mechanism design problems, we use value U^N and $v(n,m)$ as choice variables to replace consumption c. The Lagrangian can thus be written as

$$
\begin{aligned}
\max_{v,U^N,y^N,y^{MN},y^M,z,\theta} & \int_m -\phi(z(m))dF_m(m) \\
& + \int_n \int_m v(n,m)dG(n|z(m),m)dF_m(m) \\
& + \int_n U^N(n)dF_n(n) \\
& + \lambda E - \lambda \int_m z(m)d\tilde{F}_m(m) \\
& - \lambda \int_n \left[u^{-1}\left(U^N(n) + \psi\left(\frac{y^N(n)}{w_n n}\right)\right) - y^N(n)\right] d\tilde{F}_n(n) \\
& - \lambda \int_m \int_{\underline{n}}^{\theta(m)} \left[u^{-1}\left(v(n,m) + \psi\left(\frac{y^M(n,m)}{w_m m}\right)\right) - y^M(n,m)\right] dG(n|z(m),m)d\tilde{F}_m(m) \\
& - \lambda \int_m \int_{\theta(m)}^{\bar{n}} \left[u^{-1}\left(v(n,m) + \psi\left(\frac{y^{MN}(n,m)}{w_n n}\right)\right) - y^{MN}(n,m)\right] dG(n|z(m),m)d\tilde{F}_m(m) \\
& - \int_n \left[\xi'(n)U^N(n) + \xi(n)\psi'\left(\frac{y^N(n)}{w_n n}\right)\frac{y^N(n)}{w_n n^2}\right] dn \\
& - \int_m \mu'(m)\left[-\phi(z(m)) + \int_n v(n,m)dG(n|z(m),m)\right] dm \\
& - \int_m \mu(m)\left[\int_{\underline{n}}^{\theta(m)} \psi'\left(\frac{y^M}{w_m m}\right)\frac{y^M}{w_m m^2}dG(n|z(m),m) + \int_n v(n,m)\frac{\partial g(n|z(m),m)}{\partial m}dn\right] dm \\
& - \int_m \int_n \left[\frac{\partial \eta(n,m)}{\partial n}v(n,m) + \eta(n,m)\psi'\left(\frac{y^{MN}}{w_n n}\right)\frac{y^{MN}}{w_n n^2}\mathbb{I}(n > \theta(m))\right] dndm
\end{aligned}
$$

The FOC for $U^N(n)$ is

$$
f_n(n) - \lambda \frac{1}{u'(c^N(n))}\tilde{f}_n(n) - \xi'(n) = 0 \tag{7}
$$

for $y^N(n)$

$$
-\lambda\left[\frac{1}{u'(c^N(n))}\psi'\left(\frac{y^N(n)}{w_n n}\right)\frac{1}{w_n n} - 1\right]\tilde{f}_n(n) - \xi(n)\left[\psi''\left(\frac{y^N(n)}{w_n n}\right)\frac{y^N(n)}{w_n^2 n^3} + \psi'\left(\frac{y^N(n)}{w_n n}\right)\frac{1}{w_n n^2}\right] = 0 \tag{8}
$$

for $v(n,m)$

$$
\begin{aligned}
0 = & g(n|z(m),m)f_m(m) - \lambda\frac{1}{u'(c(n,m))}g(n|z(m),m)\tilde{f}_m(m) - \mu'(m)g(n|z(m),m) \\
& - \mu(m)\frac{\partial g(n|z(m),m)}{\partial m} - \frac{\partial \eta(n,m)}{\partial n}
\end{aligned} \tag{9}
$$

for $y(n,m)$ when $n > \theta(m)$ (i.e., $y^{MN}(n,m)$)

$$
\begin{aligned}
0 = & -\lambda \left(\frac{1}{u'(c(n,m))} \psi' \left(\frac{y(n,m)}{w_n n} \right) \frac{1}{w_n n} - 1 \right) g(n|z(m),m) \tilde{f}_m(m) \\
& - \eta(n,m) \left(\psi'' \left(\frac{y(n,m)}{w_n n} \right) \frac{y(n,m)}{w_n^2 n^3} + \psi' \left(\frac{y(n,m)}{w_n n} \right) \frac{1}{w_n n^2} \right)
\end{aligned} \tag{10}
$$

for $y(n,m)$ when $n < \theta(m)$ (i.e., $y^{M}(n,m)$)

$$
\begin{aligned}
0 = & -\lambda \left(\frac{1}{u'(c(n,m))} \psi' \left(\frac{y(n,m)}{w_m m} \right) \frac{1}{w_m m} - 1 \right) g(n|z(m),m) \tilde{f}_m(m) \\
& - \mu(m) \left(\psi'' \left(\frac{y(n,m)}{w_m m} \right) \frac{y(n,m)}{w_m^2 m^3} + \psi' \left(\frac{y(n,m)}{w_m m} \right) \frac{1}{w_m m^2} \right)
\end{aligned} \tag{11}
$$

for $z(m)$

$$
\begin{aligned}
& \left[-\phi'(z(m)) + \int_n v(n,m) \frac{\partial g(n|z(m),m)}{\partial z(m)} dn \right] f_m(m) - \lambda \tilde{f}_m(m) \left[1 + \int_{\underline{n}}^{\bar{n}} (c(n,m) - y(n,m)) \frac{\partial g(n|z(m),m)}{\partial z(m)} dn \right] \\
& - \mu'(m) \left[-\phi'(z(m)) + \int_n v(n,m) \frac{\partial g(n|z(m),m)}{\partial z(m)} dn \right] \\
& - \mu(m) \left(\int_{\underline{n}}^{\theta(m)} \psi' \left(\frac{y(n,m)}{w_m m} \right) \frac{y(n,m)}{w_m m^2} \frac{\partial g(n|z(m),m)}{\partial z(m)} dn + \int_n v(n,m) \frac{\partial^2 g(n|z(m),m)}{\partial z(m) \partial m} dn \right) \\
= & 0
\end{aligned} \tag{12}
$$

and finally for $\theta(m)$

$$
\begin{aligned}
& -\lambda \left[c^M(\theta(m),m) - y^M(\theta(m),m) \right] g(\theta(m)|z(m),m) \tilde{f}_m(m) \\
& + \lambda \left[c^{MN}(\theta(m),m) - y^{MN}(\theta(m),m) \right] g(\theta(m)|z(m),m) \tilde{f}(n_b) \\
& - \mu(m) \psi' \left(\frac{y^M(\theta(m),m)}{w_m m} \right) \frac{y^M(\theta(m),m)}{w_m m^2} g(\theta(m)|z(m),m) \\
& + \eta(\theta(m),m) \psi' \left(\frac{y^{MN}(\theta(m),m)}{w_n \theta(m)} \right) \frac{y^{MN}(\theta(m),m)}{w_n \theta(m)^2} \\
= & 0
\end{aligned} \tag{13}
$$

