Are Chemists Good Bankers?
Returns to the Match between Training and Occupation

Dita Eckardt*

March 18, 2022

Abstract

This paper analyzes the returns to training-occupation combinations. I use administrative panel data on apprenticeships and employment for German workers, and identify the returns using data on occupation-specific vacancies. For the estimation, I set up a Roy model and extend existing control function approaches to deal with selection in a two-stage, high-dimensional setting. I find sizable returns to training in one's occupation, and substantial bias when not controlling for selection. Returns are decreasing in the task distance between training and occupation. I argue that imperfect information leads to ex-post suboptimal training choices, and that retraining could address this ex-ante friction.

JEL codes: I26, J24, J31, J62
Keywords: *Training, occupational choice, high-dimensional selection, wage differentials, human capital specificity, task content of occupations*

*Department of Economics, University of Warwick, Email: Dita.Eckardt@warwick.ac.uk.
I am extremely grateful to my advisors Alan Manning and Johannes Spinnewijn for their support and guidance on this project. I thank Karun Adusumilli, Joseph Altonji, Miguel Bandeira Da Silva, Ethan Ilzetzki, Xavier Jaravel, Camille Landais, Eui Jung Lee, Steve Machin, Will Matcham, Guy Michaels, Steve Pischke, Daniel Reck, Yona Rubinstein, Kilian Russ, Uta Schönberg, John Van Reenen and Horng Chern Wong for helpful comments that substantially improved my work. This study uses the factually anonymous Sample of Integrated Employment Biographies (version SIAB-R 7510). Data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

1 Introduction

Since the seminal work by Becker (1964) and Mincer (1974), a large body of literature in economics has sought to causally identify the returns to education. This work has focused on estimating the returns to additional years of schooling by running so-called Mincerian regressions, using different approaches to overcome biases typically interpreted as resulting from an omitted ability variable. More recently, a growing number of papers explores the heterogeneity of these returns across college majors, generally concluding that the earnings differentials across fields of education are large (e.g. Arcidiacono (2004), Altonji *et al.* (2012)). However, while much of this literature acknowledges that average returns to major likely mask important heterogeneity across occupations, we know little about the interactions between field and occupation. In particular, the wage effects of working outside one's field are largely unknown. The present paper aims to bridge this gap in the literature. Understanding the differential returns to field-occupation combinations provides important insights on the value of field-specific human capital across occupations. Importantly, it also reveals key welfare and policy implications since workers could hold suboptimal trainings ex-post.

The challenges faced in estimating these returns are twofold. Firstly, datasets recording the field of education and occupation are often survey-based, and field-occupation matches are typically classified as "related" or "unrelated" in a subjective way. Secondly, and more importantly, causal identification requires accounting for the fact that individuals select into a field, but also subsequently select into one of many occupations. While descriptive studies show that working in an occupation related to one's field is associated with higher earnings, it is unknown to what extent this is driven by selection effects.[1] As Altonji *et al.* (2016) note, the fact that workers select at two stages and the first stage choice affects the return across options in the second stage poses a formidable estimation problem.

I address the first challenge by using administrative panel data on German apprentices and their subsequent careers for 1975-2010. Apprenticeships are the main form of upper- and post-secondary education in Germany, held by around 70% of those who obtained this level of education. An average of 40% of those are employed in an occupation different from the one they were trained in. The data contains information on all (un)employment spells. Importantly, since apprentices are partly trained in firms during their three-year apprenticeship, the data also contains the occupation workers were working in as apprentices (their training). As a result, trainings and occupations are objectively recorded, and defining training-occupation cells is straightforward. The full matrix of combinations has training as

[1] Examples of studies that find such correlations using subjective classifcations of field-occupation matches include Robst (2007), Nordin *et al.* (2010) and Lemieux (2014). Kinsler & Pavan (2015) build a structural model and find sizeable returns to working in an occupation "related" to one's college major.

row choice and occupation as column choice, with the same number of rows and columns. I estimate the returns in these cells.

Identification of Average Treatment Effects (ATEs) in models with multiple unordered treatments is complicated by a number of factors. To address the second challenge of causal identification, I combine the employment panel with data on the universe of occupation-specific apprenticeship vacancies that were posted via local employment agencies between 1978-2010. Given the institutional setting, recorded vacancies are unlikely to be driven by labor supply and instead serve as close proxy for occupation-specific demand. The identification strategy then involves using vacancies in *other options* as exogenous shifters into a particular training or occupation. In particular, I use expected vacancies in occupations other than the chosen one as instruments for a training choice, with the idea that individuals consider future returns across occupations when making their choice. Subsequent shocks to these expectations in occupations other than the chosen one are used as instruments for occupations. Using vacancies *outside* the chosen option is key to satisfy the exclusion restriction. The instruments are highly relevant, confirming the importance of earnings expectations for occupation choices (Miller (1984), Keane & Wolpin (1997), Arcidiacono *et al.* (2020)). Since the instruments lack full support, I additionally rely on distributional assumptions in order to identify ATEs, but these can be relaxed for the slope parameters in the wage equation.

To put structure on the selection problem, I set up a generalized Roy (1951) model. The model provides a behavioural justification for the identification strategy and it implies a monotonicity assumption (Vytlacil (2002)), but it does not impose further assumptions for identification. In the model, workers choose a training in an initial stage, and subsequently select into an occupation in every work life period. While training is chosen to maximize expected payoff including expected wages, occupations are chosen to maximize current payoff including current wages. Importantly, individuals have imperfect information about future labor demand and own occupation-specific abilities when choosing a training. As a result, unexpected changes in labor demand or new information about abilities may lead to individuals choosing employment outside their training.

Given the high dimensionality of the selection problem, implementing the identification strategy is not straightforward. A parametric generalization of the classic two-step Heckman (1979) approach would not be feasible in this context. Lee (1983) and Dahl (2002) develop a control function approach that deals with selection in high-dimensional settings, where the control function becomes a function of a few selection probabilities only. I extend their approach to two selection stages and implement this using a machine learning algorithm (random forests), where I predict training and occupation choices with the instruments to get consistent estimates for the selection probabilities.

The returns to training-occupation combinations depend on the granularity of the underlying occupation classification. Given the availability of the vacancy data, I use a classification with 13 categories. Using these, I find the following three main results. Firstly, focusing on effects on versus off the diagonal, my results suggest that individuals trained and working in the same occupation on average earn around 15% more than workers employed in occupations different from their training. The effect is highly significant and comparable in magnitude to estimates of the return to a year of schooling in general, and an additional year as an apprentice in the German system more specifically (Fersterer *et al.* (2008)). I find evidence that the return is strongest at the beginning of a career, but only drops by about two percentage points before stabilizing after 12 years of experience. Off-diagonal workers thus do not catch up with their on-diagonal co-workers.

Secondly, not controlling for selection leads to substantial negative bias in the estimated returns to working on versus off the diagonal such that, descriptively, on-diagonal workers have lower wages. Intuitively, only the more able workers work off the diagonal as their unobserved occupation-specific ability needs to compensate for the lack of training. The majority of this bias is visible right after the training, confirming recent results in the sorting literature that workers sort on wages early in their career (Lentz *et al.* (2021)). However, I also find evidence that the bias becomes stronger with experience, a result which is in line with the proposed model where individuals have imperfect information on their occupation-specific ability. As more information is revealed to workers about these abilities, they decide to work in an occupation different from their training if the gain in payoff exceeds the cost of lack of training. As a result, workers on the diagonal are increasingly negatively selected.

Thirdly, my results display important heterogeneity. Across trainings, I find large differences in the average returns to working on versus off the diagonal, and a strong positive correlation between these returns and the fraction of workers with the relevant training observed working on the diagonal. In line with the proposed model, relative returns thus appear to be a key determinant of the selection into occupations. Even within training, there is substantial heterogeneity in the penalty when working in different occupations. I use the estimated returns from the full training-occupation matrix to provide a microfoundation for the results in this paper by drawing on the task approach to occupations (Autor *et al.* (2003), Autor (2013)). Measures of task content have been used empirically to analyze shifts in the wage structure both between occupations (e.g. Autor *et al.* (2003) , Goos *et al.* (2014)) and within occupations (e.g. Van der Velde (2020)).[2] Most relevant to the present paper, Poletaev & Robinson (2008) and Gathmann & Schönberg (2010) argue that the transferability of human capital across occupations depends on how similar occupations are in terms of

[2]Altonji *et al.* (2014) use task content measures to study the earnings inequality across college majors.

their mix of tasks, showing that wage drops after displacement are larger for workers who move to occupations that are less related to the previous one. Based on this evidence, one may expect workers in the present context to incur larger wage penalties, the more distant the occupation is from the training. To test this conjecture, I construct training-occupation distance measures for every training-occupation cell using survey data on the task content of occupations. In a second step, I regress the estimated returns for each training-occupation match on these measures. The results suggest that a one-standard-deviation increase in task distance significantly reduces the return in a training-occupation cell by $5-7pp$. Overall, the findings provide strong evidence that workers are trained in a specific mix of tasks and face higher wage penalties, the less applicable the acquired skills are in their current occupation.

The reason underlying ex-post suboptimal training choices in the given framework is lack of information at the time of training choice. Two groups of workers are affected by this. The first group are off-diagonal workers. These workers forego the return to training in their current occupation. The second group are workers who are locked into their training. These individuals work on the diagonal, but would choose a different occupation in the absence of off-diagonal penalties. My findings suggest that almost 70% of workers either work off the diagonal or are locked in, i.e. only 30% hold the optimal training ex-post. Using these shares, I estimate the associated welfare loss to be around $6-10\%$ of wages per worker. Back-of-the-envelope calculations suggest that ex-post retraining could effectively address the ex-ante lack of information for a majority of workers.

From a more general policy perspective, my results also speak to the wider debate on Germany's apprenticeship system as a role model. It has often been argued that the system facilitates labor market entry by providing young workers with specialized skills, thereby leading to low youth unemployment rates. My findings suggest that while German apprenticeships successfully deliver occupation-specific skills, many workers cannot fully put these to use in their chosen occupation.

This paper contributes to four strands of literature. Firstly, it relates to the literature on the returns to college majors. Arcidiacono (2004) estimates a dynamic structural model of college and major choice, finding large relative earnings premiums for certain majors. Altonji *et al.* (2012) and Altonji *et al.* (2016) provide surveys of the theoretical and empirical literature on the returns to college majors and document earnings differentials that can exceed the college-high school premium. Similar results are found by Hastings *et al.* (2013) and Kirkebøen *et al.* (2016) who exploit admission cutoffs to majors in Chile and Norway in a regression discontinuity framework. I contribute to this literature by analyzing the important heterogeneity that returns to fields display across occupations.

To address the challenges arising from the selection into trainings and occupations, this

paper contributes to a second literature, on the estimation of treatment effects in selection models with multiple unordered treatments. Heckman & Robb (1985) show that, in single-index models, control functions may be written as functions of the propensity to self-select.[3] In multiple-index models, Lee (1983) and Dahl (2002) develop a control function estimator where the control function becomes a function of a small set of selection probabilities only. A recent application of this approach can be found in Ransom (2021) who studies how the returns to schooling are affected by the selection of workers into locations and occupations. With multiple unordered treatments, identification of ATEs is particularly challenging (Heckman *et al.* (2006, 2008)). I contribute to this literature by extending the Lee/Dahl approach to a two-stage selection setting, and combining it with an instrumental variable strategy to identify and estimate ATEs in a setting with multiple unordered treatments.

Thirdly, this paper relates to the literature on human capital specificity. The idea that human capital is specific was proposed by Becker (1962, 1964) and extended by Lazear (2009) in the context of the firm, and has been taken to the data to explore specificity along a number of dimensions such as industry (Neal (1995)), occupations (Shaw (1984, 1987), Kambourov & Manovskii (2009)) and skills (Poletaev & Robinson (2008), Guvenen *et al.* (2020)). Most recently, a strand of this literature considering the tasks accumulated over a work life suggests that human capital is partly task-specific, and thus more easily transferable across occupations that require a similar mix of tasks (Gathmann & Schönberg (2010), Yamaguchi (2012), Cortes & Gallipoli (2018)). The present paper contributes to this literature by linking wages in different occupations to training received in the same occupations. To the best of my knowledge, it is the first to provide such estimates.

The transferability of skills is of particular importance in the face of sectoral shocks. A final related literature documents persistent adjustment costs for workers resulting from trade shocks (e.g. Autor *et al.* (2013), Autor *et al.* (2014)) or industry regulation (Walker (2013)). Most recently, the Covid-19 pandemic has led to dramatic shifts in the structure of labor markets around the world. In the German context, Yi *et al.* (2017) show that the impact of sectoral shocks is related to the ability to reallocate jobs to a new sector. I contribute to this literature by suggesting a microfoundation for the large persistent impacts that sectoral shocks have been found to have on worker outcomes.

The remainder of this paper is organized as follows. Section 2 outlines the setting and discusses the data. Section 3 sets up the generalized Roy model. Section 4 discusses identification. Section 5 outlines the estimation using control functions. Section 6 presents and discusses the results. Section 7 relates my findings to the task distance between trainings and occupations. Section 8 discusses welfare and policy implications. Section 9 concludes.

[3] Ahn & Powell (1993) and Das *et al.* (2003) derive semi-parametric versions of such control functions.

2 Setting and Data

2.1 The German Apprenticeship System

The German apprenticeship system is a dual system where apprentices work in firms for three to four days a week and go to vocational school for the remaining one to two days. While the training in firms provides apprentices with the necessary practical skills, vocational schools teach theoretical skills in a number of different subjects. The total apprenticeship length varies between two and three and a half years depending on the apprenticeship occupation, but the majority of apprenticeships last three years.

Dual apprenticeships are the main form of upper- and post-secondary education in Germany and, in 2010, about 70% of those who obtained this education level had completed an apprenticeship in the dual system.[4] Unlike other education forms, the dual system is regulated under a federal vocational training law (*Berufsbildungsgesetz*) which implies a large degree of standardization. The system is often regarded as the key pillar of the German education system, supporting low youth unemployment rates by facilitating the transition of young individuals into the labor market.

Importantly, the dual system trains apprentices in most non-university occupations, with only a small number of exceptions in the medical and care occupations. To start a dual apprenticeship, high-shool graduates need to apply to and be offered an apprenticeship position with a firm. Once the firm accepts an apprentice, it is in charge of providing the necessary practical training which is regulated under the legally defined training regulations (*Ausbildungsordnung*). The state government is responsible for providing a place at the local vocational school to any apprentice who has been accepted by a firm. The curriculum is determined centrally by each state for each apprenticeship occupation (*Rahmenlehrplan*) and consists of general and specialized subjects which may vary depending on the apprenticeship occupation.

All dual apprenticeships are completed through a final examination which is organized and monitored by industry-specific boards (*Kammern*) and consists of a theoretical and a practical part. After completing their apprenticeship, apprentices often continue to be employed at the same firm as full-time employess ($\sim 60\%$ in 2010).[5]

[4]Source: *Statistisches Bundesamt, Bildungsstand der Bevölkerung. Ergebnisse des Mikrozensus 2016. Ausgabe 2018.* Upper- and post-secondary education corresponds to ISCED levels 3 and above. The fraction of workers who obtained this education level was around 85% among young workers in Germany in 2018 (Source: *OECD Education at a Glance, 2019*).

[5]Source: *Institut für Arbeitsmarkt- und Berufsforschung, Statista 2018.*

2.2 Data

This paper uses two main datasets: an administrative employment panel covering 1975-2010, and a dataset containing the universe of occupation-specific apprenticeship vacancies posted through local employment agencies between 1978-2010.

2.2.1 Employment Panel

The employment panel dataset consists of a 2% sample of all German social security records between 1975-2010.[6] These records are based on all German workers employed in at least one job during that time period, with the exception of the self-employed, civil servants and those serving in the military. This amounts to about 80% of the workforce. Before 1991, only West Germany is included in the sample, from 1991 the records cover both West and East Germany. Workers who are selected in the sample are followed for the entire time period. The dataset includes demographic information such as gender and date of birth as well as detailed daily information for each (un)employment spell including the start and end date, occupation, industry, location and daily wage. Wages reported in the data are capped at a time-varying threshold defined within the statutory pension scheme. In my setting, this threshold will only affect a small fraction of the data (see Section 2.5).

Importantly, since apprentices in the dual system work in firms for three days a week, they pay social security contributions and their apprenticeship spells are contained in the employment panel dataset. I therefore observe the occupation that apprentices are employed in during their apprenticeship which I refer to as the *training*.

2.2.2 Vacancy Data

I use a second dataset for my analysis which contains the universe of apprenticeship vacancies posted through local German employment agencies between 1978-2010.[7] Between 1978-1992, the data only covers West Germany. From 1993, vacancies are recorded for both West and East Germany.[8] Recorded vacancies include those filled and those not filled after a year and aggregate information is available by year, training and location. Yearly data is measured as a flow of vacancies posted between 1. October and 30. September, but most vacancies are posted to line up with the schooling leaving dates in late summer.

[6]Sample of Integrated Employment Biographies. The data was provided by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

[7]This data combines different datasets provided by the German Federal Employment Agency. Source: *Arbeitsmarkt in Zahlen, Ausbildungsstellenmarkt, Bewerber und Berufsausbildungsstellen.*

[8]In addition to firm-based vacancies, the data also includes apprenticeship vacancies outside the dual system, but these make up a small fraction of the data ($\sim 12\%$). Figures based on 2010. Source: *Arbeitsmark in Zahlen - Bewerber und Berufsausbildungsstellen Deutschland, September 2010.*

A particular advantage of using data on apprenticeship vacancies as opposed to non-apprenticeship vacancies is that it specifically refers to jobs that can be carried out by those who went through the apprenticeship training system.[9] Moreover, in contrast to general job vacancies, the degree of involvement of employment agencies for apprenticeship vacancies is high. In 2013, 71% of firms publicized their apprenticeship vacancies through an agency, while the same figure only amounted to 45% for non-apprenticeship vacancies.[10]

2.3 Field-Based Occupational Classification

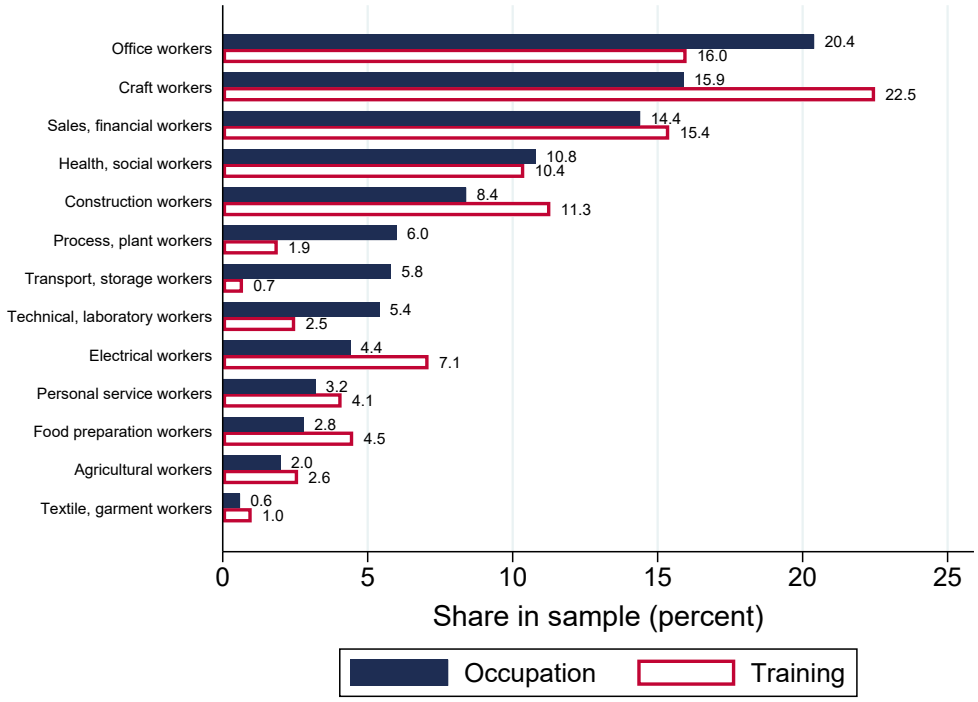
Occupations in the employment panel and the vacancy dataset are coded based on the same occupational classification called *Klassifikation der Berufe 1988 (KldB88).* This former German occupational classification system was replaced by the current system (*KldB2010*) in 2010. For the purpose of this paper, the *KldB88* has a key advantage over the newer system and other internationally used systems such as the *International Standard Classification of Occupations* or the *Standard Occupational Classification* in that it is field-based. Other systems generally contain a broad category for *Managers* and as a result, being promoted could imply that workers change their occupation in the classification. It would be impossible to accurately translate the hierarchical occupation categories of these classifications into a field-based system, and these measurement problems would be a major concern in the present analysis where the combination of training and occupation choices is of key interest. In fact, the *KldB88* was revised significantly to form the *KldB2010* precisely because, given its field-based structure, it was not comparable to most other international occupational classifications. Directly using the former field-based system *KldB88* in the present analysis therefore offers a unique opportunity to study the question of interest.

To make the analysis feasible and tractable, I use a classification level that groups occupations into 13 distinct categories, implying 169 distinct cells in the training-occupation matrix. The chosen level of categorization is the narrowest for which the historical vacancy data is available. Restricting the number of categories also ensures that the estimation will remain computationally feasible. A list of the 13 occupations and trainings together with their sample shares is shown in Figure 1.

[9] As an example, job vacancies advertised as health care occupations may refer to doctors and nurses, but only those as nurses would be relevant for a worker who does not hold a medical degree.

[10] Source: *BIBB-report 3/2014 Betriebe auf der Suche nach Ausbildungsplatzbewerberinnen und -bewerbern: Instrumente und Strategien,* and *IAB-Stellenerhebung 2013.* The degree of involvement of local employment agencies may vary over the business cycle and firms are more likely to post vacancies through the local agency when the supply of apprentices is low. The resulting time-variation will be picked up by time fixed effects which are included in all regression specifications.

Figure 1: Occupations and Trainings with Sample Shares



Notes: The figure lists the 13 occupations used, and plots their baseline sample shares by occupation and training. A detailed list of sub-categories contained in each occupation group is provided in Table A.1.

2.4 Sample Selection

Since individuals can have more than one employment relationship at a time, some of the spells overlap in the administrative data . I define the main employment spell as the highest daily wage spell and drop all secondary spells from the sample. Of the remaining individuals, I only keep those who were enrolled in exactly one apprenticeship in the dual system at some point during the sampling period. To ensure that the apprenticeship was completed, I further exclude individuals who were never classified as having completed their apprenticeship in any of their employment spells. Finally, I drop individuals whose training occupation or location is unknown when they start their apprenticeship.

For the remaining individuals, I restrict spells to full-time employment and exclude those with missing location, occupation or missing (or zero) wages.[11] Finally, I only keep employment spells which started after the end of the apprenticeship training and for which employers recorded the highest education level as vocational training. This excludes both lower education levels (apprenticeship is not recorded as completed) and higher education

[11] Zero wages indicate interrupted employment spells due to e.g. unpaid maternity leave or sabbaticals.

levels (additional university or technical college degree), to ensure that the amount of education as measured by years of schooling is comparable across the sample. The resulting baseline sample contains 291,098 individuals and 4,012,034 employment spells.

2.5 Descriptive Statistics

Table 1 provides summary statistics for the baseline sample. About 48% of individuals work outside their training occupation for at least one spell, and 38% work in more than one occupation throughout their career. Since apprenticeship spells need to fall within the sampling period for all individuals, the average worker is only 31 years old. As a result, mean daily wages are relatively low implying that less than 3% of wages exceed the upper earnings limit in the statutory pension scheme and are capped in the sample.

To give a sense of the empirical distribution across training-occupation cells, Table 2 reports the percentage of spells in each occupation for the five largest trainings. Table 3 reports the same figures for the five largest occupations.[12] Both tables are restricted to spells of workers with ten years of full-time work experience. It can be seen that, while the majority of individuals work on the diagonal, the fraction of individuals working off the diagonal is large, and displays considerable heterogeneity across trainings and occupations.

Table 1: Summary Statistics

	Mean	Min	Max	P^{10}	P^{50}	P^{90}
N of observations/spells	4,012,034					
N of individuals	291,098					
Female (% of individuals)	45.4					
Female (% of spells)	37.3					
Individuals ever off diagonal (%)	47.6					
Occupation switchers (%)	37.7					
Occ. switches per individual	0.7	0	38	0	0	1.5
Distinct occ. per individual	1.5	1	10	1	1	2.5
Age	30.6	17	62	20.5	28.5	42.5

Notes: The table reports summary statistics for the baseline sample.

Figure 2 looks at the variation in occupation choice over a career by plotting the fraction of individuals working in an occupation equal to their training by full-time work experience. While around 75% of all workers start their career after the apprenticeship working in their

[12]Since only selected categories are reported, the row percentages in Table 2 and column percentages in Table 3 do *not* sum to 100. Tables A.2 and A.3 contain equivalent figures for all trainings/occupations.

training occupation, this fraction drops to around 65% after 6 years, and around 55% after 25 years of full-time work experience. Figure B.1 shows that this decline is not due to compositional effects in the sample by plotting the fraction of individuals working on the diagonal over time for different experience levels.

Table 2: Spells as Percentage of Trainings - Selected Categories

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	80.6	0.6	12.6	1.6	0.1
	Craft workers	4.8	55.3	3.9	2.4	2.5
	Sales, financ. w.	26.5	1.6	60.6	2.1	0.3
	Health, social w.	12.2	0.7	4.3	79.0	0.2
	Construction w.	3.6	5.7	3.1	2.9	60.2

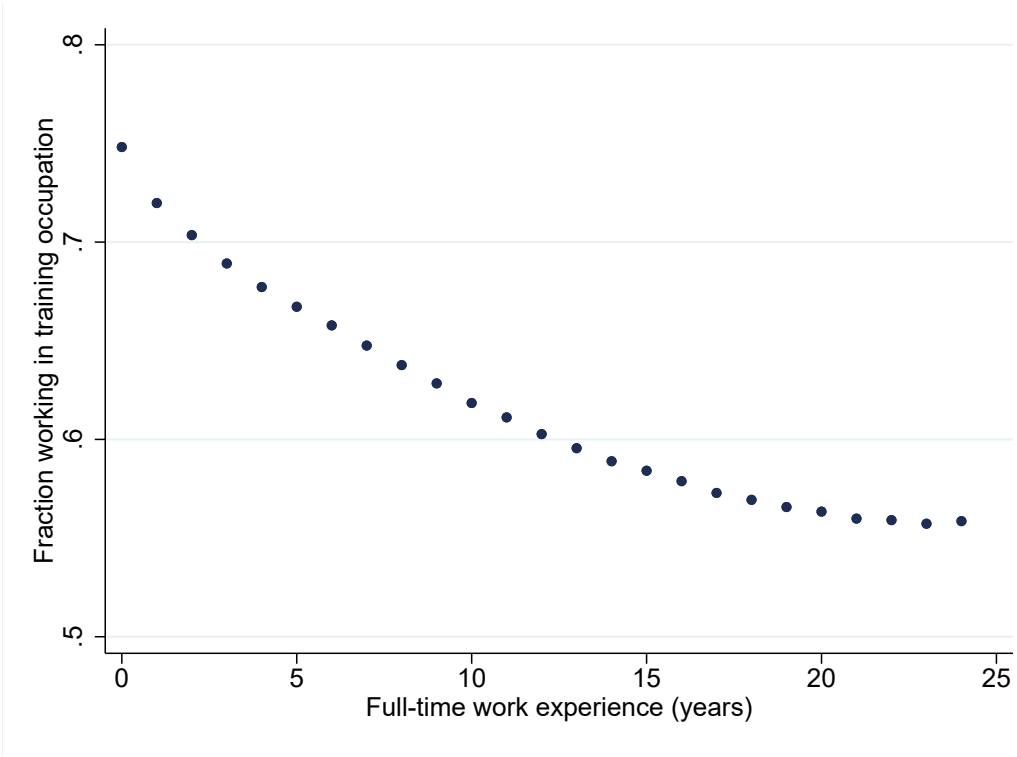
Notes: The table reports the number of spells with a particular training-occupation combination as a percentage of all spells in the training for the baseline sample. Results are restricted to individuals with ten years of work experience. Only the five most common trainings and occupations are reported. As a result, row percentages do not sum to 100. Table A.2 contains all 13 training/occupation categories.

Table 3: Spells as Percentage of Occupations - Selected Categories

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	59.4	0.7	14.4	2.8	0.3
	Craft workers	5.0	84.3	6.2	5.7	7.5
	Sales, financ. w.	18.3	1.7	64.8	3.4	0.6
	Health, social w.	5.1	0.4	2.7	76.4	0.3
	Construction w.	1.7	4.0	2.3	3.2	85.1

Notes: The table reports the number of spells with a particular training-occupation combination as a percentage of all spells in the *occupation* for the baseline sample. Results are restricted to individuals with ten years of work experience. Only the five most common trainings and occupations are reported. As a result, column percentages do not sum to 100. Table A.3 contains all 13 training/occupation categories.

Figure 2: Fraction On Diagonal by Work Experience



Notes: The figure plots the fraction of individuals working in their training occupation by full-time work experience for the baseline sample.

3 Selection Model

In this section, I present a statistical model of selection into training and occupations. I model the choice problem using a generalized two-stage Roy (1951) model, where selection is based on an underlying latent utility. The model provides a behavioral justification for the identification strategy proposed in Section 4, but (apart from implying a monotonicity assumption) it does not impose any assumptions for identification. The identification assumptions will be presented in Section 4 where I also discuss the type of economic model that would justify these assumptions.

The threshold-crossing nature of the proposed statistical model puts sufficient structure on the choice problem to implement the identification strategy using a control function approach which I describe in Section 5. Aside from that, the specification of latent utility is not restricted by the model, meaning that I do not need to take a stance on the specifics of how individuals select. Importantly, this implies that my empirical approach is robust to a range of economic selection models.

3.1 Setup and Wages

Training and occupation choices are modeled as a two-stage selection problem. In $t = t_0$ (stage 1), individual i selects into a training indexed by $j = 1, ..., J$. In $t = t_0 + 1, ..., t_0 + T$ (stage 2), individual i selects into an occupation indexed by $k = 1, ..., K$. While stage 1 involves a single selection choice in $t = t_0$, stage 2 involves T selection choices, one for each period $t = t_0 + 1, ..., t_0 + T$. Note that the set of training and occupation options individual i chooses from is identical.

In stage 2, log wages of individual i working in occupation k with training j follow:

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \tau_{jk} + \beta' X_{it} + \epsilon_{ikrt}, \tag{1}$$

where δ_r, δ_t, δ_i denote region, time and individual fixed effects, respectively. The term $f(vac_{krt})$ denotes a flexible function in log vacancies vac_{krt} posted for occupation k in region r at time t. X_{it} is a vector of individual- and time-varying controls including full-time work experience, and ϵ_{ikrt} denotes an individual error component which is assumed to vary across occupations but not trainings. This captures the idea that individuals have an unobserved occupation-specific ability affecting wages, but there is no heterogeneity in the ability to productively use a training in a particular occupation. The $(J \times K)$ fixed effects τ_{jk} are the parameters of interest. These parameters capture the log wage effect from a particular combination of training j and occupation k. Note that the wage specification from equation (1) may be derived from a standard exponential human capital production model (Griliches (1977)), where log wages are equal to the sum of a log skill price $\bar{w}_{krt} = \delta_r + \delta_t + f(vac_{krt})$ and log human capital $h_{ijkrt} = \delta_i + \tau_{jk} + \beta' X_{it} + \epsilon_{ikrt}$. I further discuss the economic model underlying the wage specification of equation (1) in Section 4.

3.2 Occupation Choice

In period $t > t_0$, occupation choices are based on an underlying latent utility which is additively separable in a sub-utility function $\tilde{U}_{i(k|j)rt}$ and an error component e_{ikrt}:

$$U_{i(k|j)rt} = \tilde{U}_{i(k|j)rt} + e_{ikrt}. \tag{2}$$

The sub-utility function captures the part of utility that depends on observable characteristics and is therefore observed to the researcher. This includes both the observed component of log wages as well as non-monetary observed preference components, t_{ikrt}, so that $\tilde{U}_{i(k|j)rt} = ln(w_{ijkrt}) - \epsilon_{ikrt} + t_{ikrt}$. The error term comprises unobserved components of utility including the log wage error term ϵ_{ikrt} as well as unobserved preference components, ψ_{ikrt}, so

that $e_{ikrt} = \epsilon_{ikrt} + \psi_{ikrt}$. Note that workers have unobserved preferences across occupations, but preferences do not vary by the match between training and occupation.

Individual i chooses occupation k to maximize the current-period utility $U_{i(k|j)rt}$. Using the above notation, individual i chooses occupation k in period t if and only if

$$(e_{ikrt} - e_{ik'rt}) \geq (\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt}), \quad \forall k' \neq k. \tag{3}$$

I define a corresponding occupation dummy variable

$$occ_{i(k|j)rt} = \begin{cases} 1 & \text{if } U_{i(k|j)rt} \geq U_{i(k'|j)rt}, \quad \forall k' \neq k, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that the occupational choice problem defined by equation (3) is static, i.e. past occupation choices do not affect current occupation-specific utility. This assumption does not preclude individuals from being forward-looking, but it restricts the type of wage equation that can be part of individuals' latent utility. For instance, if work experience in and outside of an occupation had differential effects on wages, today's occupation choice would affect the utility across options in the future, and the choice problem would be dynamic. To keep the estimation problem tractable, I abstract from such differences and only account for total work experience in the model.[13]

3.3 Training Choice

In period t_0, training choices are based on a period-t_0 training utility $U_{ijr_0t_0} = \tilde{U}_{ijr_0t_0} + e_{ijr_0t_0}$, and the expected future utility of choosing training j. Define the value of choosing j as the sum of these two components:

$$V_{ijr_0t_0} = \tilde{U}_{ijr_0t_0} + e_{ijr_0t_0} + E_{t_0}[\sum_{t>t_0} \beta^{t-t_0} U^*_{i(k|j)rt}], \tag{5}$$

where $E_{t_0}[\sum_{t>t_0} \beta^{t-t_0} U^*_{i(k|j)rt}]$ is individual i's maximal expected future reward, conditional on training choice j in $t = t_0$. Given the static nature of the occupational choice problem, the expected maximal reward depends on the probability of choosing different occupations k conditional on training j in the future. Since individuals hold imperfect information on own abilities, preferences and future labor market developments, this will generally differ

[13] The endogeneity of occupation-specific experience makes it difficult to assess the validity of this simplification, but including occupation-specific experience in the baseline regression suggests that the difference in returns inside and outside the current occupation is only about 1.8*pp* (see Table 4 in Section 6.1). Imposing equal returns therefore appears to be a reasonable approximation.

from the discounted stream of realized utilities. In line with the occupation choice problem, individual i chooses training j in $t = t_0$ if and only if

$$(e_{ijr_0t_0} - e_{ij'r_0t_0}) \geq (\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0}), \quad \forall j' \neq j, \tag{6}$$

where $\tilde{V}_{ijr_0t_0} = \tilde{U}_{ijr_0t_0} + E_{t_0}[\sum_{t>t_0} \beta^{t-t_0} U^*_{i(k|j)rt}]$ is the conditional value $V_{ijr_0t_0} - e_{ijr_0t_0}$. As before, I define a training dummy variable

$$train_{ij} = \begin{cases} 1 & \text{if } V_{ijr_0t_0} \geq V_{ij'r_0t_0}, \quad \forall j' \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where subscripts r_0 and t_0 for $train_{ij}$ are omitted for expositional clarity as individuals only choose their training once. Note that the individual takes into account that training choice j will affect the utility across different occupations in the future.

4 Identification

This section explores potential biases due to workers' selection into a training and an occupation, and provides intuition for these biases using two hypothetical experiments. To address the challenges resulting from self-selection in the given setup with multiple unordered treatments, I then propose an instrumental variables strategy. I discuss the identification assumptions, and present a range of robustness checks to support the assumptions made.

4.1 Selection Biases

Wages in a training-occupation cell are only observed for a sample of individuals who selected into that cell, and the non-random allocation into cells may lead to selection biases. Based on the definition of the training and occupation dummies from Sections 3.2 and 3.3, the selection problem in outcome equation (1) may be written as

$$E[\epsilon_{ikrt} | train_{ij} = 1, occ_{ij(k|j)rt} = 1] \neq 0. \tag{8}$$

Given the two-stage selection in the present context, identifying the effect of a particular training-occupation combination on wages requires randomizing individuals to a training-occupation cell. The ideal experiment would therefore involve initial random allocation to a training, followed by random allocation to an occupation. In order to illustrate why randomization at the training or the occupation stage alone will not identify the parameters of

interest, I consider two hypothetical experiments that look at the selection into trainings and occupations independently. The key insight from these experiments is that, while selection into training is expected to lead to *positive* bias in the estimated on- versus off-diagonal return, selection into occupations is expected to lead to *negative* bias in the return. Intuitively, the former is explained by individuals choosing a training they are relatively good at. The latter is due to the fact that, conditional on training choice, off-diagonal workers must be especially good in their chosen occupation to compensate for their lack of training.

For ease of illustration, focus on a stylized version of the selection model with two trainings and occupations, $j, k \in \{1, 2\}$, and two time periods, $t \in \{0, 1\}$. Individuals train in $t = 0$ and work in $t = 1$. Denote by $train_{ij}$ and occ_{ik} the training and occupation dummies which are equal to one if individual i is trained in j/works in k. Assume that individuals self-select into trainings and occupations based on a simplified version of the Roy (1951) model where the training is chosen in $t = 0$ to maximize expected log wages and the occupation is chosen in $t = 1$ to maximize current log wages. Assume homogenous returns to working on the diagonal, and denote this return by $\tau > 0$.[14] Finally, denote by $\epsilon_{i1}, \epsilon_{i2}$ the log wage error terms in $t = 1$ in occupations 1 and 2, and assume that these are known in $t = 0$.[15]

4.1.1 Selection into Training

Consider a first hypothetical experiment where individuals choose their training j and are subsequently randomly allocated to an occupation k. Assume that individuals do not know that occupations are randomly allocated when making their training choice. Focusing on occupation 1, the selection bias when estimating parameter τ may be written as

$$\begin{aligned} &E[\epsilon_{i1}|train_{i1} = 1, occ_{i1} = 1] - E[\epsilon_{i1}|train_{i1} = 0, occ_{i1} = 1] \\ &\quad = E[\epsilon_{i1}|train_{i1} = 1] - E[\epsilon_{i1}|train_{i1} = 0] \\ &\quad = E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) > 0}_{\text{chose training 1}}] - E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) < 0}_{\text{chose training 2}}] \geq 0, \end{aligned} \tag{9}$$

where the difference in the observed component of expected log wages has been normalized to zero. The final inequality follows from the assumptions made on the error terms (see Appendix C). Intuitively, comparing individuals in occupation 1 who previously selected into training 1 to individuals in occupation 1 who previously selected into training 2 will result in estimates of τ which are upward biased, since those with higher ability in occupation

[14] In the log wage equation, $\tau > 0$ corresponds to the parameter on a dummy variable $D_{j=k}$ equal to one if training j is the same as occupation k.