Labor wedge for incumbent N workers

First look at the labor wedge for incumbent N workers. From (8) and the definition of labor wedge, we immediately have that

$$\frac{\tau_y^N}{1-\tau_y^N} = \frac{\xi(n)u'(c(n))}{\lambda \tilde{f}_n(n)n} \frac{\psi'' \frac{y(n)}{w_n^2 n^2} + \psi' \frac{1}{w_n n}}{\psi' \frac{1}{w_n n}} = \frac{\xi(n)u'(c(n))}{\lambda \tilde{f}_n(n)n} \frac{1+\varepsilon^u(n)}{\varepsilon^c(n)}$$

where $\varepsilon^u(n)$ and $\varepsilon^c(n)$ denote the uncompensated and compensated labor supply elasticity at the labor supply demanded by the planner for skill n workers. Integrating (7) from n to $\bar{n}$ and use the boundary condition, we have that the multiplier $\xi(n)$ is given by

$$\xi(n) = \int_n^{\bar{n}} \left(\frac{\lambda}{u'(c^A(\tilde{n}_a))} - \frac{f_n(\tilde{n})}{\tilde{f}_n(\tilde{n})} \right) d\tilde{F}_n(\tilde{n})$$

Labor wedge for M-origin workers

Dividing (11) and (10) on both sides by $\frac{1}{u'(c(n,m))}\psi'\left(\frac{y(n,m)}{w_n n}\right)\frac{1}{w_n n} = 1-\tau_y(n,m)$, we have

$$\lambda g(n|z(m),m)\tilde{f}_m(m)\frac{\tau_y(n,m)}{1-\tau_y(n,m)} = \mu(m)u'(c(n,m))\frac{\left(\psi''\left(\frac{y(n,m)}{w_m m}\right)\frac{y(n,m)}{w_m^2 m^3} + \psi'\left(\frac{y(n,m)}{w_m m}\right)\frac{1}{w_m m^2}\right)}{\psi'\left(\frac{y(n,m)}{w_n n}\right)\frac{1}{w_n n}}$$

and

$$\lambda g(n|z(m),m)\tilde{f}_m(m)\frac{\tau_y(n,m)}{1-\tau_y(n,m)} = \eta(n,m)u'(c(n,m))\frac{\left(\psi''\left(\frac{y(n,m)}{w_m m}\right)\frac{y(n,m)}{w_m^2 m^3} + \psi'\left(\frac{y(n,m)}{w_m m}\right)\frac{1}{w_m m^2}\right)}{\psi'\left(\frac{y(n,m)}{w_n n}\right)\frac{1}{w_n n}}$$

which leads to

$$\frac{\tau_y(n,m)}{1-\tau_y(n,m)} = \frac{\mu(m)u'(c(n,m))}{\lambda g(n|z(m),m)\tilde{f}_m(m)m}\frac{1+\varepsilon^u(n,m)}{\varepsilon^c(n,m)}$$

for M stayers, and

$$\frac{\tau_y(n,m)}{1-\tau_y(n,m)} = \frac{\eta(n,m)u'(c(n,m))}{\lambda g(n|z(m),m)\tilde{f}_m(m)m}\frac{1+\varepsilon^u(n,m)}{\varepsilon^c(n,m)}$$

for M to N switchers, where ε^u and ε^c are defined similarly as before. To get the expression for $\mu(m)$, by integrating (9) over the support of n, we have

$$f_m(m) - \lambda\tilde{f}_m(m)\int_n \frac{1}{u'(c(n,m))}g(n|z(m),m)dn - \mu'(m) = 0 \tag{14}$$

Then, integrating again from m to $\bar{m}$, we have

$$\mu(m) = \int_m^{\bar{m}} \left(\lambda \int_n \frac{1}{u'(c(n,m))} g(n|z(m),m) dn - \frac{f_m(\tilde{m})}{\tilde{f}_m(\tilde{m})} \right) d\tilde{F}_m(\tilde{m})$$

To get $\eta(n,m)$, plug $\mu'(m)$ back into (9), we have

$$\frac{\partial \eta(n,m)}{\partial n} = \lambda \tilde{f}_m(m) g(n|z(m),m) \left[\int_n \frac{1}{u'(c(n,m))} g(n|z(m),m) dn_a - \frac{1}{u'(c(n,m))} \right] - -\mu(m) \frac{\partial g(n|z(m),m)}{\partial m}$$

Now integrating from n to $\bar{n}$, we have

$$\begin{aligned}
\eta(n,m) =& \lambda \tilde{f}_m(m) \int_n^{\bar{n}} \left[\frac{1}{u'(c(\tilde{n},m))} - \int_{\hat{n}} \frac{1}{u'(c(\hat{n},m))} g(\hat{n}|z(m),m) dn_a \right] g(\tilde{n}|z(m),m) d\tilde{n} \\
&+ \mu(m) \int_n^{\bar{n}} \frac{\partial g(\tilde{n}|z(m),m)}{\partial m} d\tilde{n} \\
=& \lambda \tilde{f}_m(m) \bar{G}(n|z(m),m) \left(\mathbb{E} \left[\frac{1}{u'(c(\tilde{n},m))} \middle| \tilde{n} > n \right] - \mathbb{E} \left[\frac{1}{u'(c(\tilde{n},m))} \right] \right) + \mu(m) \frac{\partial \bar{G}(n|z(m),m)}{\partial m}
\end{aligned}$$

where $\bar{G} := 1 - G$. □

C.5 Retraining Wedge

Proposition 8. *(Optimal retraining wedge) Denote by $\tau^{SB}(m)$ the optimal retraining wedge in the constrained problem. Then i)*

$$1 - \tau_z^{SB}(m) = \mathcal{C}(m)(1 - \tau_z^{FB}(m)) + \mathcal{I}_1(m) + \mathcal{I}_2(m)$$

where $\tau_z^{FB}(m)$ is the retraining wedge in FB, and $\mathcal{C}(m) = f_m(m) \left(\lambda \tilde{f}_m(m) \mathbb{E}^{\tilde{n}} \left[\frac{1}{u'(c(\tilde{n},m))} \right] \right)^{-1}$;

ii) for all m such that $\mu(m) > 0$, $\mathcal{I}_1(m) \leq 0$ and $\mathcal{I}_2(m) > 0$ if $G_{zm} < 0$.