[15] Further assume that these are jointly normally distributed with mean zero, standard deviation $\sigma_{\epsilon_1} = \sigma_{\epsilon_2}$, and $\sigma_{\epsilon_1\epsilon_2} = 0$. For notational simplicity, other subscripts have been omitted.

1 will have chosen it as a training. Under the given assumptions, selection into training will thus lead to *positive* bias when estimating parameter τ.

4.1.2 Selection into Occupations

Now consider a second hypothetical experiment where individuals are randomly allocated to a training, and can subsequently choose their occupation. Again focusing on occupation 1, the selection bias when estimating parameter τ may be written as

$$\begin{aligned} & E[\epsilon_{i1}|train_{i1}=1, occ_{i1}=1] - E[\epsilon_{i1}|train_{i1}=0, occ_{i1}=1] \\ & \quad = E[\epsilon_{i1}|(occ_{i1}=1|train_{i1}=1)] - E[\epsilon_{i1}|(occ_{i1}=1|train_{i1}=0)] \\ & \quad = E[\epsilon_{i1}|\underbrace{(\epsilon_{i1}-\epsilon_{i2})>-\tau}_{\substack{\text{choose occupation 1} \\ \text{cond. on training 1}}}] - E[\epsilon_{i1}|\underbrace{(\epsilon_{i1}-\epsilon_{i2})>\tau}_{\substack{\text{choose occupation 1} \\ \text{cond. on training 2}}}] \leq 0, \end{aligned} \tag{10}$$

where the difference in observed log wages net of τ has been normalized to zero. As before, the final inequality follows from the assumptions made on the error terms (see Appendix C). Workers who choose occupation 1 conditional on having been randomly allocated to training 1 in $t = 0$ are positively selected. These workers realize the benefit τ from working on the diagonal by choosing occupation 1. In contrast, workers who have previously been allocated to training 2 would realize the benefit of working on the diagonal by choosing occupation 2. The fact that they nonetheless choose occupation 1 implies they are more positively selected than their co-workers who trained in occupation 1. Intuitively, those working off the diagonal after random allocation to a training must be very productive in their chosen occupation as their ability needs to compensate for the lack of training. Under the given assumptions, selection into occupations will therefore lead to *negative* bias when estimating parameter τ.

4.2 Assumptions

The model in Section 3 is one of multiple unordered treatments. In addition, it features essential heterogeneity as individuals select into treatment based on knowledge of their idiosyncratic returns ϵ_{ikrt}. Identification in these models is complicated by a number of factors. For expositional clarity, I will simplify the model notation compared to Section 3, and discuss identification in the binary case, before moving to the case with multiple treatments.

4.2.1 Binary case

Consider the wage equation from Section 3 in a potential outcome framework. The notation is simplified in the following way: Y_k denotes the potential log wage in occupation k, X_k

denotes covariates in the outcome equation, and ϵ_k is the occupation-specific error term. Consider a model with only two occupation choices, k and k', where the training choice j is suppressed in the notation. Potential outcomes may be written as:

$$Y_k = \beta' X_k + \epsilon_k$$
$$Y_{k'} = \beta' X_{k'} + \epsilon_{k'}.$$

Selection into k, denoted by the dummy variable $occ_k = 1$, occurs if and only if:

$$e_k - e_{k'} \geq \tilde{U}_{k'} - \tilde{U}_k, \tag{11}$$

where, compared to the notation in equation 3 in Section 3, subscripts are omitted as above. The probability of selecting into occupation k is denoted by $P(occ_k = 1|X)$, where X is the vector of all characteristics, X_k and $X_{k'}$. The parameter of interest is the ATE of occupation k over k', $E[Y_k - Y_{k'}|X]$.

I make the following assumptions:

A 1 *There is a binary variable z_k that is an element of vector $X_{k'}$, but not X_k,*[16]

A 2 *z_k is conditionally independent of potential outcomes, $z_k \perp\!\!\!\perp Y_k|X_k$,*

A 3 $P(occ_k = 1|X, z_k = 0) \neq P(occ_k = 1|X, z_k = 1)$.

Assumptions 1 - 3 correspond to the well-known IV assumptions (exclusion, independence, relevance). Note that the variable z_k differs from standard "cost shifter" instruments, as these would affect $\tilde{U}_k$, not $\tilde{U}_{k'}$. Instead, there are sector-specific covariates that secure identification (d'Haultfœille & Maurel (2013)).[17] Further note that, given the additive separability of monetary and non-monetary components in $\tilde{U}_k$, the threshold-crossing nature of the selection model implies a monotonicity assumption (Vytlacil (2002)). A particular change in z_k will move all individuals either into or out of occupation k. Taken together, A1- A3 allow for identification of the Local Average Treatment Effect (LATE), $E[Y_k - Y_{k'}|X, occ_k(z_k = 1) = 1, occ_k(z_k = 0) = 0]$ (Imbens & Angrist (1994)).

In contrast to the LATE, identification of the ATE requires a large support assumption, also referred to as identification at infinity (IAI) step, even in the binary case (Chamberlain (1986), Heckman (1990)). In particular, there needs to exist some value of z_k such that $E[\epsilon_k|X, z_k] = 0$. Intuitively, the instrument needs to have enough support so that, for some value of z_k, the selection goes to zero.

[16]Note that z_k is binary for expositional purpose only.

[17]This type of identification strategy is common in the IO literature where characteristics of competing products are used as instruments in demand estimation (Berry *et al.* (1995)).

4.2.2 Multinomial case

Additional complications for identification arise in the multinomial case. Consider a model with three instead of two occupations (k, k', k''). As before, consider a variable z_k that satisfies assumptions A1-A3. Monotonicity still holds as switching on z_k will move all individuals either into or out of occupation k. The issue in the multinomial setup is that, aside from moving individuals into or out of occupation k, changes in z_k will also move individuals between occupations k' and k''. In order to separate these flows, one needs to again rely on large support assumptions.[18] Intuitively, as the value of occupation k'' becomes arbitrarily small, a shift in z_k will only lead to changes in the choice between k and k'. This argument naturally extends to the case with more than three choice alternatives.

An alternative way of thinking about this problem is in terms of the difference in indeces $\tilde{U}_{k'k} = \tilde{U}_{k'} - \tilde{U}_k$ from equation 11. An exogenous shift in this difference, holding all other utility contrasts fixed, would identify treatment effects of occupation k versus k'. Since $\tilde{U}_{k'k} = \tilde{U}_{k'k''} - \tilde{U}_{kk''}$, such a shift is impossible. Heckman *et al.* (2008) make this argument in the context of an exogenous "cost shifter" s_k that affects the own index $\tilde{U}_k$, but is not included in other indeces or the outcome equation. Even in this case, identification of treatment effects of alternative k versus k' cannot proceed without additional assumptions.[19] This is because changes in s_k affect any utility contrast that contains $\tilde{U}_k$, and so individuals are drawn into or out of option k from different second-best alternatives, not necessarily k'.

In summary, identification of the ATE of occupation k versus k' in the given context thus requires strong support for two reasons. Since there are more than two choice alternatives, one needs to invoke an IAI step to identify the LATE of option k versus k'. In addition, as in the binary case, an IAI step is required to recover the ATE.[20] I will discuss the strong support requirement in Section 5, after discussing assumptions A1-A3 in the given context.

4.2.3 Given context

The empirical part of this paper uses apprenticeship vacancies (henceforth *vacancies*) in outside options $vac_{k'rt}, \forall k' \neq k$, as instruments z_k. Since individuals select into trainings and occupations, an instrument is needed for each of these choices. I address this challenge by splitting vacancies into expectations and shocks, where the idea is that expected vacancies will affect the choice across trainings (as individuals compare the expected payoffs), and

[18] Note that this requirement is in addition to the IAI argument needed to identify ATEs instead of LATEs.

[19] Alternatively, one can rely on additional data to achieve identification. This is done by Kirkebøen *et al.* (2016) who consider the returns to choosing one field of education versus a particular next-best alternative using rich information on individuals' next-best alternatives.

[20] See Heckman *et al.* (2008) for details on this argument.

shocks to these expectations in the current time period will affect the occupation choice. Section 5 discusses how vacancies are split empirically.

For now, denote by vac_{krt} the vacancies in occupation k in region r at time t, and by vac_{jrt} the vacancies in region r at time t in the occupation that training j is training in. Define the expected vacancies for occupation k at time t of an individual deciding on a training in region r_0 at time t_0 as $E[vac_{kt}|\Omega_{r_0t_0}]$, where $\Omega_{r_0t_0}$ summarizes the individual's information set.[21] Vacancies at time t are given by $vac_{krt} = E[vac_{kt}|\Omega_{r_0t_0}] + (vac_{krt} - E[vac_{kt}|\Omega_{r_0t_0}])$, where the latter term is the shock to vacancies relative to the expectation of an individual making a training choice in region r_0 at time t_0. Using the above, I formally define the set of instruments for the training and occupation choices j and k, respectively, as:

$$z^j_{r_0t_0,j'(t_0+\tau)} = E[vac_{j'(t_0+\tau)}|\Omega_{r_0t_0}] \ \forall j' \neq j, \ \forall \tau = 0, ..., 30, \tag{12}$$

$$z^k_{r_0t_0,k'rt} = (vac_{k'rt} - E[vac_{k't}|\Omega_{r_0t_0}]) \ \forall k' \neq k. \tag{13}$$

The instruments for a training choice j in equation 12 are the predictions up to 30 years ahead of vacancies in occupations other than the one that training j is training in. The instruments for occupation choice k at time t in equation 13 are the shocks to vacancies in occupations other than k, relative to what was expected at the time of training choice.[22]

Note that for an individual making a training choice in region r_0 at time t_0,

$$vac_{k'rt} = z^j_{r_0t_0,j't} + z^k_{r_0t_0,k'rt}, \text{ for } j = k, \forall k' \neq k,$$

i.e. for $j = k$, the instruments for training choice j and occupation choice k sum to vacancies in the other available options. As discussed in Section 4.2.1 and 4.2.2, identification relies on $vac_{k'rt}$ to satisfy assumptions A1-A3. I will discuss these in turn, moving to a discussion of the large support requirement in Section 5.

A1-A2 Exclusion and conditional independence. The key identifying assumption equivalent to A1 (exclusion) and A2 (independence) in the given context is that occupation-specific vacancies at time t are a sufficient statistic for random changes to occupation-specific labor demand at time t, and that demand in occupation k is perfectly elastic.

To assess the plausibility of assumptions A1 and A2, it is important to recall that I use vacancies in outside options, not the option itself, as instruments. For example, A1 requires that *conditional on vacancies for craft workers*, vacancies for *electrical workers* are excluded

[21] Note that individuals' expectations do not vary across regions.

[22] Given the 13 training/occupation categories, there are $12 \times 31 = 372$ instruments for each training choice and 12 instruments for each occupation choice.

from the wage equation for *craft workers*. Assumption A2 requires that vacancies are a sufficient statistic for occupation-specific labor demand since vacancies for *electrical workers* may otherwise be correlated with the unobserved part of wages for *craft workers*.

A potential threat to assumptions A1 and A2 are confounding effects through labor supply. Since labor supply across occupations is linked through worker self-selction, if vacancies reflected occupation-specific supply instead of demand, vacancies for *electrical workers* would be correlated with vacancies for *craft workers*, which could in turn lead to confounding impacts in the wage equation. It is therefore important that vacancies reflect changes in labor demand. Recall that the empirical measure of vacancies corresponds to the total number of *apprenticeship* vacancies posted in a particular year by training firms, regardless of whether they become filled or remain unfilled at the end of the year. I find that these vacancies are positively correlated with non-apprenticeship job ads. However, since apprenticeship vacancies are administered separately from job ads for other applicants in firms, they are unlikely to be affected by changes in the occupation-specific supply of workers. In addition, the majority of vacancies is posted to match up with the school leaving date, so the measure does not react to supply fluctuations of apprentices in the same year. This implies that vacancies likely serve as proxy for occupation-specific demand in the given context.

Even if vacancies reflect occupation-specific changes in demand, assumption A1 could be violated if demand were not perfectly elastic, since general equilibrium effects through labor supply may lead to an impact of vacancies for *electrical workers* on wages for *craft worekrs*. While it is difficult to rule out such feedback effects, empirical support against them may be provided by using the fact that supply-side responses would be expected not to impact wages with a time lag. I therefore show that my results are robust to including occupation *times* time fixed effects in the baseline model (see Table A.7). The fact that the results are quantitatively very similar when only using variation within an occupation-time cell suggests that supply feedback effects are not a major concern in the present setting.

Moving to assumption A2, a key requirement is that vacancies at time t are a sufficient statistic for occupation-specific labor demand at time t. If this were not the case, it could be that, conditional on vacancies for *craft workers*, vacancies for *electrical workers* are correlated with the unobserved part of wages for *craft workers*. Most obviously, this would happen if productivity shocks were industry-specific, and variation in occupation-specific vacancies only captured part of the shock that affects wages in a particular occupation. To address this concern, I show that my results are robust to the inclusion of industry *times* time fixed effects in the estimation (see Table A.7). By only using variation within a particular industry, this specification gives the instruments a Bartik interpretation where productivity shocks (the shift) affect all occupations within an industry and identification comes from the

different occupational composition (the share) within each industry. The fact that the results are quantitatively very similar supports the assumption that occupation-specific vacancies serve as a sufficient statistic for occupation-specific labor demand.[23]

An additional concern regarding assumption A2 would be non-random changes in vacancies with respect to the outcome. To rule out strategic vacancy setting by firms (e.g. firms set fewer vacancies for *electrical workers* to increase the supply of *craft workers*), each firm needs to be sufficiently small relative to the market. This is true empirically where around three quarters of apprentices are trained in small and medium-sized firms.[24] On the worker side, conditional random assignment rules out systematic relocation of individuals as a result of labor market conditions. For instance, individuals who are particularly able in a specific occupation could choose to move to a state with a high number of vacancies in a given year. Empirically, however, mobility is low. On average, over 93% of apprentices start their apprenticeship in their state of residence.[25] Moreover, only about $10-15\%$ of all occupation changes in the sample correspond to changes of the region in which the employer is located. Section 6.5 provides a robustness check excluding these spells from the sample, showing that the results are almost unchanged by this restriction.

A3 Relevance. The relevance or first stage assumption states that the set of instruments needs to be sufficiently related to the training and occupation choices. In the context of categorical endogenous variables, a natural way of assessing the relevance assumption is to look at the variation in selection probabilities generated by the instruments (e.g. Hull (2018)) (see Section 5.3 for details on the estimation of the selection probabilitities). Figures B.3 and B.4 show histograms of the selection probabilities into the five largest trainings and occupations. It can be seen that, for both trainings and occupations, there is considerable variability in the selection probabilities, indicating a substantial degree of first stage variation in the sample. At the same time, selection probabilities do not typically reach extreme values of 0.9 or above. This suggests that a fully non-parametric estimator will be infeasible in the given context, justifying the additional structure imposed in the estimation (see Section 5.2).

[23] As a complementary check, I also examine the pairwise correlations of vacancies across all occupations and do not find a pattern of higher correlations across vacancies for occupations that belong to arguably similar industries. For example, while the two main occupations belonging to the manufacturing sector (*craft workers* and *process and plant workers*) display a vacancy correlation of 0.24 across the sampling period, the correlation between vacancies for *electrical workers* and *sales and financial workers* is 0.88.

[24] Around 50% are trained in small firms with less than 50 employes, 23% are trained in medium-sized firms with more than 50 and less than 250 employees (see Figure B.2). Source: *Bundesagentur für Arbeit.*

[25] Population-weighted average across states in 2016. Source: *Datenreport zum Berufsbildungsbericht 2016.*

5 Estimation

A common approach to estimate ATEs in selection models with essential heterogeneity is a control function estimator. The general idea behind this method is to model the endogenous component of the regression error term using its dependence on the instruments and control for it in the estimation. In the present context, define $\lambda_{jk}(...)$ as the appropriate control function using equation (8):

$$\lambda_{jk}(...) = E[\epsilon_{ikrt}|train_{ij} = 1, occ_{ij(k|j)rt} = 1], \tag{14}$$

where $\lambda_{jk}(...)$ depends on a set of variables further defined below.

Based on the above control function, we can provide further intuition for the assumptions required for identification of the LATE from Section 4. Intuitively, the instruments ensure that $\lambda_{jk}(...)$ can vary in a sufficiently independent way from the expected outcome, and this variation identifies the expected outcome up to a location parameter (d'Haultfœille & Maurel (2013)). For separate identification of the intercept of expected wages and of the control function, an IAI step or distributional assumptions are necessary (see Section 4.2).

Standard parametric control function approaches are computationally infeasible in high-dimensional settings. For instance, a two-step Heckman (1979) estimator would require the integration of a $(J \times K)$-fold integral over the joint distribution of outcome and selection error terms. To address this challenge, my approach is to reduce the dimensionality of the selection problem by using assumptions on the joint distribution of the outcome and selection errors. This control function method builds on Lee (1983) and Dahl (2002), and extends their insights in settings with high-dimensional selection to a case with two selection stages.

5.1 Reduction of Dimensionality

Below, I briefly outline how the Lee (1983) and Dahl (2002) approach can be extended to reduce the dimensionality of the given selection problem. Details can be found in Appendix D. As a first step, Lee (1983) shows that the selection rules defined by equations (3) and (6) may be re-written in terms of maximum order statistics:

$$train_{ij} = 1 \quad \text{iff} \quad \max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq 0, \tag{15}$$

$$occ_{i(k|j)rt} = 1 \quad \text{iff} \quad \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq 0. \tag{16}$$

The control function defined by equation (14) will depend on the conditional joint distribution of the outcome error ϵ_{ikrt} and the two maximum order statistics, where the conditioning is on the set of observed utility and value function differences $(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt}), \forall k' \neq k$, and $(\tilde{V}_{ij'krt} - \tilde{V}_{ijkrt}), \forall j' \neq j$. Since there exists a one-to-one mapping between these observed utility and value function differences, and selection probabilities, the joint distribution can instead be conditioned on the vector of selection probabilities. Based on this result, it has been shown that control functions in single-index models may be written as a function of the probability of selection only (Heckman & Robb (1985); Ahn & Powell (1993)). In the present multiple-index setting, equation (1) may be written as

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \tau_{jk} + \lambda_{jk}(p_{i1r_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) + u_{ikrt}, \tag{14}$$

where $p_{ijr_0t_0}$ is the probability of selecting into training j, $p_{i(k|j)rt}$ is the probability of selecting into occupation k conditional on training j, and u_{ikrt} is a mean zero error term.

Note that, given the sequential nature of the selection problem, the functions $\lambda_{jk}(...)$ depend only on those occupation probabilities $p_{i(k|j)rt}$ that condition on the observed training choice j. Nonetheless, estimating this equation fully flexibly would require a flexible function in $(J + K)$ probabilities to be included in $(J \times K)$ control functions which is infeasible.

Following Lee (1983) and Dahl (2002), I impose a distributional assumption to further reduce the dimensionality of the problem. In its strongest form, this assumption imposes a specific functional form for the joint distribution of outcome errors and maximum order statistics. As further discussed in Section 5.2, the estimation of parameters τ_{jk} from equation (1) relies on such a parametric distributional assumption.[26] On the other hand, some parameters in the heterogeneity analysis may be estimated non-parametrically, based on a weaker so-called index sufficiency assumption (Dahl (2002)).[27] I make use of this result and provide robustness checks where I use both the parametric and the non-parametric control function to estimate the slope parameters (see Section 5.2).

5.2 Distributional Assumptions

As outlined in Section 4, identifying the jk-specific intercepts from equation (1) separately from the intercepts in the control function relies on strong support requirements. Intuitively,

[26] The parametric assumption replaces the IAI step discussed in Section 4.

[27] The index sufficiency assumption states that all information about the joint distribution of outcome errors and maximum order statistics is summarized by a small set of selection probabilities. The parameters that can be estimated using this assumption are those that vary within a particular jk-cell. This includes the differences in parameters τ^{exp}, i.e. the slope of the on- versus off-diagonal effect by experience.

the instruments need to have enough support so that for some values, individuals select into a specific group with probability close to one. At these values of the instruments, the selection bias goes to zero and OLS consistently estimates the ATE. In practice, it may be difficult to meet such strong support requirements. This is especially true in settings with a large number of choice alternatives. An alternative to this approach is to make parametric assumptions on the distribution of the outcome and selection errors. Intuitively, this method extrapolates the selection probabilities by imposing a functional form assumption in order to separately identify the parameters of interest from the intercepts of the control function. I follow the approach by Lee (1983) to simplify the estimation and impose standard normality assumptions for the relevant distributions.[28] The control function is then given by the well-known function of the inverse Mill's ratio (Heckman (1976, 1979)):

$$\lambda_{jk}(p_{i1r_0t_0}, ...p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) = -\rho_{jk}\frac{\phi[\Phi^{-1}(p_{ijkrt})]}{p_{ijkrt}}, \tag{19}$$

where $p_{ijkrt} = p_{ijr_0t_0} \times p_{i(k|j)rt}$ is the probability of selecting into the *observed* training-occupation cell jk, ρ_{jk} is the correlation between the outcome error and a random variable constructed as part of the Lee (1983) approach, and $\phi(.)$ and $\Phi(.)$ denote the standard normal probability density and cumulative density function, respectively. Given consistent estimates for the selection probabilities p_{ijkrt}, the parametric approach may be implemented by evaluating the inverse Mill's ratio at these estimates, and including an interaction of this expression with selected jk-cells in the outcome equation.[29] Log wages in equation (1) may then be written as

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \tau_{jk} + \beta' X_{it} - \rho_{jk}\frac{\phi[\Phi^{-1}(p_{ijkrt})]}{p_{ijkrt}} + u_{ikrt}, \tag{20}$$

where $E[u_{ikrt}|train_{ij} = 1, occ_{ij(k|j)rt} = 1] = 0$.

Lee's (1983) approach makes estimation in high-dimensional selection problems feasible, but directly applying his transformation in the present setting simplifies the selection problem by abstracting from its sequential nature. This simplification implies that the same control function estimator would have been used in a static context with $(J \times K)$ choice alternatives even though the joint distribution of the outcome and transformed errors would likely have been different. Kline & Walters (2019) show that a wide class of control function estimators yield estimates of the LATE that are identical to non-parametric IV estimates, but the same does not necessarily hold for the estimation of ATEs which may be more sensitive to the

[28] See Appendix D for details on the Lee (1983) approach in the present setting.

[29] The selected jk-cells include all cells where $j = k$, and a cell for each occupation k where $j \neq k$.

choice of distributional assumptions.

To provide some justification for the distributional assumptions required for the estimation of the intercept parameters τ_{jk} in the present setting, I use the fact that the slope parameters may be estimated non-parametrically using the weaker index sufficiency assumption described in Section 5.1. A comparison of the estimates for these parameters from the parametric and the non-parametric approach thus serves as a robustness check for the distributional assumptions required to estimate the intercept parameters. As shown by the results of this comparison (see Section 6.2), the slope estimates from both approaches coincide almost exactly, lending support to the distributional assumptions made.

5.3 Estimating the Selection Probabilities

Implementing the control function approach requires consistent estimates for the selection probabilities $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$. To avoid imposing assumptions on the form of individual preferences, I choose a flexible approach for the estimation of selection probabilities, and use a machine learning algorithm (random forests) to predict the selection into trainings and occupations based on observables and the instruments discussed in Section 4.2.3.

To obtain the instruments, vacancies need to be split into expectations and shocks. To do so, I estimate separate linear time trend models in each region-time cell, where log vacancies for each occupation are explained using five years of previous data.[30] The details of this approach are described in Appendix D. Intuitively, individuals use five years of past vacancy data at the time of training choice to predict vacancies for 30 years into the future. Each individual gets assigned the predictions for vacancies in each occupation based on demographics, and the region and time they started the apprenticeship in. Subsequent vacancy shocks are defined at the individual level as the difference between realized vacancies and the prediction at the time of training choice.

In a second step, I predict training and occupation choices using a random forest. Random forests are among the most accurate classifiers available (Breiman (2001)). Besides avoiding functional form assumptions, they have the advantage of naturally allowing for the large number of independent variables in the present setting. The algorithm predicts variables $train_{ij}$ and $occ_{i(k|j)rt}$ using optimal splitting rules on the explanatory variables, and problems of overfitting are avoided through a bootstrapping procedure.

To account for sampling variation due to the estimation of selection probabilities in my analysis, I randomly select 50% of the individuals in the baseline sample to train the random

[30]Note that this implies that predictions and shocks will not be available during the first five years of the sample, 1978-1981. Moreover, due to regional classification changes following German reunification, data are not available for four regions for 1994-1997. This reduces the number of observations used in the estimation.

forest, and only use the remaining 50% of individuals and their probability predictions in the regression analysis. Details on the random forest algorithm as well as the implementation of the algorithm in the present context can be found in Appendix D. With the estimated probabilities at hand, the control function estimation proceeds by replacing the selection probabilties from Section 5.2 with their estimates $\hat{p}_{ijr_0t_0}$ and $\hat{p}_{i(k|j)rt}$.

6 Results

This section discusses the results for regression equation (1), where I parameterize τ_{jk} using different specifications to estimate returns to working on versus off the diagonal within occupations (Sections 6.1 and 6.2) and across occupations for each training (Sections 6.3 and 6.4). All results are based on the baseline sample, excluding observations used to train the random forest as well as years where the instruments are not available (see Section 5.3). Following Abadie *et al.* (2017), standard errors allow for clustering at both the region and time level. To generate meaningful averages, regression observations are weighted using the empirical training-occupation distribution in the most recent sampling year, 2010. In all baseline estimations, $f(vac_{krt})$ is approximated using a fourth order polynomial in vac_{krt}. Robustness using a higher order polynomial is provided in Section 6.5.

6.1 Average Return to Working On versus Off the Diagonal

This section reports and discusses the results for the outcome equation where $\tau_{jk} = \delta_k + \tau D_{j=k}$, and parameter τ captures the average on- versus off-diagonal return within occupations.[31] Table 4 shows the regression results. The main variable of interest is $D_{j=k}$ which is equal to one if the individual works on the diagonal. Columns (1) and (2) report the results from estimations that do not control for occupation-specific experience, columns (3) and (4) condition on exp_k.[32] Both specifications are estimated without controlling for selection (columns (1) and (3)), and using the control function estimator (columns (2) and (4)).

The results from column (1) show that working on the diagonal is associated with a small negative wage effect. This effect becomes more negative after controlling for exp_k (column (3)), which is in line with workers on the diagonal having more experience in their current occupation. When accounting for selection using the parametric control function

[31] Note that since all regressions contain individual fixed effects and individuals only complete one training, the within-occupation comparison here assumes that the effect of a particular training that is common to all occupations is the same across trainings.

[32] Note that, since occupation-specific experience corresponds to past selection into occupations, it may be an endogenous regressor. The given control function does not explicitly take this additional potential endogeneity into account. The results from columns (3) and (4) should therefore be taken with caution.

estimator (columns (2) and (4)), the effect of $D_{j=k}$ becomes positive and significant, implying a sizable negative selection bias of around 15 percentage points. The coefficients on the control function are highly significant, confirming the importance of the selection bias. The negative sign of the bias suggests that the selection into occupations dominates the selection into training in the given setup (see Section 4.1.2). Intuitively, workers on the diagonal may be negatively selected relative to workers off the diagonal as the latter must compensate for the lack of training with higher occupation-specific ability.

Table 4: Average On- versus Off-Diagonal Returns

	(1)	(2)	(3)	(4)
$D_{j=k} = 1$	-0.0126	0.1510^{***}	-0.0349^{***}	0.1355^{***}
	(0.0079)	(0.0193)	(0.0087)	(0.0214)
exp	0.0597^{***}	0.0593^{***}	0.0431^{***}	0.0432^{***}
	(0.0022)	(0.0023)	(0.0025)	(0.0025)
exp^2	-0.0010^{***}	-0.0010^{***}	-0.0004^{***}	-0.0004^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
exp_k			0.0182^{***}	0.0176^{***}
			(0.0013)	(0.0014)
exp_k^2			-0.0008^{***}	-0.0008^{***}
			(0.0001)	(0.0001)
Indiv. FE	yes	yes	yes	yes
Occ./Reg./Time FE	yes	yes	yes	yes
Parametric cf	no	yes	no	yes
p-value cf		0.000		0.000
N	1,140,518	1,140,518	1,140,518	1,140,518

Notes: The table reports regression results for equation (1) with $\tau_{jk} = \delta_k + \tau D_{j=k}$. Standard errors (in parentheses) are clustered at the region and time level. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Results from column (2) suggest that working on the diagonal leads to a significant wage increase of 15.1%. This figure should be interpreted as the full effect of having received training in the current occupation, including potentially higher experience in that occupation that was accumulated as a result of the training. As before, controlling for occupation-specific experience lowers the effect of $D_{j=k}$ to about 13.6% (column (4)). Albeit smaller than the full effect, the results from column (4) suggest that most of the positive effect of working on the diagonal is due to the training itself, not the subsequent effect that training may have on the accumulation of occupation-specific experience. As outlined above, given the potential

endogeneity of occupation-specific experience, these results should however be taken with caution. As a result, only the full-effect specification corresponding to columns (1) and (2) will be used in the heterogeneity and full-matrix analysis that follows, i.e. regressions will control for total work experience *exp* but not for occupation-specific work experience exp_k.

6.2 Heterogeneity by Experience

This section reports and discusses the results for the outcome equation where $\tau_{jk} = \delta_k + \tau^{exp} D_{j=k}$, which looks at the heterogeneity in on- versus off-diagonal returns across different levels of full-time work experience. Figure 3 plots separate coefficient estimates for τ^{exp}, where experience levels have been binned into yearly categories. Each coefficient compares workers with a specific level of full-time work experience who were trained in their occupation to workers with the same level of experience who were not trained in their occupation.

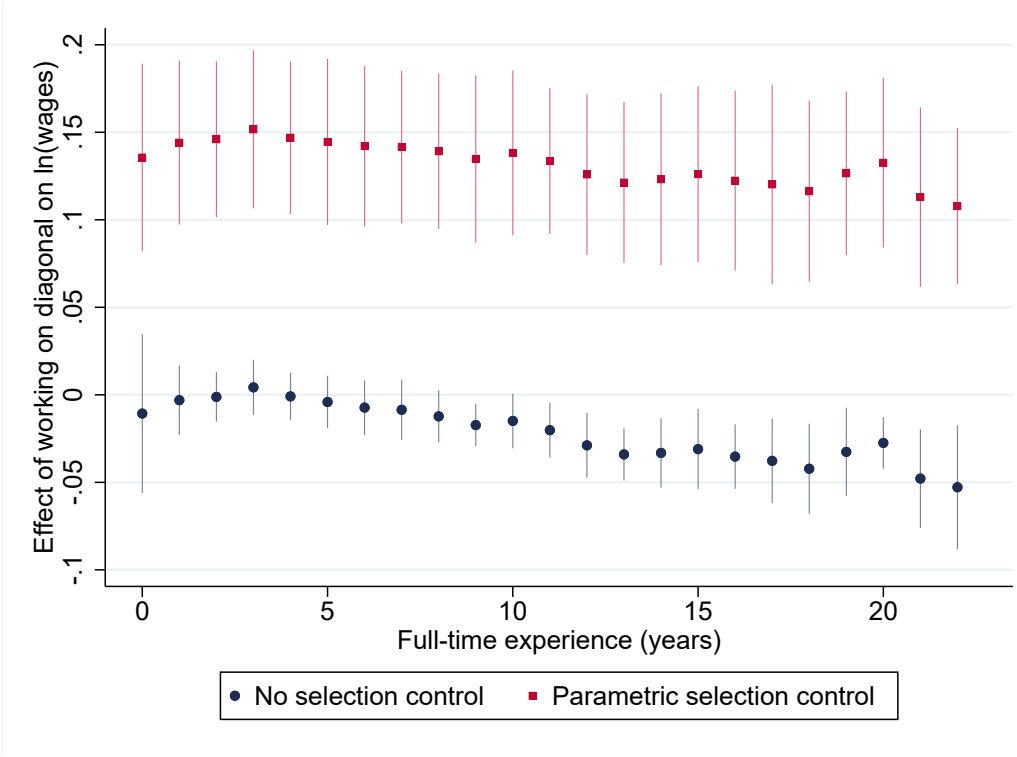
In line with the results from Table 4, Figure 3 shows that not controlling for selection leads to a sizable negative bias in the coefficient estimates. The set of control function estimates suggests that the effect of $D_{j=k}$ first increases slightly from around 13% to 15%, then falls to about 12% after 12 years of work experience where it stabilizes. While the early increase is in line with an initially stronger accumulation of occupation-specific work experience for on-diagonal workers, the subsequent decline suggests that off-diagonal workers partly catch up with their co-workers who received the relevant training. However, there is no full catch-up and sizable differences remain after 20 years of work experience.

Figure B.5 plots the same set of coefficients as Figure 3, adding the coefficients estimated using the non-parametric control function estimator described in Sections 5.1 and 5.2. Since the latter identifies the slope but not the level of the parameters of interest, all coefficients are normalized to zero at zero years of work experience. Comparing the set of coefficients estimated without selection control to those estimated with the parametric control function shows that not controlling for selection leads to an increasingly negative bias in the estimated coefficients, such that final levels are underestimated by about 2% more than those at low levels of work experience. This increase in bias is in line with workers receiving better information about their occupation-specific abilities over time.

Moreover, Figure B.5 shows that the normalized coefficients from the parametric and non-parametric control function estimator are almost identical. This lends support to the distributional assumptions made to implement the parametric approach.[33]

[33]Figure B.8 reports a similar comparison, adding on-diagonal probability terms to the non-parametric control function (see Dahl (2002)). The results provide further support to the assumptions made.

Figure 3: On- versus Off-Diagonal Returns by Experience



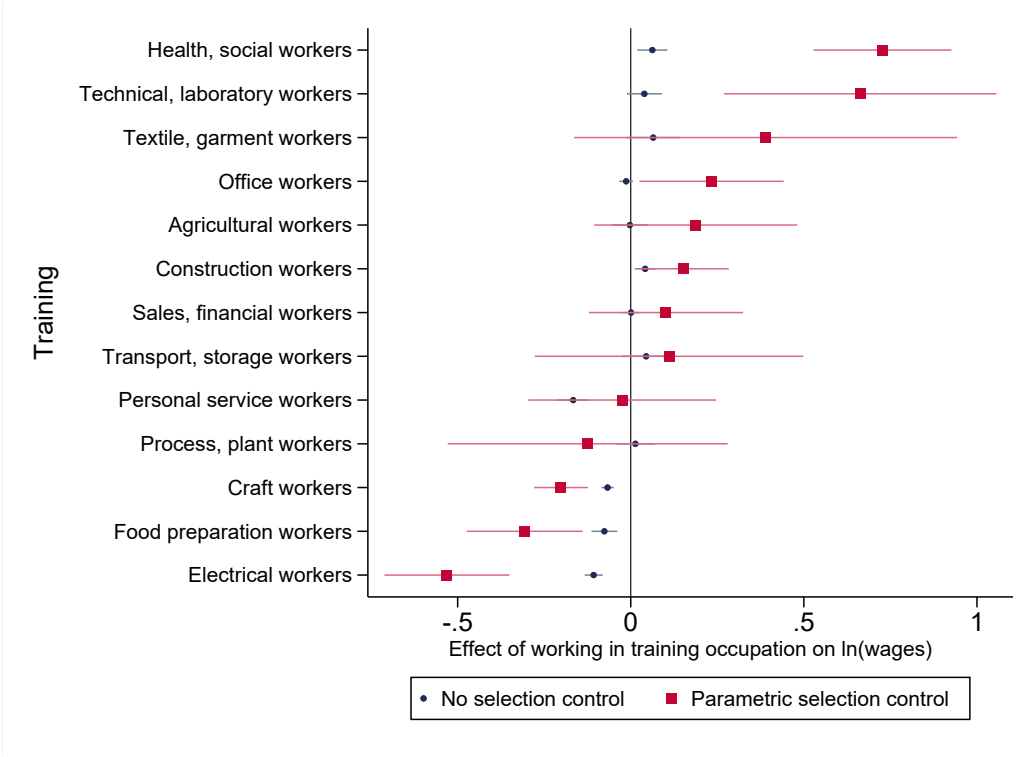
Notes: The figure plots regression coefficient estimates for τ^{exp} in a version of equation (1) with $\tau_{jk} = \delta_k + \tau^{exp} D_{j=k}$, where experience levels have been binned into yearly categories. Standard errors are clustered at the region and time level. 95% confidence intervals are shown.

6.3 Heterogeneity by Training

This section presents and discusses the results for the outcome equation where $\tau_{jk} = \tau_j D_{j=k}$, which explores the on- versus off-diagonal returns for each training. Note that this model aims at estimating the parameters that determine the occupational selection conditional on each training and therefore does not contain occupation fixed effects. Figure 4 plots the coefficients τ_j estimated with and without the parametric control function. The coefficient estimates are highly heterogeneous and, out of the five largest trainings, *health and social workers* have the highest, *craft workers* the lowest return to working in their training occupation. Given the regression specification, negative coefficient estimates $\hat{\tau}_j$ can be rationalized by other occupations $k \neq j$ providing better opportunities, regardless of the training.