Proof. The definition of retraining wedge implies that

$$-\phi'(z) + \int_n v(n,m,z) d \frac{\partial G(n|z,m)}{\partial z} = (1 - \tau_z^{SB}(m)) \int u' g dn$$

Recall the retraining wedge in FB satisfies

$$\lambda \tilde{f}_m(m)\left[1+\int_{\underline{n}}^{\bar{n}}(c(n,m)-y(n,m))\frac{\partial g(n|z(m),m)}{\partial z}dn\right]=(1-\tau_z^{FB}(m))f_m(m)\int u'gdn$$

Plugging them in equation (12) , i.e., the FOC for z, we have

$$(1-\tau_z^{SB}(m))f_m(m)\int u'gdn=(1-\tau_z^{FB}(m))f_m(m)\int u'gdn+\mu'(m)(1-\tau_z^{SB}(m))\int u'gdn+\tilde{\mathcal{I}}_1(m)+\tilde{\mathcal{I}}_2(m)$$

where

$$\tilde{\mathcal{I}}_1(m):=\mu(m)\int_{\underline{n}}^{\theta(m)}\psi'\left(\frac{y(n,m)}{w_m m}\right)\frac{y(n,m)}{w_m m^2}\frac{\partial g(n|z(m),m)}{\partial z}dn$$

and

$$\tilde{\mathcal{I}}_2(m):=\mu(m)\int_n v(n,m)\frac{\partial^2 g(n|z(m),m)}{\partial z\partial m}dn$$

Together,

$$\begin{aligned}\tilde{\mathcal{I}}_1(m)+\tilde{\mathcal{I}}_2(m)&=\mu(m)\frac{\partial}{\partial z}\left[\int_{\underline{n}}^{\theta(m)}\psi'\frac{l}{n}gdn+\int_n vg_m dn\right]\\&=\mu(m)\frac{\partial}{\partial z}\left(\frac{\partial U^M(m)}{\partial m}\right)\end{aligned}$$

measures changes in incentive cost to induce truthful report after extra training dz at skill m in the first stage. Rearranging leads to

$$1-\tau_z^{SB}(m)=\left(\frac{f_m(m)-\mu'(m)}{f_m(m)}\right)^{-1}(1-\tau_z^{FB}(m))+\frac{\tilde{\mathcal{I}}(m)}{[f_m(m)-\mu'(m)]\int u'gdn}$$

Using equation (14) to replace $\mu'(m)$, we have

$$1-\tau_z^{SB}(m)=\left(\frac{\tilde{f}_m(m)\int\frac{g}{u'}dn}{f_m(m)}\right)^{-1}(1-\tau_z^{FB}(m))+\frac{\tilde{\mathcal{I}}(m)}{\tilde{f}_m(m)\int\frac{g}{u'}dn\int u'gdn}$$

Redefine $\mathcal{I}_i(m)$ by dividing $\tilde{\mathcal{I}}_i(m)$ by $\tilde{f}_m(m)\int\frac{g}{u'}dn\int u'gdn, i=1,2,$ gives the formula in the proposition. Because $\mathcal{I}_i(m)$ has the same sign as $\tilde{\mathcal{I}}_i(m), i=1,2,$ we only need to discuss the sign of the latter. Since $\mu(m)$ measures the local shadow cost of compensating for truthful report at m, it is positive as long as as long as there are information rents at and above m. Note

$$\begin{aligned}
\int_{\underline{n}}^{\theta(m)} \psi'\left(\frac{y}{w_m m}\right)\frac{y}{w_m m^2} g_z dn &= \frac{\partial}{\partial z}\left(\int_{\underline{n}}^{\theta(m)} \psi' \frac{l^m}{m} g dn\right) \\
&= \frac{\partial}{\partial z}\left[\psi'\frac{l^m}{m}G\bigg|_{\underline{n}}^{\theta(m)} - \int_{\underline{n}}^{\theta(m)} G\frac{d}{dn}\left(\psi'\frac{l^m}{m}\right)dn\right] \\
&= \psi'\frac{l^m}{m}G_z\bigg|_{\theta(m)} - \int_{\underline{n}}^{\theta(m)} G_z\frac{d}{dn}\left(\psi'\frac{l^m}{m}\right)dn \\
&= \psi'\frac{l^m}{m}G_z\bigg|_{\theta(m)} \\
&\leq 0
\end{aligned}$$

where the first equality changes the order of integration and differentiation, the second uses integration by parts, and the last equality holds because when a worker stays in M, the new draw of n is inactive, and the allocation to him does not respond to any change in n. Finally, the last inequality holds because we assume $G_z \leq 0$. Therefore, we have $\mathcal{I}_1(m) \leq 0$.

As for $\mathcal{I}_2(m)$, we have

$$\begin{aligned}
\int_n v(n,m)\frac{\partial^2 g(n|z(m),m)}{\partial z(m)\partial m}dn &= \int_n v(n,m)dG_{zm}(n|z(m),m) \\
&= v(n,m)G_{zm}(n|z(m),m)\bigg|_{\underline{n}}^{\bar{n}} - \int_n v_n(n,m)G_{zm}(n|z(m),m)dn \\
&= -\int_n v_n(n,m)G_{zm}(n|z(m),m)dn
\end{aligned}$$

where the last equality holds because $G(\underline{n}|z(m),m) \equiv 0$ and $G(\bar{n}|z(m),m) \equiv 1$, so the cross partial is zero. Now, notice that

$$v_n(n,m) \geq 0$$

which is necessary for global IC constraint to hold. Hence, this $\mathcal{I}_2(m) > 0$ if $G_{zm} < 0$. This completes the proof. □

C.6 Occupational Choice

Proposition 9. *(Switching threshold and comparative advantage) Suppose the planner can observe n in the second stage (but not m in the first stage), and worker's preference is $u(c,l) = c - \frac{l^{1+\sigma}}{1+\sigma}$. Then, when*