Since workers choose their occupations taking into account the return to working on versus off the diagonal, the heterogeneity in these returns across trainings should affect the fraction of individuals choosing to work on the diagonal ex-post. Conditional on training choice, the Roy model predicts that more workers will select onto the diagonal, the higher

Figure 4: Average On- versus Off-Diagonal Returns by Training

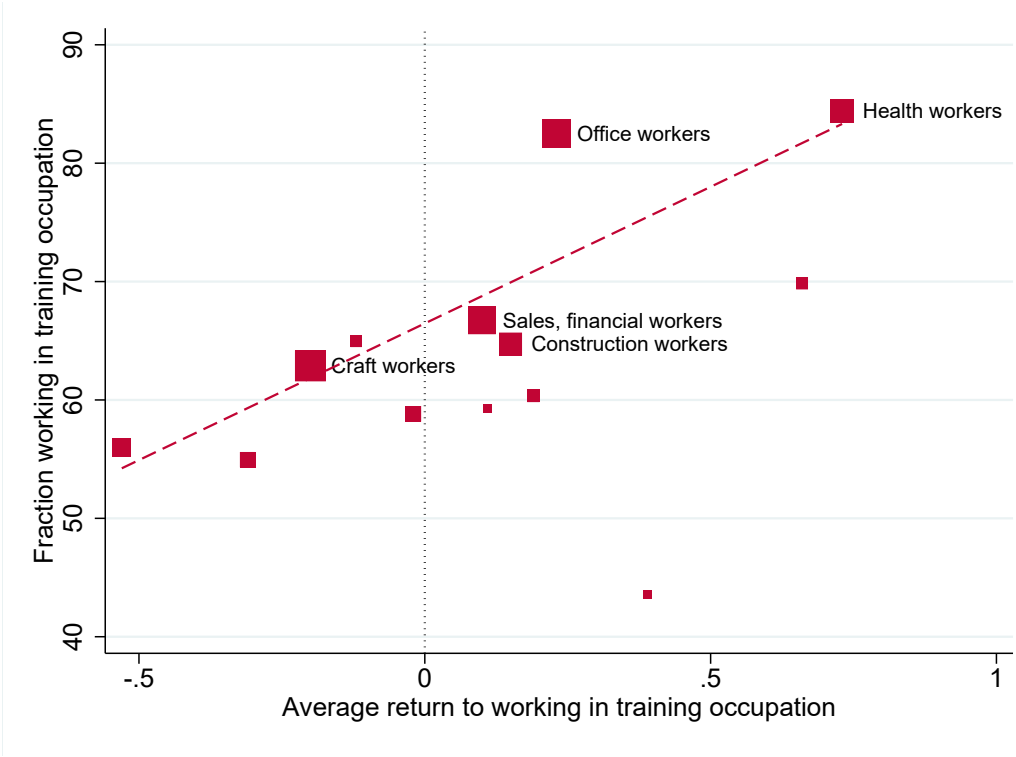


Notes: The figure shows regression coefficient estimates for τ_j in a version of equation (1) with $\tau_{jk} = \tau_j D_{j=k}$ for each training. Standard errors are clustered at the region and time level. 95% confidence intervals shown.

the on- versus off-diagonal return in that training. Figure 5 explores this relationship by plotting the average return to working on the diagonal from Figure 4 together with the fraction working on the diagonal for each training. The positive slope is consistent with the Roy model predictions outlined above, and suggests that relative returns are an important determinant of the selection into occupations.

A further implication of the Roy model is that heterogeneity in the average return to working on versus off the diagonal across trainings is related to the size of the selection bias. As outlined in Section 6.1, the strong negative bias in average returns suggests that the selection into occupations dominates the selection into training, and that off-diagonal workers compensate for their lack of training with higher occupation-specific ability. As a result, one may expect a negative correlation between the on-diagonal return and the estimated selection bias across trainings. The higher the return to working on the diagonal for a particular training, the more occupation-specific ability outside the training is required to work off the diagonal, so returns in high-return trainings will be more strongly underestimated when not controlling for selection. Figure B.6 confirms this by showing a negative correlation between the estimated return and the selection bias across trainings.

Figure 5: Average Return and Fraction Working in Training Occupation



Notes: The figure plots average on- versus off-diagonal returns for each training from Figure 4 against the fraction of individuals working in their training occupation. The fitted line corresponds to a weighted OLS regression using the sample fraction in each training as weights. Marker size is proportional to the weights.

6.4 Full Training-Occupation Matrix

This section reports and discusses the results for the fully parameterized outcome equation which contains separate parameters τ_{jk} for all cells in the training-occupation matrix. Table 5 shows the results using the parametric control function estimator for the five largest occupations. Tables A.5 and A.6 report the results for all coefficients, with and without selection control. As outlined before, the inclusion of individual fixed effects in the estimation implies that all coefficients should be interpreted relative to the diagonal within the same training.

In line with the positive on- versus off-diagonal returns for four out of five of the largest trainings in Figure 4, Table 5 shows that, with the exception of training as a *craft worker*, most coefficients are negative suggesting that individuals incur wage penalties when working outside their training occupation. Nonetheless, there is considerable heterogeneity in the magnitudes of off-diagonal returns across trainings. For instance, the results suggest that while trained *office workers* incur moderate penalties when working in a different occupation, much larger penalties are incurred by trained *health and social workers*, with estimates

ranging between −0.74 and −0.48 log points. Table 5 also shows that returns are highly asymmetric. While trained *office workers* incur sizable penalties when working as *craft workers*, trained *craft workers* receive wage gains when working as *office workers*. In line with this finding, the fact that *all* trainings incur penalties when working as *craft workers* suggests that craft occupations provide worse opportunities, regardless of the training (see Section 6.3). Similar to Figure 5, Figure B.7 plots the estimated off-diagonal returns against the fraction of individuals choosing to work in the relevant occupation conditional on their training. Albeit noisier than Figure 5, the positive correlation confirms the importance of returns in determining the selection into occupations.

While hard to interpret individually, the estimates from Table 5 provide an opportunity to study a mechanism underlying the results presented in this paper. In Section 7, I use data on the task content of occupations to explore the heterogeneity in estimated returns, thereby providing a microfoundation for the empirical findings.

Table 5: Full Matrix of Returns - Within-Training Comparisons

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	0	−0.41*	−0.06	−0.18	−0.12
			(0.20)	(0.09)	(0.16)	(0.14)
	Craft workers	0.21***	0	0.37***	0.26**	0.28**
		(0.04)		(0.06)	(0.09)	(0.11)
	Sales, financ. w.	−0.10	−0.01	0	−0.02	0.02
		(0.10)	(0.16)		(0.11)	(0.15)
	Health, social w.	−0.73***	−0.74***	−0.48***	0	−0.72***
		(0.10)	(0.10)	(0.11)		(0.18)
	Construction w.	−0.10	−0.14*	0.02	−0.04	0
		(0.06)	(0.07)	(0.06)	(0.14)	

Notes: The table shows regression coefficient estimates for τ_{jk} in equation (1), estimated using the parametric control function estimator. Results are shown for the five largest occupations. Table A.6 shows the full set of coefficients. Standard errors (in parentheses) are clustered at the region and time level. Given the low number of clusters, critical values of the t(9)-distribution are used. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

6.5 Robustness

Table 6 provides robustness checks for the main result in column (2) of Table 4 by restricting the sample in a number of different ways.[34] Column (1) excludes employment spells where workers change the region in which their employer is located (see Section **??**); column (2) restricts the sample to individuals with an apprenticeship length between two and a half and three years; column (3) excludes wages that could potentially be capped in the dataset (see Section 2.5); column (4) excludes individuals who switched firms during their apprenticeship; column (5) excludes all spells where workers were employed in their apprenticeship firm. All results are obtained using the parametric control function estimator.

Table 6 shows that the effect of working on versus off the diagonal is positive and significant in all columns, with most results being quantitatively very similar to the main sample estimate of 15.1%. Columns (1)-(4) all report coefficient estimates of around 15%. While column (1) alleviates concerns regarding the conditional random assignment assumption (see Section **??**), columns (2)-(4) suggest that differences in the length across apprenticeships, the presence of institutional wage caps in the data or the sample of apprentices switching firms during their training are not driving the main result. The point estimate is slightly lower at 10.5% in column (5). While this is partly due to the fact that the main coefficient is lower at higher levels of experience, and that spells in apprenticeship firms are concentrated early in a worker's career, it also points to potential complementarities of working both in the occupation and the firm one has been trained in. At the same time, column (5) shows that such complementarities are small relative to the effect purely associated to the training.

7 Task Content

This section provides a microfoundation for the results presented in this paper by drawing on the literature on the task content of occupations. The task approach considers tasks as inputs to production, and skills as the human capital required to carry out these tasks (Autor (2013), Autor & Handel (2013)). Occupations, as discrete classification units, correspond to vectors of tasks that are carried out by workers.[35]

Based on this concept, it is possible to construct measures of task distance between occupations. Poletaev & Robinson (2008) and Gathmann & Schönberg (2010) argue that, if

[34] Further robustness results focusing on the estimation method can be found in Table A.7.

[35] More generally, task vectors can be carried out by labor or capital and changes in relative prices may lead to changes in the allocation of tasks to labor or capital (Acemoglu & Autor (2010), Autor (2013)). I focus on the task vectors which are carried out by workers within each occupation.

Table 6: Average On- versus Off-Diagonal Returns - Sample Restrictions Robustness

	(1) no movers	(2) app. length $2.5-3$ years	(3) no capped wages	(4) no app.-firm-switchers	(5) no spells in app. firm
$D_{j=k}=1$	0.1492^{***}	0.1461^{**}	0.1434^{***}	0.1558^{***}	0.1052^{***}
	(0.0200)	(0.0481)	(0.0189)	(0.0229)	(0.0245)
exp	0.0595^{***}	0.0552^{***}	0.0599^{***}	0.0594^{***}	0.0569^{***}
	(0.0023)	(0.0027)	(0.0024)	(0.0023)	(0.0024)
exp^2	-0.0010^{***}	-0.0009^{***}	-0.0010^{***}	-0.0010^{***}	-0.0009^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Indiv. FE	yes	yes	yes	yes	yes
Occ./Reg./T. FE	yes	yes	yes	yes	yes
Parametric cf	yes	yes	yes	yes	yes
p-value cf	0.000	0.000	0.000	0.000	0.000
N	1,118,123	279,249	1,118,404	1,024,332	827,353

Notes: The table reports regression results for equation (1) with $\tau_{jk}=\delta_k+\tau D_{j=k}$. Each column restricts the baseline sample as indicated in the column header. Standard errors (in parentheses) are clustered at the region and time level. $^{*}p<0.1,^{**}p<0.05,^{***}p<0.01$.

human capital is task-specific, it should be more easily transferable across occupations that require a similar mix of tasks.[36] In the present context, these findings suggest an explanation for the heterogeneity in estimated returns in the training-occupation matrix. If workers are trained in a specific mix of tasks (equal to the mix of tasks performed in the occupation they are trained for), then one would expect the penalty in a different occupation to be larger, the more distant in terms of the task content the occupation is from the original training.

7.1 Measuring Task Distance

The measure of task distance is constructed using data from the German Qualification and Career Survey (GQS), a representative telephone survey of around 20.000 individuals conducted by the German Federal Institute for Vocational Training and Education (*Bundesinstitut für Berufsbildung - BiBB*). This data has been used to study the skill requirements

[36]Using samples of displaced workers, they find that wage penalties are larger the more distant the occupational switch after displacement. Yamaguchi (2012) sets up a structural model to formalize these findings. Similarly, Cortes & Gallipoli (2018) estimate a structural model and show that task difference is a significant component of the cost of switching occupations.

across occupations in Germany in a variety of different contexts (e.g. DiNardo & Pischke (1997), Spitz-Oener (2006) and Gathmann & Schönberg (2010)). For the present analysis, I use four survey waves that fall within the sampling period used in this study (1985/86, 1991/92, 1998/99 and 2005/06).

The survey records information on workers' occupations and asks them to pick from a list of tasks the ones they perform in their current occupation. A summary table of the tasks together with the percentage of individuals working in the two largest occupations (*office and craft workers*) performing these tasks is presented in Table 7.[37] An advantage of the GQS task data is that, unlike the Dictionary of Occupational Title (DOT) which is the primary source of task data in the US, it makes a clear distinction between tasks and skills.[38] As a result, the task measures in the GQS all refer to *activities* that are required in specific occupations (e.g. operate machines) as opposed to *capabilities* of workers which are required to carry these out (e.g. manual dexterity).

Table 7: List of Tasks and Fraction Performing

Task	Office workers	Craft workers
Cultivate	0%	1%
Manufacture, install or construct	4%	47%
Publish, present or entertain others	5%	0%
Serve or accommodate	6%	2%
Clean	8%	27%
Secure	9%	15%
Repair, renovate, reconstruct	9%	72%
Equip or operate machines	14%	66%
Nurse or treat others	14%	11%
Pack, ship or transport	25%	35%
Execute laws or interpret rules	25%	3%
Design, plan, sketch	38%	29%
Employ, manage personnel, organize, coordinate	38%	14%
Calculate or do bookkeeping	41%	6%
Research, evaluate or measure	47%	50%
Sell, buy or advertise	48%	21%
Teach or train others	51%	39%
Program	55%	25%
Correct texts or data	74%	9%

Notes: The table shows the average fraction of individuals indicating they perform the given task.

[37] To construct averages, observations in each wave are weighted using survey weights and subsequently combined giving equal weight to each wave. See Table A.8 for an equivalent list for all occupation categories.

[38] See Yamaguchi (2012) and Robinson (2018) for a recent discussion of the job measures in the DOT.

Following Gathmann & Schönberg (2010), I use the task data to construct a measure of distance between training j and occupation k, $Dist_{jk}$, which is based on the angular separation between task vectors j and k (see Appendix E for details).[39] The resulting distance measure ranges from zero to one, and is decreasing in the degree of overlap between the two task vectors (two orthogonal task vectors having distance one).

Excluding on-diagonal training-occupation cells where $Dist_{jk} = 0$, the empirical distance in the given setting varies between 0.01 and 0.46 with a mean of 0.23. When weighting training-occupation cells by their sample fractions, the mean distance drops to 0.19 suggesting that, on average, workers who leave their training occupation work in occupations which are more similar to their training than the average occupation. Figure B.9 explicitly shows the negative correlation between training-occupation distance and the fraction of workers in the relevant occupation. Tables A.9 and A.10 report the distance measure for the five most similar and five most distant training-occupation pairs, as well as for the five largest trainings and occupations.

7.2 Match Returns and Task Distance

I model the estimated returns to a training-occupation combination $\hat{\tau}_{jk}$ from Section 6.4 using the following simple specification:

$$\hat{\tau}_{jk} = \alpha + \beta Dist_{jk} + \eta_{jk}, \tag{21}$$

where $Dist_{jk}$ is the measure of task distance between training j and occupation k described in Section 7.1, standardized to have mean zero and standard deviation equal to one, and η_{jk} is a match-specific error term.

Table 8 presents the results for equation (21). Column (1) shows that higher task distance is significantly related to lower returns in training-occupation cells. Specifically, the results suggest that a one-standard-deviation increase in task distance is associated with a reduction in the return to the match of around 7*pp*, or more than 50% of the average $\hat{\tau}_{jk}$. To control for the fact that some occupations provide better opportunities regardless of training, column (2) includes occupation fixed effects in the regression. This slightly reduces the coefficient on $Dist_{jk}$ to 5.5*pp*. Columns (3) and (4) show the results for the same regression models where the left-hand-side returns τ_{jk} have been estimated *without* selection control. It can be

[39] Measuring similarity between two vectors by the angular separation was first proposed by Jaffe (1986, 1989a) who estimated R&D spillovers across technologically similar firms. Subsequently, a number of studies have used the measure in various contexts such as spillovers of university research to commercial innovation (Jaffe (1989b)), knowledge-relatedness in technological diversification (Breschi *et al.* (2003)) and similarity of tasks performed across occupations (Gathmann & Schönberg (2010), Cortes & Gallipoli (2018)).

seen that the effect of $Dist_{jk}$ is much smaller in magnitude and only marginally significant.

Tables A.11 and A.12 show that the above findings are robust to excluding on-diagonal observations where $\hat{\tau}_{jk} = 0$ and $Dist_{jk} = 0$, and even to restricting the sample to the five largest trainings and occupations. Overall, the results are in line with the proposed hypothesis that apprentices are trained to carry out a specific mix of tasks and their returns across occupations are lower, the less applicable the acquired skills are to that occupation.

Table 8: Match Returns and Task Distance

τ_{jk} estimated	with parametric control fcn.		without selection control	
	(1)	(2)	(3)	(4)
$Dist_{jk}$	-0.0738^{**}	-0.0545^{**}	-0.0172^{*}	-0.0026
	(0.0281)	(0.0243)	(0.0097)	(0.0079)
Occ. FE		yes		yes
Mean of $\hat{\tau}_{jk}$	-0.1277	-0.1277	-0.0138	-0.0138
N	169	169	169	169

Notes: The table reports regression results from equation (21). $Dist_{jk}$ is scaled by its standard deviation. Robust standard errors are reported. $^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$.

8 Welfare and Policy

The results from Section 6.1 suggest sizable penalties from working off the diagonal. Using the model from Section 3, this section explores the welfare losses from ex-post suboptimal training choices which are implied by these estimates. In line with the model, the reason why training choices may be suboptimal ex-post is imperfect information at the time of training choice. This means that new information about the labor market and own abilities may be revealed, causing workers to seek work in an occupation different from their training. In addition to off-diagonal workers, a second group is affected by the ex-ante lack of information. These are workers who are locked into their training, meaning that they would choose a different occupation in the absence of off-diagonal penalties but currently work on the diagonal as their payoff elsewhere is insufficient to compensate for the lack of training.

Section 8.1 quantifies the welfare loss for off-diagonal and locked-in workers. Section 8.2 considers a potential policy intervention, retraining programs. Importantly, such programs

target ex-post wage outcomes and do not require any assumptions on preferences beyond those given in Section 3. Back-of-the-envelope calculations suggest that retraining could be very effective in addressing the information friction in the present context.

8.1 Welfare Losses

Using the selection model from Section 3, this section looks at the partial equilibrium per-period welfare losses due to imperfect information at the time of training choice. These losses are equivalent to the gains from changing training to the optimal option ex-post. All calculations are based on the empirical estimates for the average on- versus off-diagonal returns within occupations (see Table 4 in Section 6.1). Total losses are computed as the product of loss per worker and share of affected workers.

Focusing first on off-diagonal workers, the loss due to ex-ante imperfect information is equivalent to the gain associated with being trained in the current occupation. Note that since the model does not allow for individual-specific heterogeneity in the welfare returns to matching a training to a particular occupation, this gain is given by τ.[40] My findings suggest that τ is about 15%.[41] Figure 2 in Section 2.5 shows that the average share of off-diagonal workers is about 40%. This implies that the welfare loss from off-diagonal workers amounts to 6% of wages per worker in the system.

For each locked-in worker, the welfare loss is bounded from above by the on- versus off-diagonal return τ. Intuitively, locked-in workers choose not to change their occupation given the on- versus off-diagonal return, so their gain from a more suitable training choice can be at most τ. Since occupation-specific abilities are unobserved, locked-in workers are not directly observed in the data. To estimate the share of these workers, I use variation in the on- versus off-diagonal return and in the fraction of on-diagonal workers by experience. Assuming that changes in the fraction of workers on the diagonal are partly driven by the decrease in the return, this variation allows for a counterfactual estimate of workers who would not be on the diagonal in the absence of on-diagonal returns. Details of this calculation are provided in Appendix F. The resulting estimated fraction of locked-in workers is 30%. Combining this share with the upper bound on losses per worker, these results suggest that the welfare loss from locked-in workers is given by at most 4.5% of wages per worker. This implies total

[40] By revealed preference a worker would not choose a different occupation after retraining in their current occupation. Note that, in a model of heterogeneous returns across the training-occupation matrix (a model with τ_{jk} instead of $\tau \times D_{j=k}$), τ will be a lower bound on the gains from retraining since retraining in the current occupation may not be the first best outcome.

[41] Note that the estimate controlling for occupation-specific experience could instead be used here, since a hypothetical retraining scenario would not lead to higher accumulated work experience in the newly chosen occupation. In practice, this makes little difference in the ensuing calculations.

losses from off-diagonal and locked-in workers of about $6 - 10.5\%$ of wages per worker in the dual apprenticeship system.

8.2 Retraining Programs

The calculations in Section 8.1 show that the welfare losses due to imperfect information at the time of training choice are large. This section briefly considers retraining programs as a potential policy intervention.[42] Since workers are trained in occupation-specific subjects for two thirds of their apprenticeship, I assume that retraining programs would last two thirds of the initial training time. Details on all calculations below are presented in Appendix F.