$\mu(m) > 0$, *we have*

$$\theta(m) < \frac{w_m m}{w_n}$$

Proof. Below, I denote $x^j = x^j(\theta(m), m)$ where $x \in \{c, y, l\}$, $j \in \{M, MN\}$ as the allocation to a marginal worker at M and at N. The assumption that n is observable amounts to that the second-stage multiplier $\eta = 0$. The FOC for $\theta(m)$ reduces to (again, I suppress the arguments of some functions as long as there is no confusion)

$$\lambda T^M = \lambda T^{MN} + \tilde{\mu}(m)\psi' \frac{l^M}{m}$$

where $T^j := y^j - c^j$, and

$$\tilde{\mu}(m) := \frac{\mu(m)}{\tilde{f}_m(m)}$$

is the net social cost of redistribution/insurance for an individual m worker. When $\mu(m) > 0$, we have

$$T^{MN} < T^M \iff y^{MN} - c^{MN} < y^M - c^M$$

Note the marginal worker must have the same value in the two sectors, implying

$$c^M - \psi\left(\frac{y^M}{w_m \theta(m)}\right) = c^{MN} - \psi\left(\frac{y^{MN}}{w_n n}\right)$$

Adding above the two equations, we have

$$y^{MN} - \psi\left(\frac{y^{MN}}{w_n \theta(m)}\right) < y^M - \psi\left(\frac{y^M}{w_m m}\right) \tag{15}$$

Define the function $\varphi(y) = y - \psi\left(\frac{y}{wn}\right)$. Then,

$$\varphi'(y) = 1 - \frac{y^\sigma}{(wn)^{1+\sigma}} = 1 - \frac{l^\sigma}{wn}$$

Note in the first-best case under the assumed preference, we have

$$l = \lambda^{\frac{1}{\sigma}}(wn)^{\frac{1}{\sigma}}$$

Now, since $\mu(m) > 0$, labor wedge is positive. As a result, labor supply is distorted downward, which implies

$$l < \lambda^{\frac{1}{\sigma}} (wn)^{\frac{1}{\sigma}}$$

Hence,

$$\varphi'(y) > 1 - \lambda = 0$$

where we use the result that under quasilinear preference, $\lambda = 1$. This implies $\varphi(y)$ is a increasing function. Therefore, if $y^{MN} > y^M$, from (15) we must have

$$w_n \theta(m) < w_m m$$

which is what we want to prove. On the other hand, if $y^{MN} \leq y^M$, we go back to the FOC of y. From (11) and (10), we have, under the assumed preference,

$$0 = \lambda \left(\psi' \left(\frac{y^{MN}}{w_n \theta(m)} \right) \frac{1}{w_n \theta(m)} - 1 \right) g(\theta(m)|z(m), m) \tilde{f}_m(m)$$

and

$$\begin{aligned} 0 =& \lambda \left(\psi' \left(\frac{y^M}{w_m m} \right) \frac{1}{w_m m} - 1 \right) g(\theta(m)|z(m), m) \tilde{f}_m(m) \\ &+ \mu(m) \left(\psi'' \left(\frac{y(n, m)}{w_m m} \right) \frac{y(n, m)}{w_m^2 m^3} + \psi' \left(\frac{y(n, m)}{w_m m} \right) \frac{1}{w_m m^2} \right) \end{aligned}$$

Since ψ is convex, the second term in the LHS of the above equation is positive. Therefore, we have

$$\psi' \left(\frac{y^M}{w_m m} \right) \frac{1}{w_m m} < \psi' \left(\frac{y^{MN}}{w_n \theta(m)} \right) \frac{1}{w_n \theta(m)}$$

Because we assume $y^{MN} \leq y^M$, this implies

$$w_m m \geq w_n \theta(m)$$

Combining the two parts completes the proof. □

C.7 Sectoral Redistribution

Proposition 10. *(Redistribution between the two sectors) The multiplier λ on the resource constraint satisfies*

$$\frac{F_n(\bar{n})}{\int \frac{1}{u'(c^N(n))}\tilde{f}_n(n)dn} = \lambda = \frac{F_m(\bar{m})}{\int_m \tilde{f}(m)\int_n \frac{1}{u'(c(n,m))}g(n|z(m),m)dndm}$$

Proof. Integrating (7) from $\underline{n}$ and $\bar{n}$ and using the boundary condition that $\xi(\underline{n}) = \xi(\bar{n}) = 0$, we have

$$\lambda = \frac{F_n(\bar{n})}{\int \frac{1}{u'(c^N(n))}\tilde{f}_n(n)dn}$$

Integrating (14) from $\underline{m}$ and $\bar{m}$ and using the boundary condition that $\mu(\underline{m}) = \mu(\bar{m}) = 0$, we have

$$\lambda = \frac{F_m(\bar{m})}{\int_m \tilde{f}(m)\int_n \frac{1}{u'(c(n,m))}g(n|z(m),m)dndm}$$

which completes the proof. □

C.8 Implementation

Proposition 11. *(Implementation by a history-dependent tax schedule) Any incentive compatible allocation can be implemented by a retraining schedule $\mathcal{Z} : \mathbb{R}_+ \to \mathbb{R}_+, w_- \mapsto z$ and a history-dependent nonlinear tax schedule $\mathcal{T} : \mathbb{R}_+ \times \mathbb{R}_+ \times \{N, M, MN\} \to \mathbb{R}$, $(w_-, y, h) \mapsto T$, where w_- is the worker's past wage, y is a worker's income, h is his occupational history, and z and T are his retraining level and tax payment, respectively.*

Proof. The implementation contains two steps. Let $z^*(m), h^*(n,m), y^*(n,m)$ denote the optimal training level, optimal sectoral choice, and optimal income in the constrained problem. In the first step, workers from sector M are asked to report their past wage level w_-, and then they get a voucher which allows them to get training z such that

$$\mathcal{Z}(w_-) = z^*\left(\frac{w_-}{w_m}\right)$$

In the second step, based on their reported past wage w_-, they are faced with the tax schedule $\mathcal{T}(w_-, y, h)$ defined below. For sector M-origin workers, i.e., $\mathcal{S} \in \{M, MN\}$, for a (y, h) allocation that is on the equilibrium path, suppose skill n leads to this allocation. Then, define

$$\mathcal{T}(w_-, y, h) = y^*\left(n, \frac{w_-}{w_m}\right) - c^*\left(n, \frac{w_-}{w_m}\right)$$

For a (y, h) allocation that is not on the equilibrium path, simply assume $\mathcal{T}(w_-, y, h)$ is extremely large so that it will not be chosen in any case. Then, M-origin workers choose their preferred sector. After that, along with incumbent N workers who face the tax schedule

$$\mathcal{T}(y^*(n)) = y^*(n) - c^*(n)$$

on the equilibrium path and extremely large if not, they choose their labor supply subject to their individual tax schedule.

Since the utility function is separable in c and l, which implies the single crossing condition is satisfied, given retraining history and occupational history in the labor supply stage, the taxation principle immediately implies that $(c(n, m), y(n, m))$ solves

$$\max_{c,y} u(c) - \psi\left(\frac{y}{w_s s}\right) \; s.t. \; c + \mathcal{T}(w_-, y, h) \leq y$$

Because the allocation is incentive compatible, facing the menu of choices, the worker would choose $\mathcal{S}$ in the occupational choice stage and report w_- before the retraining stage, as they would truthfully report their skills in the direct mechanism. □

Given a $(\mathcal{Z}, \mathcal{T})$ combination that implements an incentive compatible allocation, one can decompose the tax schedule into different components. For example, one can split it into a history-independent tax schedule and a history-contingent retraining loan. The latter resembles many real-world instruments like income-contingent student loans. This decomposition is formally stated below.

Corollary 1. *(Decomposing the tax schedule) Given $(\mathcal{Z}, \mathcal{T})$ that implements an incentive compatible allocation. We can decompose $\mathcal{T}(w_-, y, h)$ as any tax schedule $T(y)$ that only depends on income, and a* *history-contingent retraining subsidy $\mathcal{R}(w_-, y, h)$ such that*

$$\mathcal{T}(w_-, y, h) = T(y) + \mathcal{Z}(w_-) - \mathcal{R}(w_-, y, h)$$

where $\mathcal{Z}(w_-)$ is both the normalized training monetary cost, which is the same as the training level.

Therefore, one can think $T(y)$ is the standard income tax payment, which is the same for workers in both sectors. Workers pay for their training cost z. Finally, they receive subsidy $\mathcal{R}(w_-, y, h)$ that depends on their history.

C.9 Independent Instruments

Proposition 12. *(Misaligned income tax and retraining subsidy) Suppose the planner uses an income tax schedule $T(y) = T_0 + T_1(y)$ that depends only on current income, and retraining is fully subsidized, i.e., retraining wedge $\tau_z = 1$. Then in the decentralized economy, optimal training level increases with T_0.*

Proof. Given the two policy instruments, the FOC for retraining level z of an M-origin worker with productivity m is

$$-\phi'(z) + \int_{\underline{n}}^{\bar{n}} v_z(n, m, z) g(n|m, z) dn + \int_{\underline{n}}^{\bar{n}} v(n, m, z) g_z(n|m, z) dn = 0$$

Since $\tau_z = 1$, v is independent of z, so $v_z = 0$. We take derivative with respect to T_0 on the LHS, which is

$$\begin{aligned}
\int_{\underline{n}}^{\bar{n}} v_{T_0}(n, m, z) g_z(n|m, z) dn &= v_{T_0}(n, m, z) G_z(n|m, z)\Big|_{\underline{n}}^{\bar{n}} - \int_{\underline{n}}^{\bar{n}} v_{T_0 n}(n, m, z) G_z(n|m, z) dn \\
&= -\int_{\underline{n}}^{\bar{n}} v_{T_0 n}(n, m, z) G_z(n|m, z) dn
\end{aligned}$$

Notice

$$v_{T_0 n}(n, m, z) = -u''(y - T_0 - T_1(y(n))) T_1'(y(n)) y'(n)$$

Since $T_1'(y(n)) > 0$, $y'(n) > 0$, we have $v_{T_0 n}(n, m, z) > 0$. Combining with the assumption that $G_z < 0$, we have

$$\int_{\underline{n}}^{\bar{n}} v_{T_0}(n, m, z) g_z(n|m, z) dn > 0$$

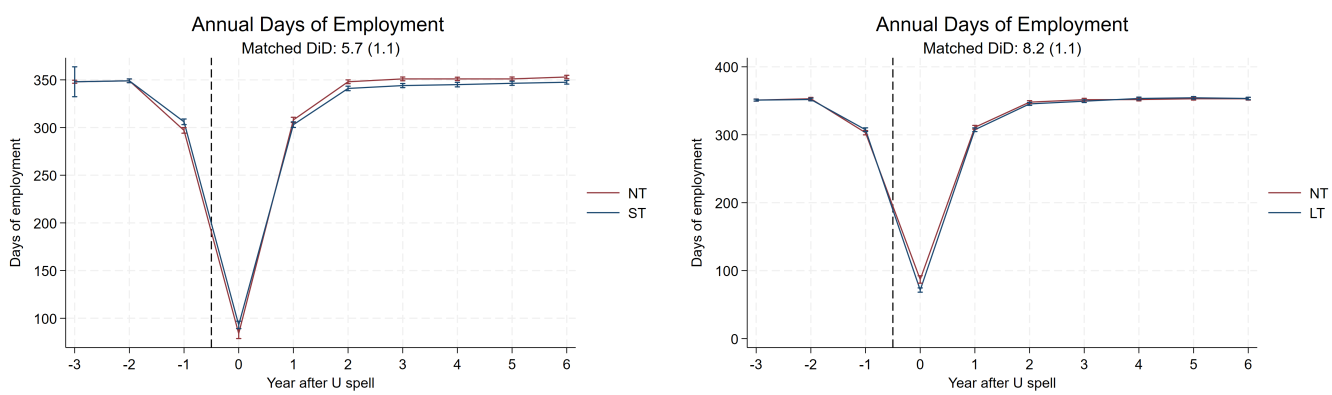
Hence, by monotone comparative statics, the optimal retraining level $z^*(m)$ must increase with T_0. In other words, if there is more direct subsidy, i.e., T_0 goes down, so does $z^*(m)$. $\square$

D Data, Empirical Evidence, and Calibration

D.1 Matching Event Study for Employment

In the main text, I show some suggestive evidence about the effectiveness of training programs in Germany using matching method. In particular, I show the wage path for a worker before and after unemployment who participated in a retraining program compared to those who didn't. Similar pattern appears if we look at the annual days of employment, though not as significant as earnings. The difference is summarized by the matching diff-in-diff estimator: compared to those who never participated in any training program during unemployment, those who did had slightly higher days of employment, and this effect is larger for those who took long-term training.