Retraining programs will be costly to the government and training firms. In addition, workers will forgo earnings while retraining. In 2010, total costs to train an apprentice, including training, schooling costs and foregone earnings amounted to 31,960 Euros. As outlined in Section 8.1, the per worker annual gain from retraining is τ for off-diagonal workers, and at most τ for locked-in workers. My results suggest that this implies annual benefits net of foregone work experience of $3,510$ Euros in 2010.

Note that, while the costs of retraining need to be paid upfront, the benefits accrue for every subsequent year spent working with a more suitable training. Whether or not retraining is associated with a net benefit therefore depends on the career stage of a worker. Following this logic, my calculations indicate that retraining costs would be recovered for workers with at most 10 years of work experience. Since about two thirds of off-diagonal workers leave their training occupation in the first 10 years after completing the apprenticeship (see Figure 2 in Section 2.5) and only switch occupations once (see Table 1 in Section 2.5), the findings suggest that retraining could pass a cost-benefit test for a large majority of the workers ever working off the diagonal. Moreover, an additional 85% of workers ever locked in could benefit from retraining (see Appendix F).

The above results imply that ex-post retraining programs could be highly effective in addressing the imperfect information workers face at the time of training choice. While a small number of firms already offer shorter training programs for career switchers, my findings suggest that these programs should be substantially expanded to facilitate retraining in response to new information on abilities, preferences and labor market circumstances.[43] In addition, further research is required to better understand potential barriers to entry into retraining such as liquidity constraints.

[42] The effects of a potential *ex-ante* provision of information by the government at the time of training choice are harder to quantify. See Appendix F for a discussion.

[43] The sample fraction of individuals enrolled in two apprenticeships in distinct occupations is only 4.02% (see Section 2.5). This is likely an overestimate of the fraction retraining as it includes spells for which the occupation is missing and non-completed apprenticeships.

9 Conclusion

This paper combines a large administrative employment panel with data on historical occupation-specific vacancies to identify and estimate the returns to different training-occupation combinations. To this end, I extend previous methodological approaches in the presence of high-dimensional selection to the given context where individuals select amongst a large number of alternatives in two stages. I provide a behavioral justification for the identification strategy, and implement the estimation approach by setting up a generalized two-stage Roy (1951) model where individuals seek relative advantage when choosing their training and occupation. To the best of my knowledge, this work is the first to present estimates on the returns to different training-occupation combinations which are well identified.

The results suggest significant returns of 15% to working in the occupation one has been trained for, with considerable heterogeneity across trainings and occupations. Combining the returns with data on the task content of occupations shows that returns in a particular training-occupation cell are lower, the higher the task distance between the training and the occupation. These findings provide a microfoundation for the estimated returns, and contribute to the literature on the task content of occupations by directly relating tasks workers are trained in to the value of human capital across occupations.

Given the magnitude of the estimates, my findings suggest that imperfect information at the time of training choice leads to important welfare losses. These losses are economically meaningful and may seem surprising in the context of the German apprenticeship system which has repeatedly been termed a role model for vocational training in other economies in Europe, the US, China and India. My findings show that young workers' ex-ante imperfect information on own abilities and future labor market developments should be addressed by policy makers, and that ex-post retraining programs could generate sizable net welfare gains. For the presented policy analysis, I took the existing training system as given and looked at improvements in the allocation of workers to training choices. Building on the estimates of the effect of task distance on the returns to training-occupation combinations, an interesting future avenue for research could relax this constraint and consider optimal training programs.

References

Abadie, Alberto, Athey, Susan, Imbens, Guido W., & Wooldridge, Jeffrey. 2017. When Should You Adjust Standard Errors for Clustering? *NBER Working Paper 24003*.

Acemoglu, Daron, & Autor, David. 2010. Skills, Tasks and Technologies: Implications for Employment and Earnings. *Chap. 12, pages 1043–1171 of:* Ashenfelter, Orley, & Card, David (eds), *Handbook of Labor Economics*, vol. 4b. North Holland.

Ahn, Hyungtaik, & Powell, James L. 1993. Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism. *Journal of Econometrics*, **58**(1-2), 3–29.

Altonji, Joseph G., Blom, Erica, & Meghir, Costas. 2012. Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers. *Annual Review of Economics*, **4**, 185–223.

Altonji, Joseph G., Kahn, Lisa B., & Speer, Jamin D. 2014. Trends in Earnings Differentials across College Majors and the Changing Task Composition of Jobs. *American Economic Review: Papers & Proceedings*, **104**(5), 387–393.

Altonji, Joseph G., Arcidiacono, Peter, & Maurel, Arnaud. 2016. The Analysis of Field Choice in College and Graduate School: Determinants and Wage Effects. *Chap. 7, pages 305–396 of:* Hanushek, Erik A., Machin, Stephen, & Woessmann, Ludger (eds), *Handbook of the Economics of Education*, vol. 5. North Holland.

Arcidiacono, Peter. 2004. Ability Sorting and the Returns to College Major. *Journal of Econometrics*, **121**(1-2), 343–375.

Arcidiacono, Peter, Hotz, V. Joseph, Maurel, Arnaud, & Romano, Teresa. 2020. Ex Ante Returns and Occupational Choice. *Journal of Political Economy*, **128**(12), 4475–4522.

Autor, David H. 2013. The "Task Approach" to Labor Markets: An Overview. *Journal for Labour Market Research*, **46**(3), 185–199.

Autor, David H., & Handel, Michael J. 2013. Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics*, **31**(S1, Part 2), S59–S96.

Autor, David H., Levy, Frank, & Murnane, Richard J. 2003. The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, **118**(4), 1279–1333.

Autor, David H., Dorn, David, & Hanson, Gordon H. 2013. The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review*, **103**(6), 2121–2168.

Autor, David H., Dorn, David, Hanson, Gordon H., & Song, Jae. 2014. Trade Adjustment: Worker-Level Evidence. *The Quarterly Journal of Economics*, **129**(4), 1799–1860.

Becker, Gary S. 1962. Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy*, **70**(5), 9–49.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Columbia University Press, New York.

Berry, Steven, Levinsohn, James, & Pakes, Ariel. 1995. Automobile Prices in Market Equilibrium. *Econometrica*, **63**(4), 841–890.

Breiman, Leo. 2001. Random Forests. *Machine Learning*, **45**, 5–32.

Breschi, Stefano, Lissoni, Francesco, & Malerba, Franco. 2003. Knowledge-Relatedness in Firm Technological Diversification. *Research Policy*, **32**(1), 69–87.

Chamberlain, Gary. 1986. Asymptotic Efficiency in Semi-Parametric Models with Censoring. *Journal of Econometrics*, **32**, 189–218.

Cortes, Guido Matias, & Gallipoli, Giovanni. 2018. The Costs of Occupational Mobility: An Aggregate Analysis. *Journal of the European Economic Association*, **16**(2), 275–315.

Dahl, Gordon B. 2002. Mobility and the Return to Education: Testing a Roy Model with Multiple Markets. *Econometrica*, **70**(6), 2367–2420.

Das, Mitali, Newey, Whitney K., & Vella, Francis. 2003. Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies*, **70**(1), 33–58.

d'Haultfœille, Xavier, & Maurel, Arnaud. 2013. Inference on an Extended Roy Model, with an Application to Schooling Decisions in France. *Journal of Econometrics*, **174**(2), 95–106.

DiNardo, John E., & Pischke, Jörn-Steffen. 1997. The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too? *The Quarterly Journal of Economics*, **112**(1), 291–303.

Fersterer, Josef, Pischke, Jörn-Steffen, & Winter-Ebmer, Rudolf. 2008. Returns to Apprenticeship Training in Austria: Evidence from Failed Firms. *The Scandinavian Journal of Economics*, **110**(4), 733–753.

Gathmann, Christina, & Schönberg, Uta. 2010. How General is Human Capital? A Task-Based Approach. *Journal of Labor Economics*, **28**(1), 1–49.

Goos, Maarten, Manning, Alan, & Salomons, Anna. 2014. Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. *American Economic Review*, **104**(8), 2509–2526.

Griliches, Zvi. 1977. Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica*, **45**(1), 1–22.

Guvenen, Fatih, Kuruscu, Burhan, Tanaka, Satoshi, & Wiczer, David. 2020. Multidimensional Skill Mismatch. *American Economic Journal: Macroeconomics*, **12**(1), 210–244.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, New York.

Hastings, Justine S., Neilson, Christopher A., & Zimmerman, Seth D. 2013. Are Some Degrees Worth More than Others? Evidence from College Admission Cutoffs in Chile. *NBER Working Paper 19241*.

Heckman, James J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **5**(4), 475–492.

Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica*, **47**(1), 153–161.

Heckman, James J. 1990. Varieties of Selection Bias. *American Economic Review*, **80**(2), Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, 313–318.

Heckman, James J., & Robb, Richard. 1985. Alternative Methods for Evaluating the Impact of Interventions. *Chap. 4, pages 156–246 of:* Heckman, James J., & Singer, Burton (eds), *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, New York.

Heckman, James J., Urzua, Sergio, & Vytlacil, Edward. 2006. Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, **88**(3), 389–432.

Heckman, James J., Urzua, Sergio, & Vytlacil, Edward. 2008. Instrumental Variables in Models with Multiple Outcomes: the General Unordered Case. *Annales d'Economie et de Statistique*, **91/92**, 151–174.

Hull, Peter. 2018. Estimating Hospital Quality with Quasi-Experimental Data. *Unpublished manuscript*.

Imbens, Guido W., & Angrist, Joshua D. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**(2), 467–475.

Jaffe, Adam B. 1986. Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value. *American Economic Review*, **76**(5), 984–1001.

Jaffe, Adam B. 1989a. Characterizing the "Technological Position" of Firms, with Application to Quantifying Technological Opportunity and Research Spillovers. *Research Policy*, **18**(2), 87–97.

Jaffe, Adam B. 1989b. Real Effects of Academic Research. *American Economic Review*, **79**(5), 957–970.

Kambourov, Gueorgui, & Manovskii, Iourii. 2009. Occupational Specificity of Human Capital. *International Economic Review*, **50**(1), 63–115.

Keane, Michael P., & Wolpin, Kenneth I. 1997. The Career Decisions of Young Men. *Journal of Political Economy*, **105**(3), 473–522.

Kinsler, Josh, & Pavan, Ronny. 2015. The Specificity of General Human Capital: Evidence from College Major Choice. *Journal of Labor Economics*, **33**(4), 933–972.

Kirkebøen, Lars J., Leuven, Edwin, & Mogstad, Magne. 2016. Field of Study, Earnings, and Self-Selection. *The Quarterly Journal of Economics*, **131**(3), 1057–1111.

Kline, Patrick, & Walters, Christopher R. 2019. On Heckits, LATE, and Numerical Equivalence. *Econometrica*, **87**(2), 677–696.

Lazear, Edward P. 2009. Firm-Specific Human Capital: A Skill-Weights Approach. *Journal of Political Economy*, **117**(5), 914–940.

Lee, Lung-Fei. 1983. Generalized Econometric Models with Selectivity. *Econometrica*, **51**(2), 507–512.

Lemieux, Thomas. 2014. Occupation, Fields of Study and Returns to Education. *Canadian Journal of Economics*, **47**(4), 1047–1077.

Lentz, Rasmus, Piyapromdee, Suphanit, & Robin, Jean-Marc. 2021. The Anatomy of Sorting - Evidence from Danish Data. *Unpublished manuscript*.

Miller, Robert A. 1984. Job Matching and Occupational Choice. *Journal of Political Economy*, **92**(6), 1086–1120.

Mincer, Jacob A. 1974. *Schooling, Experience, and Earnings*. National Bureau of Economic Research; distributed by Columbia University Press, New York.

Neal, Derek. 1995. Industry-Specific Human Capital: Evidence from Displaced Workers. *Journal of Labor Economics*, **13**(4), 653–677.

Nordin, Martin, Persson, Inga, & Rooth, Dan-Olof. 2010. Education-Occupation Mismatch: Is there an Income Penalty? *Economics of Education Review*, **29**(6), 1047–1059.

Poletaev, Maxim, & Robinson, Chris. 2008. Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984-2000. *Journal of Labor Economics*, **26**(3), 387–420.

Ransom, Tyler. 2021. Selective Migration, Occupation Choice, and the Wage Returns to College Major. *Annals of Economics and Statistics*.

Robinson, Chris. 2018. Occupational Mobility, Occupational Distance, and Specific Human Capital. *The Journal of Human Resources*, **53**(2), 513–551.

Robst, John. 2007. Education and Job Match: The Relatedness of College Major and Work. *Economics of Education Review*, **26**(4), 397–407.

Roy, Andrew D. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, **3**(2), 135–146.

Shaw, Kathryn L. 1984. A Formulation of the Earnings Function Using the Concept of Occupational Investment. *The Journal of Human Resources*, **19**(3), 319–340.

Shaw, Kathryn L. 1987. Occupational Change, Employer Change, and the Transferability of Skills. *Southern Economic Journal*, **53**(3), 702–719.

Spitz-Oener, Alexandra. 2006. Technical Change, Job Tasks, and Rising Educational Demands: Looking outside the Wage Structure. *Journal of Labor Economics*, **24**(2), 235–270.

Van der Velde, Lucas. 2020. Within Occupation Wage Dispersion and the Task Content of Jobs. *Oxford Bulletin of Economics and Statistics*, **82**(5), 1161–1197.

Vytlacil, Edward. 2002. Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, **70**(1), 331–341.

Walker, W. Reed. 2013. The Transitional Costs of Sectoral Reallocation: Evidence From the Clean Air Act and the Workforce. *The Quarterly Journal of Economics*, **128**(4), 1787–1835.

Yamaguchi, Shintaro. 2012. Tasks and Heterogeneous Human Capital. *Journal of Labor Economics*, **30**(1), 1–53.

Yi, Moises, Müller, Steffen, & Stegmaier, Jens. 2017. Industry Mix, Local Labor Markets, and the Incidence of Trade Shocks. *Unpublished manuscript*.

Appendix A. Tables

Table A.1: List of Occupations

KldB88 Code	Occupation label	Sub-label	% in code
75-78	Office workers	Office workers	73.1
		Other	26.9
19-30, 32	Craft workers	Vehicle mechanics	14.4
		Machine fitters	10.7
		Plumbers	10.7
		Other	64.2
68-70	Sales, financial workers	Salespeople	34.3
		Banking experts	24.3
		Wholesalers, retail dealers	16.6
		Other	24.8
79-89	Health, social workers	Medical receptionists	25.9
		Nurses, midwives	23.0
		Nursery, childcare w.	10.2
		Other	40.9
44-51	Construction workers	Bricklayers, concrete w.	21.9
		Carpenters	21.2
		Decorators, painters	15.7
		Other	41.2
10-18, 52-54	Process, plant workers	Chemical, plastics proc. w.	26.4
		Unskilled laborers	19.2
		Other	54.4
71-74	Transport, storage workers	Vehicle drivers	39.7
		Movers, warehousers	22.0
		Stock clerks	17.9
		Other	20.4
60-63	Technical, laboratory workers	Other technicians	22.6
		Technical drawers	17.0
		Electrical technicians	16.0
		Other	44.4

Table A.1 continued: List of Occupations

KldB88 Code	Occupation label	Sub-label	% in code
31	Electrical workers	Electricians	69.5
		Telephone technicians	17.2
		Electrical appliance fitters	13.3
		Other	0
90-93	Personal service workers	Hairdresssers, body care occ.	40.8
		Hospitality workers	28.4
		Other	30.8
39-43	Food preparation workers	Cooks, ready meal producers	39.0
		Bakers, confectioners	28.6
		Butchers, fish processing w.	21.7
		Coopers, brewers, food prod.	10.8
		Other	0
01-09	Agricultural workers	Gardeners, florists, foresters	57.9
		Miners, oil production w.	22.9
		Farmers, zookeepers	19.2
		Other	0
33-37	Textile, garment workers	Tailors, textile ind. w.	59.6
		Spinners, leather good/shoem.	40.4
		Other	0

Notes: The table lists all occupations contained in the baseline sample by fraction in the sample. Sub-labels are provided for all within-code shares greater than 10%.

Table A.2: Spells as Percentage of Trainings

			Occupation											
			01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	Agricultural	51.8	6.3	4.9	0.8	0.2	0.4	4.8	3.3	5.3	9.6	6.5	4.7	1.4
	10-54	Process, plant	0.7	57.3	4.2	0.6	0.1	0.2	1.5	12.0	4.7	5.7	8.7	3.6	0.6
	19-32	Craft	0.9	9.5	55.3	1.6	0.2	0.4	2.5	8.9	3.9	9.2	4.8	2.4	0.6
	31	Electrical	0.6	5.3	8.7	47.0	0.1	0.2	1.2	17.1	3.8	4.7	8.0	2.8	0.5
	33-37	Textile, garment	0.4	9.7	8.0	0.4	35.6	1.5	3.6	7.0	8.3	5.6	12.7	4.7	2.5
	39-43	Food preparation	1.1	8.6	6.2	0.7	0.3	43.1	3.6	1.7	7.8	13.2	6.7	3.6	3.4
	44-51	Construction	1.1	7.5	5.7	0.5	0.3	0.4	60.2	4.3	3.1	9.4	3.6	2.9	0.9
	60-63	Technical, lab.	0.3	2.7	2.5	3.2	0.0	0.1	0.7	68.7	4.1	1.6	12.8	2.8	0.5
	68-70	Sales, financial	0.2	2.3	1.6	0.2	0.2	0.6	0.3	0.8	60.6	3.4	26.5	2.1	1.1
	71-74	Transport, storage	0.1	5.3	3.5	0.9	0.0	0.3	2.5	2.1	7.6	55.2	18.9	2.9	0.7
	75-78	Office	0.1	0.8	0.6	0.1	0.0	0.0	0.1	1.1	12.6	2.0	80.6	1.6	0.4
	79-89	Health, social	0.2	0.8	0.7	0.1	0.0	0.2	0.2	0.8	4.3	1.0	12.2	79.0	0.7
	90-93	Personal service	0.4	5.2	3.6	0.5	0.3	3.0	0.6	1.0	10.6	3.9	20.8	4.9	45.2

Notes: The table reports the number of spells with a particular training-occupation combination as a percentage of all spells in the training for the baseline sample. Results are restricted to individuals with ten years of work experience.

Table A.3: Spells as Percentage of Occupations

			Occupation												
			01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	Agricultural	70.8	2.5	0.8	0.5	0.9	0.4	1.5	1.3	0.9	3.9	0.7	1.2	1.5
	10-54	Process, plant	0.8	19.2	0.6	0.3	0.3	0.2	0.4	3.7	0.7	1.9	0.8	0.8	0.6
	19-32	Craft	11.3	36.9	84.3	8.9	8.8	3.6	7.5	32.3	6.2	36.2	5.0	5.7	6.7
	31	Electrical	2.6	6.4	4.1	84.1	0.9	0.7	1.2	19.2	1.9	5.7	2.6	2.1	1.6
	33-37	Textile, garment	0.2	1.5	0.5	0.1	70.1	0.6	0.4	1.0	0.5	0.9	0.5	0.5	1.1
	39-43	Food preparation	2.6	6.1	1.7	0.8	2.8	82.8	2.0	1.1	2.3	9.6	1.3	1.6	6.6
	44-51	Construction	7.0	13.5	4.0	1.2	6.3	2.1	85.1	7.2	2.3	17.0	1.7	3.2	4.3
	60-63	Technical, lab.	0.4	1.2	0.4	2.1	0.2	0.1	0.2	27.6	0.7	0.7	1.5	0.8	0.6
	68-70	Sales, financial	1.7	5.9	1.7	0.9	6.1	4.3	0.6	1.8	64.8	9.0	18.3	3.4	8.1
	71-74	Transport, storage	0.0	0.6	0.1	0.1	0.0	0.1	0.2	0.2	0.3	5.9	0.5	0.2	0.2
	75-78	Office	1.0	2.3	0.7	0.3	1.0	0.3	0.3	2.8	14.4	5.5	59.4	2.8	3.1
	79-89	Health, social	0.8	1.3	0.4	0.3	0.6	0.7	0.3	1.2	2.7	1.6	5.1	76.4	2.8
	90-93	Personal service	0.7	2.7	0.7	0.4	2.2	4.1	0.2	0.5	2.2	2.0	2.8	1.6	62.9

Notes: The table reports the number of spells with a particular training-occupation combination as a percentage of all spells in the occupation for the baseline sample. Results are restricted to individuals with ten years of work experience.