Figure 10: Annual days of employment for different worker groups in the matched sample



NT stands for no-training group, ST for short-term training group, and LT for long-term training group.

D.2 Model Inverting

Proposition 13. *(Identification) Conditional on n_B, assume there is no selection into different retraining levels except the extreme value preference shock, and the income tax schedule depends on income alone. Then, the μ's and σ's in the skill production technology, the scale parameter of the extreme value shock, ν, and training disutility ϕ^{ST}, ϕ^{LT} (normalizing $\phi^{NT} = 0$) can be identified from $\mathbb{P}[z^* = z|m]$, $\mathbb{P}[S = 1|z, m]$, and $\mathbb{E}[n|z, m, S = 1]$ at each z and at least three different m values.*

Proof. Suppose we know $\mathbb{P}[z^* = z|m]$, $\mathbb{P}[S = 1|z, m]$, and $\mathbb{E}[n|z, m, S = 1]$for $m = m^i, i = 1, 2, 3$. I first prove that the μ's and σ's in the skill production technology are identifiable. Since the income tax schedule does not depend on occupation, the occupation switching threshold must satisfy

$$\theta(m^i) = \frac{w_m m^i}{w_n}$$

Then, the switching probability can be written as

$$\begin{aligned} \mathbb{P}\left[S = 1|m^i, z\right] &= \int_{\theta(m^i)}^{\bar{n}} g(|m^i, z) dn_A \\ &= \bar{\Phi}\left(\frac{\log\theta(m^i) - \mu(m^i, z)}{\sigma(m^i, z)}\right) := x_1^{i,z} \end{aligned}$$

where $\Phi(.)$ is the cdf of a standard normal random variable, and $\bar{\Phi} := 1 - \Phi$. The second line follows from the formula for the truncated mean of a log-normal random variable. Similarly, the conditional mean of n can be written as

$$\begin{aligned} \mathbb{E}\left[n|S = 1, m^i, z\right] &= \frac{\int_{\theta(m^i)}^{\bar{n}} n g(n|m^i, z) dn}{\mathbb{P}[S = 1|m^i, z]} \\ &= \mu(m^i, z) + \sigma(m^i, z) \frac{\varphi(\frac{\log\theta(m^i) - \mu(m^i, z)}{\sigma(m^i, z)})}{\bar{\Phi}(\frac{\log\theta(m^i) - \mu(m^i, z)}{\sigma(m^i, z)})} := x_2^{i,z} \end{aligned}$$

where $\varphi(.)$ is the pdf of a standard normal random variable. Since $\Phi(.)$ is monotonic, we know from the first equation that

$$\frac{\log\theta(m^i) - \mu(m^i, z)}{\sigma(m^i, z)} = \bar{\Phi}^{-1}(x_1^{i,z})$$

Plug into the second one,

$$\mu(m^i, z) + \sigma(m^i, z) \frac{\varphi(\bar{\Phi}^{-1}(x_1^{i,z}))}{x_1^{i,z}} = x_2^{i,z}$$

Eliminating $\mu(m^i, z)$, we have

$$\log\theta(m^i) + \sigma(m^i, z) \frac{\varphi(\bar{\Phi}^{-1}(x_1^{i,z}))}{x_1^{i,z}} = x_2^{i,z} + \sigma(m^i, z)\bar{\Phi}^{-1}(x_1^{i,z})$$

Hence, as long as $\frac{\varphi(\bar{\Phi}^{-1}(m_1^{i,z}))}{m_1^{i,z}} \neq \bar{\Phi}^{-1}(m_1^{i,z})$, both $\mu(m^i, z)$ and $\sigma(m^i, z)$ have a unique solution:

$$\sigma(m^i, z) = \frac{x_2^{i,z} - \log \theta(m^i)}{\frac{\varphi(\bar{\Phi}^{-1}(x_1^{i,z}))}{x_1^{i,z}} - \bar{\Phi}^{-1}(x_1^{i,z})}$$

$$\mu(m^i, z) = x_2^{i,z} - \sigma(m^i, z)\frac{\varphi(\bar{\Phi}^{-1}(x_1^{i,z}))}{x_1^{i,z}}$$

Finally, knowing $\sigma(m^i, z)$ and $\mu(m^i, z)$ for $i = 1, 2, 3$, we can recover all the μ's and σ's assumed in the quadratic parametrization, assuming those three points are not colinear.

Given the new skill production function, I next show that the scaler parameter ν and the training disutility ϕ^{ST} and ϕ^{LT} are identifiable up to a normalization that $\phi^{NT} = 0$. Let $v(n, m^i)$ be the ex post value of a B-origin worker after n is revealed. It does not depend on the retraining level since the income tax does not. Let $V^z(m^i)$ be the ex ante value of taking retraining level z given the worker's skill m^i, excluding the preference shock. Then,

$$V^z(m^i) = \int v(n, m^i) g(n|m^i, z) dn - \phi^z$$

where ϕ^z is (additive) training disutility. Note the first part, $\int v(n, m^i) g(n|m^i, z) dn$, is known given the new skill production function and the tax schedule.

The actual expected value of choosing training level z is $V^z(m^i) + \xi^z$, where ξ^z is an independent preference shock drawn from a type-I extreme value distribution with location parameter zero and scale parameter ν. The type-I extreme value shock then leads to

$$\mathbb{P}\left[z^* = z|m^i\right] = \frac{e^{V^z(m^i)/\nu}}{\sum_{z^* \in \{NT, ST, LT\}} e^{V^{z^*}(m^i)/\nu}} := x_3^{i,z}$$

To identify ν, note that

$$\begin{aligned}
\frac{x_3^{1,ST}}{x_3^{1,NT}} &= \frac{\mathbb{P}\left[z^* = ST|m^1\right]}{\mathbb{P}\left[z^* = NT|m^1\right]} = e^{\left[V^{ST}(m^1)-V^{NT}(m^1)\right]/\nu} \\
\frac{x_3^{2,ST}}{x_3^{2,NT}} &= \frac{\mathbb{P}\left[z^* = ST|m^2\right]}{\mathbb{P}\left[z^* = NT|m^2\right]} = e^{\left[V^{ST}(m^2)-V^{NT}(m^2)\right]/\nu} \\
\Longrightarrow \frac{x_3^{1,ST}/x_3^{1,NT}}{x_3^{2,ST}/x_3^{2,NT}} &= e^{\frac{\left[V^{ST}(m^1)-V^{NT}(m^1)\right]-\left[V^{ST}(m^2)-V^{NT}(m^2)\right]}{\nu}} \\
&= e^{\frac{\left[V^{ST}(m^1)-V^{ST}(m^2)\right]-\left[V^{NT}(m^1)-V^{NT}(m^2)\right]}{\nu}}
\end{aligned}$$

Notice that training disutility $\phi^{NT}$and ϕ^{ST} are differenced out, and only ν is an unknown parameter, which is identified.

Finally, since

$$\frac{x_3^{1,ST}}{x_3^{1,NT}} = \frac{\mathbb{P}\left[z^* = ST|m^1\right]}{\mathbb{P}\left[z^* = NT|m^1\right]} = e^{\left[V^{ST}(m^1)-V^{NT}(m^1)\right]/\nu}$$

by normalizing ϕ^{NT} to zero, ϕ^{ST} is identified. So is ϕ^{LT} by the same argument. □

D.3 Moments Construction

The identification result shows that (1) $\mathbb{P}\left[z^* = z|m\right]$ for $z = NT, ST, LT$, the participation rate in each training type, (2) $\mathbb{P}\left[S = 1|z, m\right]$, the occupational switching rate conditional on manual productivity and training type, and (3) $\mathbb{E}\left[n|z, m, S = 1\right]$, the average non-manual productivity conditional on former manual productivity, training type, and having switched occupation, are important moments to identify the training technology. I also compute $var\left[n|z, m, S = 1\right]$ to test model fit. I now describe how to construct those moments.

I first discretize the manual productivity distribution into 10 bins based the quantiles, i.e., $[0, 0.1), [0.1, 0.2), ..., [0.9, 1]$. For each bin, I use the center point as the value of m, e.g., $m = 0.05$ for the first bin.

Next, within each bin, I calculate those moments. The primary goal is to identify the retraining technology for those non-manual-related programs, However, one limitation of the data is that I cannot observe the actual contents of those retraining programs. In other words, I do not know whether unemployed workers participate in manual-related training or non-manual-related training, while both types of training programs exist.

Therefore, some sample selections and assumptions are required to construct these moments. I assume for those who take training and then switch from M to N, they participate in non-manual training. Then, I select the sample such that only those who switch from M to N are included. With the assumption made, this is the relevant sample to identify the non-manual-related retraining technology.

One remark here is that in the model, a switch must imply wage increase. However, in the data, some occupational switch is accompanied with wage drop, which can never happen in the model. Therefore, to be consistent with the model definition, I label those who switch from M to N AND have a wage increase as successful switch, which corresponds to $S = 1$ in the model. With these assumptions, I can now compute $\mathbb{E}\left[n|z, m, S = 1\right]$ and $var\left[n|z, m, S = 1\right]$.

However, these assumptions are still not enough to compute the switching rate $\mathbb{P}\left[S = 1|z, m\right]$, as there may be some workers who take training z but stay in M. They are left out in the sample selection, but they should ener the demoninator. To deal with this issue, I assume for those who take non-manual training but do not draw a satisfying n, they are indifferent between going to N and staying in M. Then, I can adjust the switching rate in the following way. Assume conditional on z and m, the total number of workers that switch from M to N is s, and x of them have successful switch. Then, $s - x$ have unsuccessful switch. By the assumption, there are also $s - x$ who take training z but stay at M. Then, the actual switching rate is

$$\mathbb{P}\left[S = 1|z, m\right] = \frac{x}{x + s - x + s - x} = \frac{x/s}{2 - x/s}$$

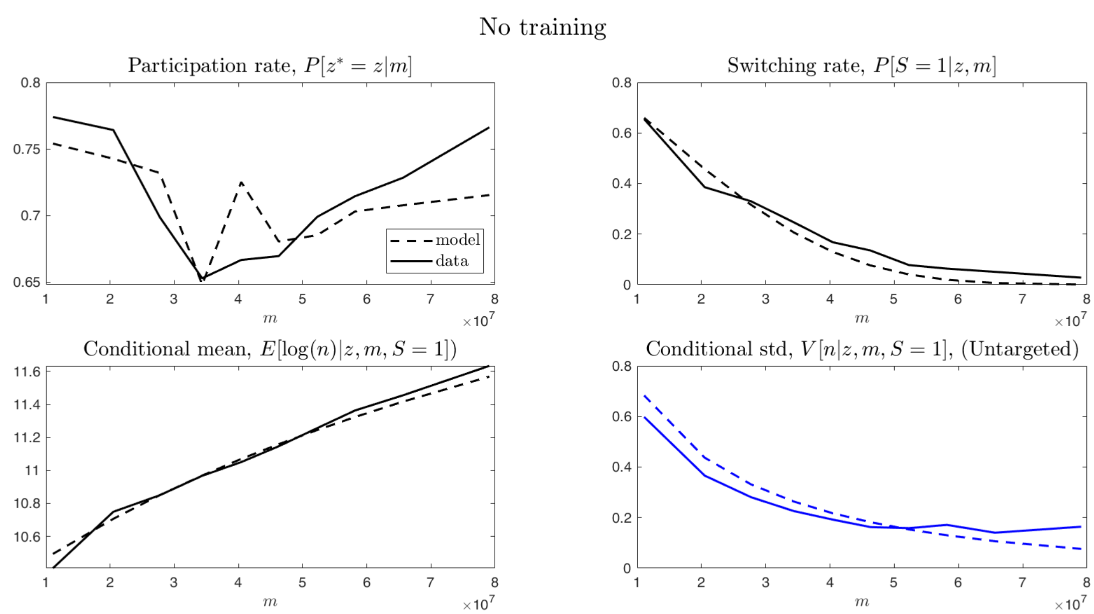
where $\frac{x}{s}$ can be read directly from the data.

Finally, I compute $\mathbb{P}\left[z^* = z|m\right]$ within the selected sample. Ideally, I want to know the intended-to-treat group, i.e., those who should be in the denominator. There are two potential biases for this moment from the sample selection. First, those who are potential training z participants but stay in M are left out. This leads to an upward bias in the computed statistic compared to the true participation rate. However, there are also many workers who never consider participating in training z even if they switch. They enter the denominator and lead to a downward bias in the computed statistic compared to the true switching rate. Hence, the direction of the bias is ambiguous. Without better data, a complete investigation is impossible. with a caveat, I use the moment computed here from the selected sample while acknowledging its potential bias.