Table A.4: Suggestive First Stage Regressions

	Occupation												
	01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
$ln(vac_{in})$	0.80	1.82	9.86	0.04	0.49	1.10	2.37	1.69	2.43	1.20	5.92	1.91	1.08
	(0.02)	(0.04)	(0.12)	(0.10)	(0.00)	(0.04)	(0.04)	(0.05)	(0.15)	(0.03)	(0.09)	(0.05)	(0.04)
$ln(vac_{out})$	-0.77	-2.21	-9.76	0.27	-0.43	-1.23	-2.78	-1.12	-2.68	-1.2	-5.17	-2.65	-1.49
	(0.02)	(0.05)	(0.14)	(0.10)	(0.01)	(0.04)	(0.05)	(0.06)	(0.15)	(0.03)	(0.08)	(0.06)	(0.04)

Notes: N=3,964,883 in all regressions. This corresponds to the main sample, excluding observations where no vacancy data is available (1975-1978 and 1992-1993 in East Germany). The dependent variable in each column is a dummy variable equal to one for the given occupation and zero otherwise, vac_{in} denotes vacancies in the given occupation, and vac_{out} denotes the mean of vacancies in other occupations. Regressions control for gender and experience. Coefficients are scaled by 100. Robust standard errors are reported.

Table A.5: Full Matrix of Returns - No Selection Control

			Occupation												
			01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	Agricultural	0	−0.02	0.03	0.03	−0.02	0.04	0.07	0.08	0.07	−0.02	0.04	0.02	−0.21
				(0.03)	(0.04)	(0.07)	(0.06)	(0.11)	(0.05)	(0.10)	(0.03)	(0.04)	(0.04)	(0.04)	(0.07)
	10-54	Process, plant	−0.08	0	−0.03	−0.01	0.27	0.00	−0.02	0.08	−0.02	−0.06	0.00	−0.01	−0.45
			(0.07)		(0.05)	(0.10)	(0.14)	(0.14)	(0.06)	(0.03)	(0.07)	(0.04)	(0.05)	(0.04)	(0.17)
	19-32	Craft	−0.09	0.02	0	0.04	0.13	−0.01	0.03	0.15	0.07	−0.01	0.08	−0.03	−0.13
			(0.04)	(0.01)		(0.02)	(0.08)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.06)
	31	Electrical	−0.05	0.07	0.09	0	−0.07	0.05	−0.03	0.17	0.13	−0.02	0.19	0.04	−0.09
			(0.08)	(0.02)	(0.02)		(0.14)	(0.07)	(0.06)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)
	33-37	Textile, garment	0.07	0.00	0.08	−0.04	0	−0.20	0.06	0.04	−0.03	−0.02	−0.05	−0.09	−0.57
			(0.21)	(0.03)	(0.06)	(0.13)		(0.15)	(0.09)	(0.10)	(0.12)	(0.12)	(0.10)	(0.09)	(0.09)
	39-43	Food preparation	0.00	0.03	0.17	0.10	0.06	0	0.11	0.21	0.13	0.07	0.15	0.07	−0.03
			(0.05)	(0.03)	(0.03)	(0.10)	(0.05)		(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.04)
	44-51	Construction	−0.09	−0.07	0.01	−0.10	−0.22	−0.03	0	0.06	−0.06	−0.07	0.01	−0.11	−0.21
			(0.03)	(0.02)	(0.02)	(0.03)	(0.06)	(0.04)		(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.04)
	60-63	Technical, lab.	−0.22	−0.08	0.02	−0.09	−0.17	0.08	0.05	0	0.05	−0.06	0.03	−0.04	−0.66
			(0.19)	(0.06)	(0.04)	(0.03)	(0.31)	(0.15)	(0.07)		(0.05)	(0.08)	(0.03)	(0.07)	(0.20)
	68-70	Sales, financial	−0.24	−0.03	0.11	0.05	0.06	−0.01	−0.04	0.05	0	−0.01	0.01	−0.09	−0.26
			(0.05)	(0.03)	(0.02)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)		(0.02)	(0.01)	(0.03)	(0.03)
	71-74	Transport, storage	−0.40	−0.07	0.10	−0.19	0	−0.42	0.04	−0.04	−0.08	0	−0.02	0.04	−0.26
			(0.25)	(0.07)	(0.05)	(0.17)		(0.22)	(0.08)	(0.05)	(0.10)		(0.04)	(0.11)	(0.15)
	75-78	Office	−0.22	−0.09	−0.10	−0.11	0.01	−0.20	0.01	0.11	0.07	−0.08	0	−0.07	−0.38
			(0.09)	(0.04)	(0.05)	(0.08)	(0.17)	(0.12)	(0.07)	(0.03)	(0.01)	(0.03)		(0.03)	(0.07)
	79-89	Health, social	−0.29	−0.01	−0.01	0.11	0.08	−0.20	−0.14	−0.00	0.03	−0.08	−0.03	0	−0.49
			(0.09)	(0.04)	(0.04)	(0.16)	(0.13)	(0.13)	(0.11)	(0.04)	(0.02)	(0.04)	(0.03)		(0.07)
	90-93	Personal service	0.05	0.22	0.41	0.46	0.45	0.07	0.10	0.20	0.16	0.17	0.18	0.13	0
			(0.07)	(0.04)	(0.04)	(0.10)	(0.12)	(0.04)	(0.12)	(0.05)	(0.03)	(0.04)	(0.02)	(0.05)	

Notes: The table shows coefficient estimates $\hat{\tau}_{jk}$ from equation (1), estimated without selection control. Standard errors (in parentheses) are clustered at the region and time level.

Table A.6: Full Matrix of Returns - Parametric Selection Control

			Occupation											
			01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	Agricultural	0	−0.39	−0.21	−0.67	−0.14	−0.23	−0.04	0.26	−0.09	−0.31	−0.16	0.02	−0.46
				(0.13)	(0.17)	(0.25)	(0.68)	(0.36)	(0.25)	(0.17)	(0.10)	(0.18)	(0.13)	(0.19)	(0.25)
	10-54	Process, plant	0.04	0	0.02	−0.47	0.45	0.03	0.28	0.62	0.34	−0.07	0.11	0.33	−0.40
			(0.50)		(0.23)	(0.24)	(0.73)	(0.48)	(0.34)	(0.21)	(0.20)	(0.15)	(0.19)	(0.22)	(0.30)
	19-32	Craft	0.07	0.03	0	−0.23	0.34	0.07	0.28	0.57	0.37	0.05	0.21	0.26	−0.02
			(0.31)	(0.09)		(0.16)	(0.50)	(0.21)	(0.11)	(0.05)	(0.06)	(0.06)	(0.04)	(0.09)	(0.14)
	31	Electrical	0.38	0.32	0.46	0	0.41	0.40	0.53	0.93	0.74	0.30	0.60	0.64	0.28
			(0.35)	(0.10)	(0.11)		(0.62)	(0.29)	(0.15)	(0.11)	(0.08)	(0.10)	(0.09)	(0.11)	(0.19)
	33-37	Textile, garment	−0.23	−0.51	−0.28	−0.93	0	−0.59	0.07	0.16	−0.11	−0.45	−0.37	−0.20	−0.94
			(0.75)	(0.31)	(0.43)	(0.39)		(0.49)	(0.45)	(0.37)	(0.26)	(0.39)	(0.39)	(0.28)	(0.36)
	39-43	Food preparation	0.23	0.06	0.33	−0.21	0.35	0	0.48	0.80	0.56	0.18	0.36	0.48	0.14
			(0.38)	(0.11)	(0.11)	(0.20)	(0.55)		(0.16)	(0.11)	(0.09)	(0.08)	(0.10)	(0.12)	(0.21)
	44-51	Construction	−0.18	−0.33	−0.14	−0.66	−0.26	−0.19	0	0.28	0.02	−0.27	−0.10	−0.04	−0.36
			(0.27)	(0.09)	(0.07)	(0.20)	(0.51)	(0.18)		(0.09)	(0.06)	(0.09)	(0.06)	(0.14)	(0.21)
	60-63	Technical, lab.	−0.83	−0.91	−0.67	−1.27	−0.71	−0.62	−0.42	0	−0.35	−0.81	−0.62	−0.48	−1.35
			(0.46)	(0.18)	(0.22)	(0.19)	(0.78)	(0.41)	(0.27)		(0.17)	(0.20)	(0.18)	(0.21)	(0.38)
	68-70	Sales, financial	−0.30	−0.27	−0.01	−0.48	0.05	−0.14	0.02	0.31	0	−0.18	−0.10	−0.02	−0.38
			(0.31)	(0.16)	(0.16)	(0.20)	(0.44)	(0.20)	(0.15)	(0.15)		(0.13)	(0.10)	(0.11)	(0.20)
	71-74	Transport, storage	−0.46	−0.34	−0.02	−0.85	0	−0.57	0.15	0.33	0.09	0	−0.10	0.17	−0.39
			(0.42)	(0.20)	(0.20)	(0.39)		(0.43)	(0.27)	(0.19)	(0.18)		(0.18)	(0.24)	(0.32)
	75-78	Office	−0.47	−0.52	−0.41	−0.85	−0.19	−0.53	−0.12	0.18	−0.06	−0.44	0	−0.18	−0.68
			(0.34)	(0.11)	(0.20)	(0.21)	(0.45)	(0.35)	(0.14)	(0.11)	(0.09)	(0.10)		(0.16)	(0.16)
	79-89	Health, social	−0.98	−0.86	−0.74	−1.10	−0.53	−0.93	−0.72	−0.37	−0.48	−0.87	−0.73	0	−1.22
			(0.33)	(0.10)	(0.10)	(0.36)	(0.56)	(0.23)	(0.18)	(0.12)	(0.11)	(0.11)	(0.10)		(0.18)
	90-93	Personal service	−0.10	−0.12	0.19	−0.23	0.34	−0.17	0.08	0.41	0.17	−0.10	−0.00	0.13	0
			(0.33)	(0.17)	(0.17)	(0.28)	(0.61)	(0.25)	(0.24)	(0.15)	(0.13)	(0.17)	(0.12)	(0.17)	

Notes: The table shows coefficient estimates $\hat{\tau}_{jk}$ from model (1), estimated using the parametric control function estimator. Standard errors (in parentheses) are clustered at the region and time level.

Table A.7: Average On- versus Off-Diagonal Returns - Estimation Robustness

	(1) 10th-order polynomial	(2) full set of cf cells	(3) occ. x time FE	(4) ind. x time FE
$D_{j=k} = 1$	0.1515^{***}	0.1088^{***}	0.1491^{***}	0.1451^{***}
	(0.0191)	(0.0288)	(0.0217)	(0.0181)
exp	0.0593^{***}	0.0591^{***}	0.0581^{***}	0.0571^{***}
	(0.0023)	(0.0023)	(0.0022)	(0.0042)
exp^2	-0.0010^{***}	-0.0010^{***}	-0.0010^{***}	-0.0010^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Indiv./Reg. FE	yes	yes	yes	yes
Occ. FE	yes	yes		yes
Time FE	yes	yes		
Occ. x Time FE			yes	
Ind. x Time FE				yes
Parametric cf	yes	yes	yes	yes
p-value cf	0.000	0.000	0.000	0.000
N	1,140,518	1,140,518	1,140,512	1,139,022

Notes: The table reports regression results for equation (1) with $\tau_{jk} = \delta_k + \tau D_{j=k}$. Column (1) controls for a tenth order polynomial in own vacancies. Column (2) allows the control function to vary for the full set of training-occupation cells. Column (3) includes a full set of 14 industry fixed effects in the regression. Standard errors are clustered at the region and time level (columns (1), (2) and (3)), or at the region and time and industry level (column (4)). $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

Table A.7 shows further robustness results relating to the estimation method. Column (1) approximates the skill price in occupation k using a tenth order polynomial in log vacancies instead of the fourth order polynomial used throughout Chapter 6. This functional form change leaves the baseline estimate almost unchanged. Column (2) allows the parametric control function estimator to vary across *all* training-occupation cells. The resulting on- versus off-diagonal is only slightly lower at around 11%. Columns (3) and (4) address two of the identification concerns discussed in Section **??**, showing that the inclusion of occupation *times* time or industry *times* time fixed effects does not appreciably change the baseline result of 15.1%.

Table A.8: List of Tasks and Fraction Performing

Task		01-09	10-54	19-32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
1	Cultivate	80	3	1	1	0	2	4	1	1	2	0	2	1
2	Manufacture	30	39	47	39	66	67	50	19	6	5	4	7	10
3	Publish	1	0	0	1	0	1	0	6	4	1	5	17	2
4	Serve	8	2	2	1	2	38	2	2	8	2	6	15	32
5	Clean	38	29	27	24	35	50	33	11	20	25	8	26	70
6	Secure	19	15	15	18	6	13	16	16	9	17	9	25	9
7	Repair	52	38	72	84	41	19	71	31	11	31	9	20	17
8	Equip machines	50	65	66	61	53	48	43	39	12	30	14	20	18
9	Nurse	23	7	11	9	10	20	8	11	20	14	14	49	30
10	Pack	50	46	35	31	22	38	43	21	41	81	25	24	21
11	Execute laws	5	2	3	6	1	5	2	20	8	6	25	20	2
12	Design	31	19	29	34	22	29	32	62	33	18	38	42	21
13	Employ	21	11	14	20	8	21	16	49	33	14	38	36	14
14	Calculate	16	4	6	9	4	15	10	31	34	5	41	9	9
15	Research	39	43	50	57	42	39	41	68	43	27	47	54	23
16	Sell	48	10	21	27	22	37	27	42	86	23	48	39	39
17	Teach	39	29	39	46	31	37	39	54	52	30	51	72	31
18	Program	19	25	25	32	13	12	13	49	38	21	55	35	11
19	Correct texts	20	11	9	16	7	10	10	37	50	21	74	44	12

Notes: The table shows the average percentage of individuals indicating they perform the given task. Task 1: cultivate; task 2: manufacture, install or construct; task 3: publish, present or entertain others; task 4: serve or accommodate; task 5: clean; task 6: secure; task 7: repair, renovate, reconstruct; task 8: equip or operate machines; task 9: nurse or treat others; task 10: pack, ship or transport; task 11: execute laws or interpret laws; task 12: design, plan, sketch; task 13: employ, manage personnel, organize, coordinate; task 14: calculate or do bookkeeping; task 15: research, evaluate or measure; task 16: sell, buy or advertise; task 17: teach or train others; task 18: program; task 19: correct texts or data.

Table A.9: Training-Occupation Distances - Selected Categories

Statistics	Training j	Occupation k	$Dist_{jk}$
Overall mean			0.2314
Standard dev.			0.1108
Weight. mean			0.1863
	Craft workers	Electrical workers	0.0111
	Craft workers	Construction workers	0.0235
	Construction workers	Electrical workers	0.0316
	Craft workers	Process, plant workers	0.0379
	Craft workers	Textile, garment workers	0.0492
	.	.	.
	.	.	.
	.	.	.
	Office workers	Construction workers	0.4000
	Textile, garment workers	Sales, financial workers	0.4108
	Office workers	Process, plant workers	0.4089
	Office workers	Craft workers	0.4113
	Office workers	Textile, garment workers	0.4631

Notes: The table reports summary statistics on the distance measure $Dist_{jk}$, and distances for the five most similar and the five most distant training-occupation pairs.

Table A.10: Training-Occupation Distances - Five Largest Occupations

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	0				
	Craft workers	0.41	0			
	Sales, financ. w.	0.07	0.38	0		
	Health, social w.	0.12	0.29	0.12	0	
	Construction w.	0.40	0.02	0.33	0.28	0

Notes: The table reports the distance measure $Dist_{jk}$ for the five largest occupations.