D.4 Targeted and Untargeted Moments for NT and ST Groups

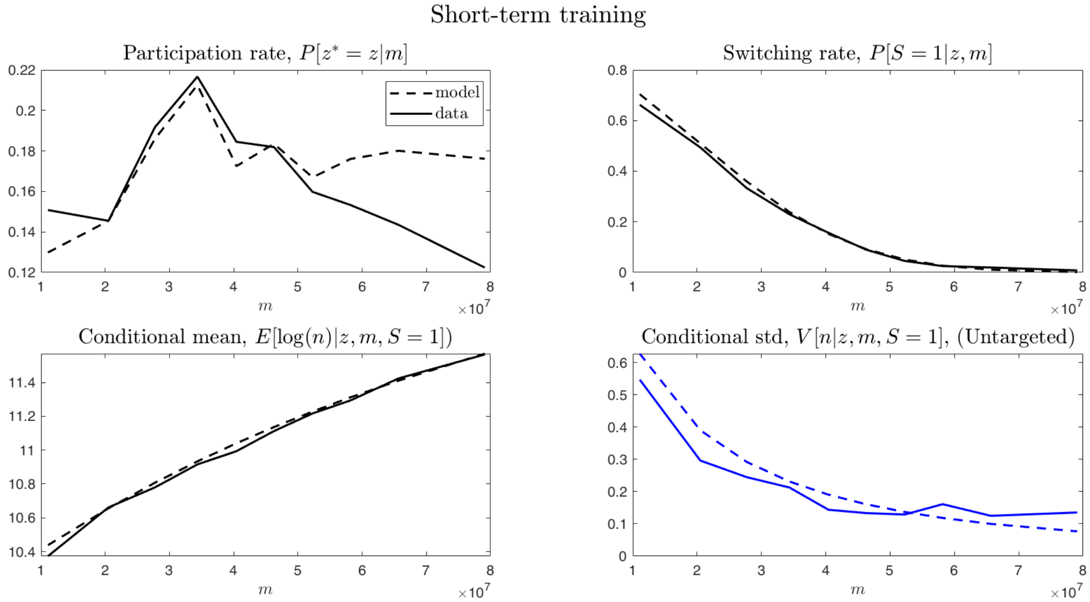
In the main text, I show the targeted and untargeted moments for the long-term training group, which demonstrates that the relatively few parameters can capture the empirical patterns quite well. For completeness, I show the same objects for the no-training and the short-term training group here. Similarly as the long-term group, the model can target the empirical moments quite well, especially for the untargeted one, the conditional variance of skills after occupational switch (in blue). The only place that may be slightly off the target is the short-term training participation rate for workers with median-to-high manual productivities. The model predicts that the participation rate should be relatively stable at 17 percent, while in the data, it drops from 17 percent to 13 percent. Still, the magnitude is very close, and the model can well capture the hump-shaped pattern of the participation rate.

Figure 11: Targeted and untargeted moments for no-training group



This figure shows the data and model-generated moments for short-term training, including the participation rate, the occupation switching rate, the average and the standard deviation of the non-manual productivity conditional on short-term training and switch, for different manual productivity levels. The conditional standard deviation is not targeted moment.

Figure 12: Targeted and untargeted moments for short-term training group



This figure shows the data and model-generated moments for no training group, including the participation rate, the occupation switching rate, the average and the standard deviation of the non-manual productivity conditional on no training and switch, for different manual productivity levels. The conditional standard deviation is not targeted moment.

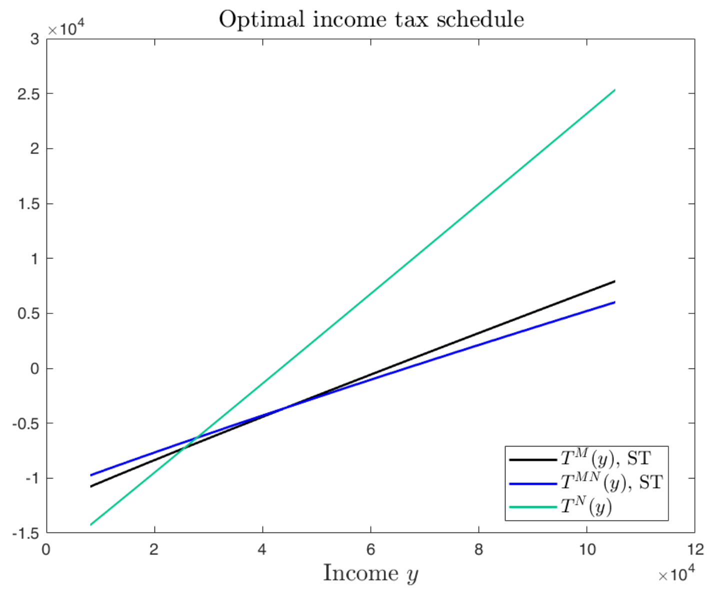
E Nonlinear Tax Schedule that Implements the Constrained Optimal Allocation

After discussing the optimal allocation in the constrained problem, Figure 13 shows the optimal history-dependent tax schedule that implements that allocation, where the green line is for incumbent N workers, the black and blue line for M stayers and M-to-N switchers, respectively, averaged over m. Four comments follow. First, for each group, the tax schedule looks quite linear. But this may hide the nontrivial labor wedges. Also, the nonlinearity across groups is quite significant. Second, the tax schedule for M-origin workers is much flatter than incumbent N workers,[36] which means the tax schedule is less aggressive to M-origin workers as they are greatly affected by the technology shock. Third, $T^{MN}(y)$ is higher than $T^{M}(y)$ for low-income workers but becomes lower for high-income workers. This is consistent with the optimal occupation switching

[36] For workers with extremely low income, incumbent N workers receive more subsidy than M-origin workers. This is because 1) M-origin workers on average have higher productivity from the initial distribution and 2) the restriction that incumbent N workers have no access to the training technology. So the planner wants to compensate for that.

threshold. Finally, one can also decompose and re-combine these tax schedules to construct, for example, sector-independent tax and history-dependent subsidies as in Findeisen and Sachs (2016) and Stantcheva (2017). The allocation and wedges will not change.

Figure 13: Optimal tax schedule



Notes: This figure plots the optimal non-linear income tax schedule, grouped at those who take short-term training and stay in M, who take short-term training and switch to N, and incumbent N workers.