Table A.11: Match Returns and Task Distance - Exclude On-Diagonal Observations

τ_{jk} estimated	with parametric control fcn.		without selection control	
	(1)	(2)	(3)	(4)
$Dist_{jk}$	-0.0753^{**}	-0.0573^{*}	-0.0189	-0.0037
	(0.0348)	(0.0309)	(0.0116)	(0.0090)
Occ. FE		yes		yes
Mean of $\hat{\tau}_{jk}$	-0.1383	-0.1383	-0.0150	-0.0150
N	156	156	156	156

Notes: The table reports regression results from equation (21). $Dist_{jk}$ is scaled by its standard deviation. Robust standard errors are reported. $^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Table A.12: Match Returns and Task Distance - Five Largest Trainings/Occupations

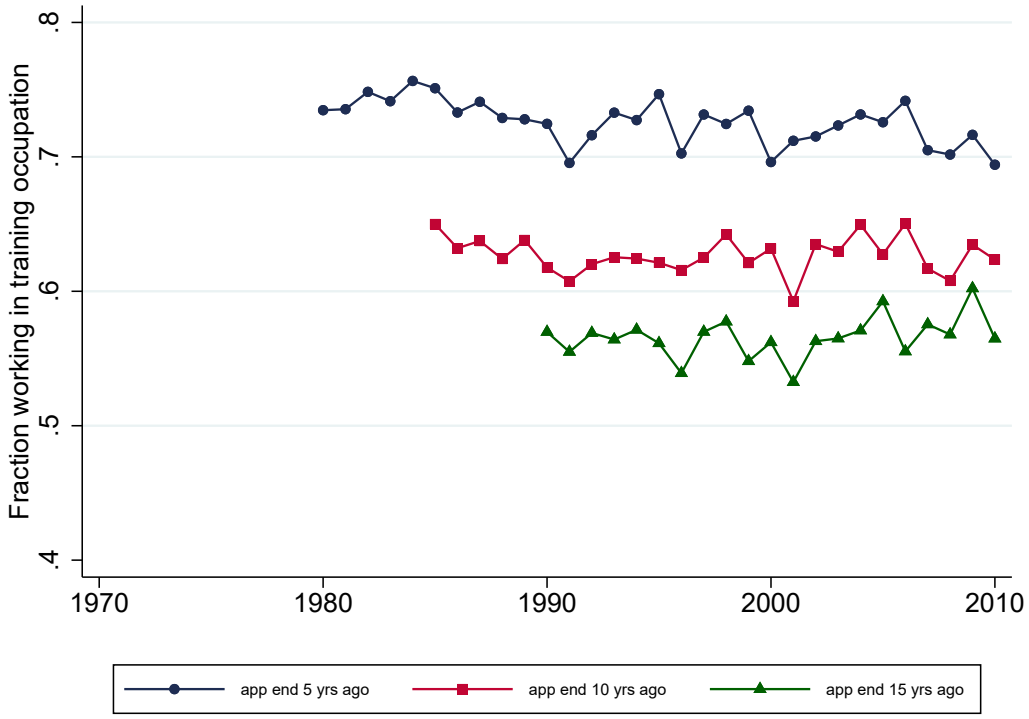
τ_{jk} estimated	with parametric control fcn.		without selection control	
	(1)	(2)	(3)	(4)
$Dist_{jk}$	-0.0802^{**}	-0.0488	-0.0240	-0.0062
	(0.0330)	(0.0331)	(0.0113)	(0.0090)
Occ. FE		yes		yes
Mean of $\hat{\tau}_{jk}$	-0.2382	-0.2382	-0.0478	-0.0478
N	65	65	65	65

Notes: The table reports regression results from equation (21). $Dist_{jk}$ is scaled by its standard deviation. Robust standard errors are reported. $^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Appendix B. Figures

Fraction On Diagonal Over Time

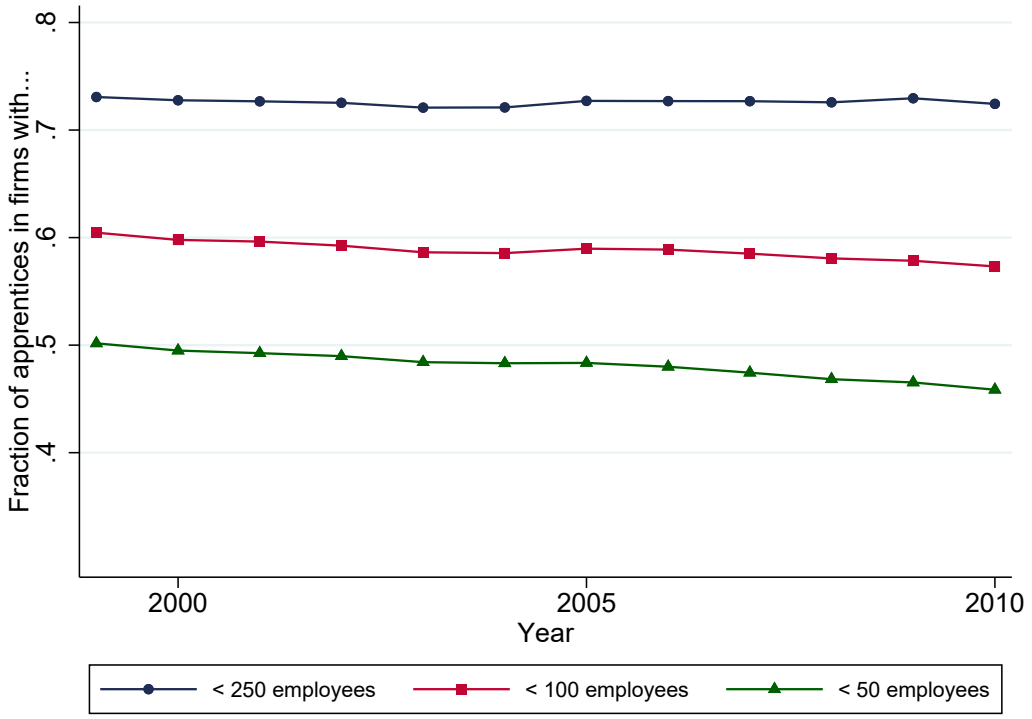
Figure B.1: Fraction On Diagonal over Time



Notes: The figure plots the fraction of individuals working in an occupation equal to their training occupation over time for the baseline sample. The three lines plot this fraction for individuals who finished their apprenticeship 5, 10, or 15 years prior to the date shown on the x-axis.

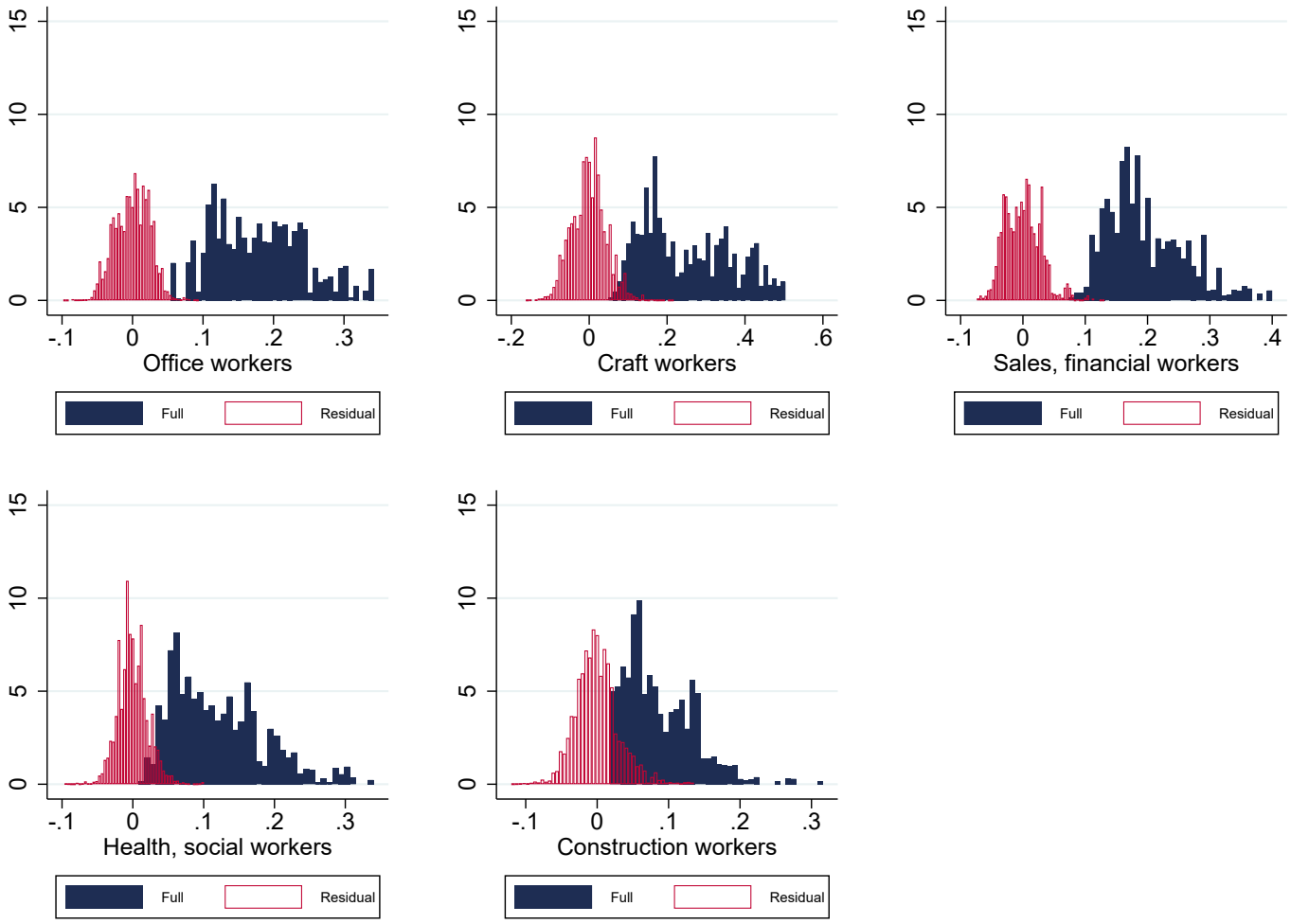
Training Firm Statistics

Figure B.2: Fraction of Apprentices by Firm Size



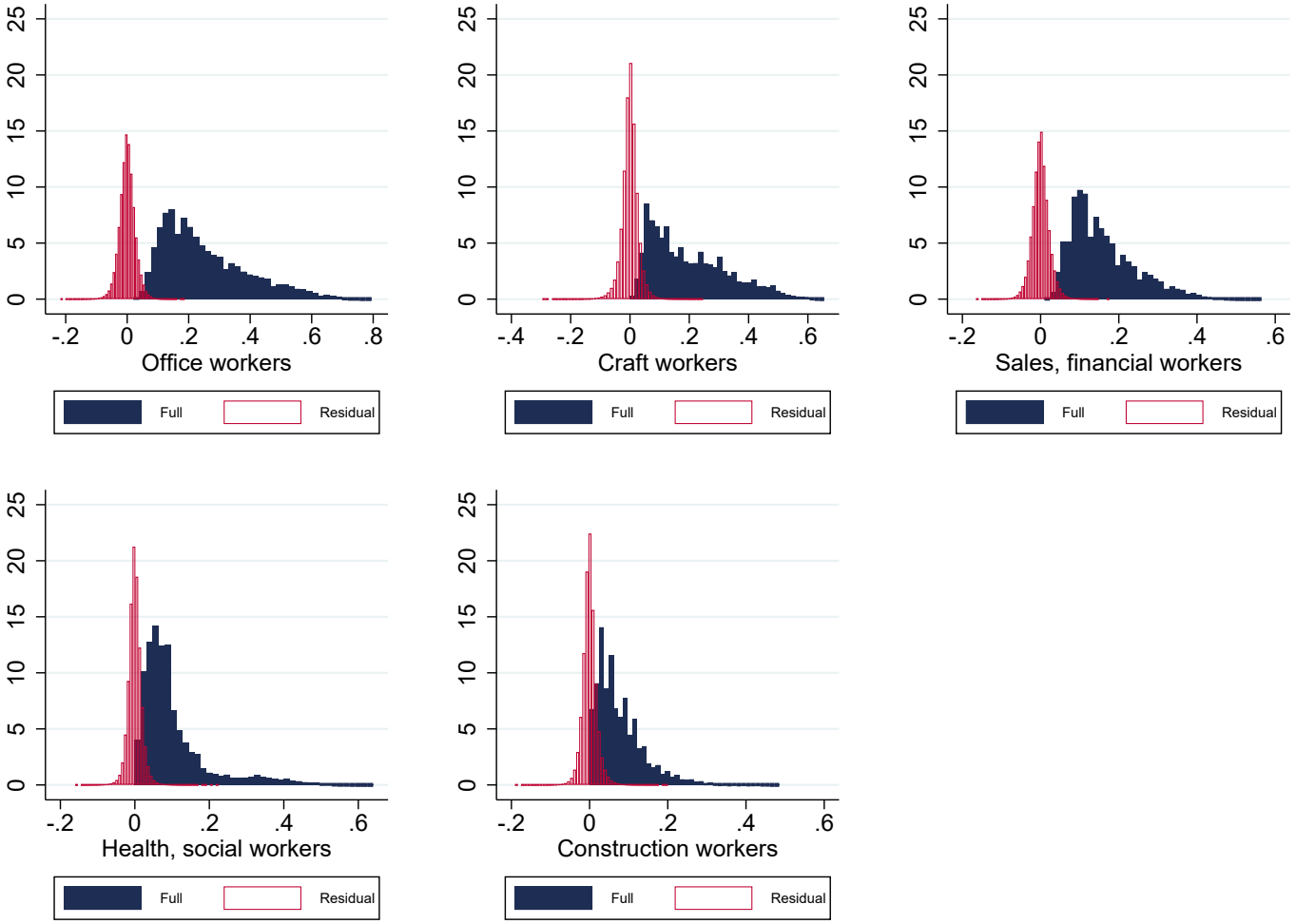
Notes: The figure plots the fraction of apprentices trained in firms with less than 50 (small firms), 100 and 250 (medium-sized firms) employees over time. Source: *Bundesagentur für Arbeit*.

Figure B.3: First Stage Variation in Selection Probabilities - Training



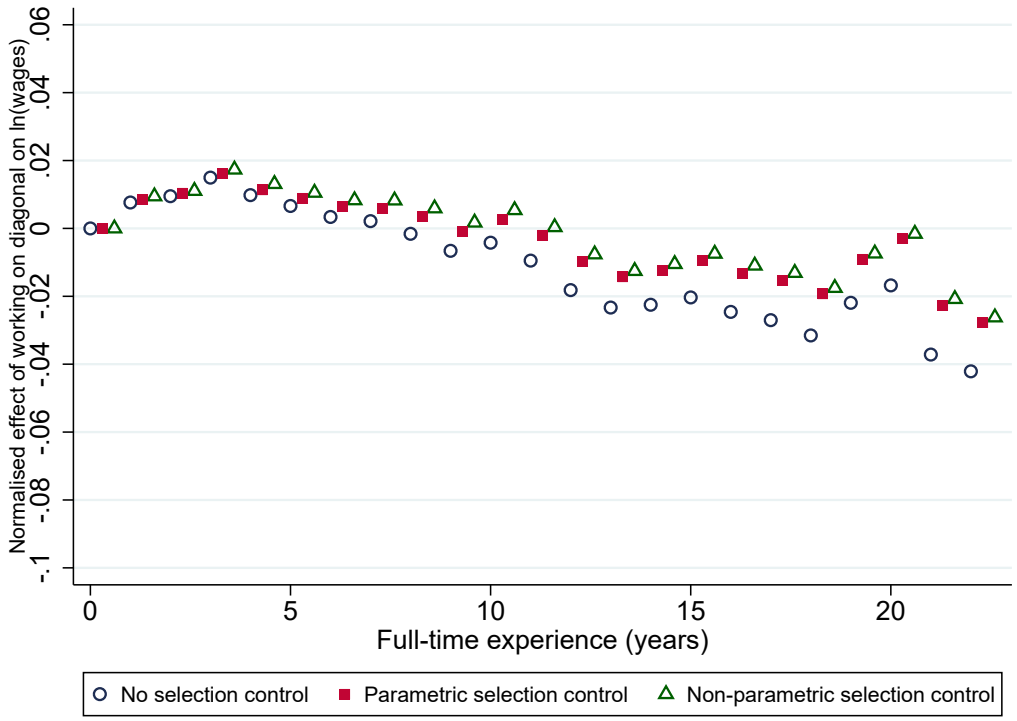
Notes: The figure shows a set of histograms of the estimated selection probabilities for the five largest trainings. Histograms in blue show the full variability in estimated selection probabilities. Histograms in red are restricted to male workers, and residualized using location and time of training fixed effects.

Figure B.4: First Stage Variation in Selection Probabilities - Occupation



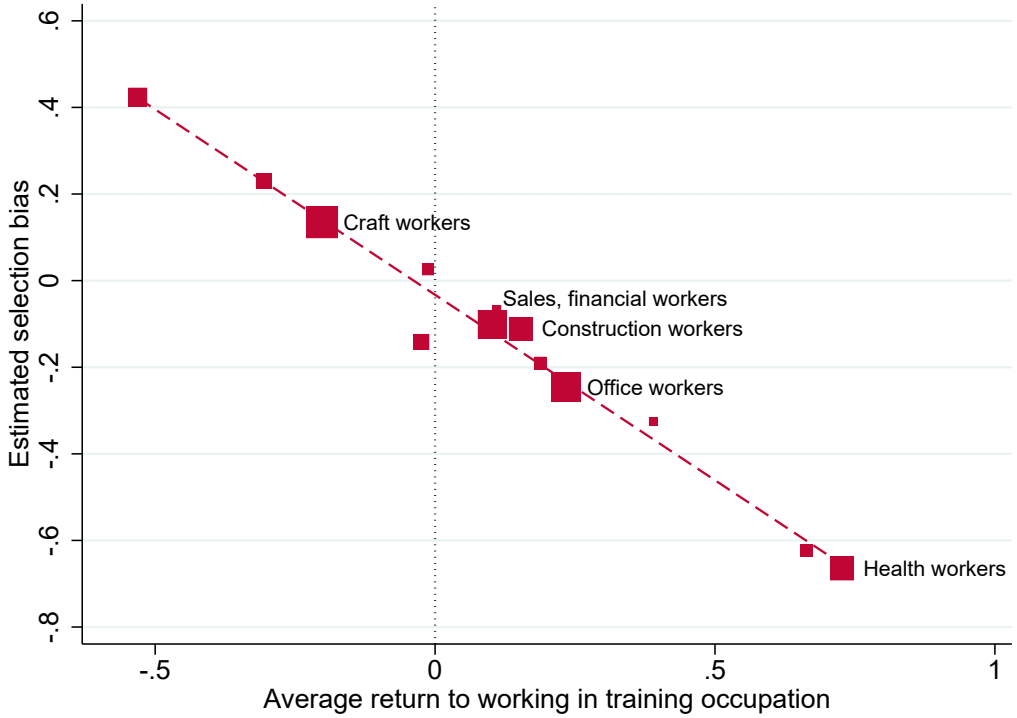
Notes: The figure shows a set of histograms of the estimated selection probabilities for the five largest occupations. Histograms in blue show the full variability in estimated selection probabilities. Histograms in red are residualized using region, time and individual fixed effects.

Figure B.5: Normalized On- versus Off-Diagonal Returns by Experience



Notes: The figure plots regression coefficient estimates for τ^{exp} in a version of equation (1) with $\tau_{jk} = \delta_k + \tau^{exp} D_{j=k}$, where experience levels have been binned into yearly categories. All coefficient estimates are normalized to zero at zero years of work experience.

Figure B.6: Average Return and Estimated Selection Bias by Training

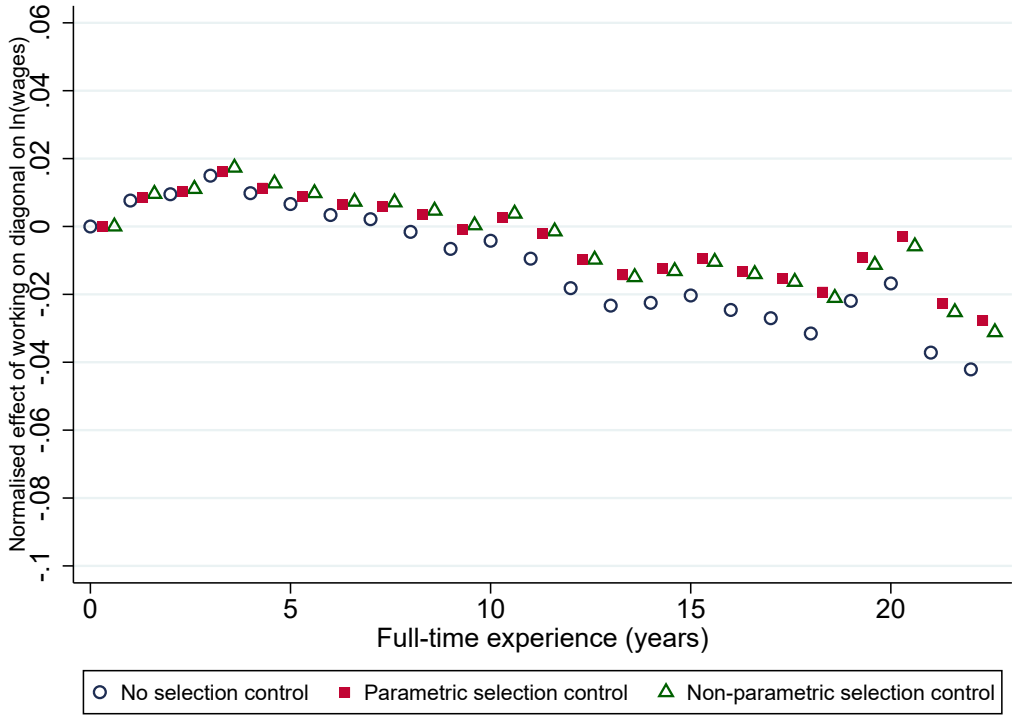


Notes: The figure plots average on- versus off-diagonal returns for each training from Figure 4 against the estimated selection bias, i.e. the difference between the returns estimated without selection control and those using the parametric control function estimator. The fitted line corresponds to a weighted OLS regression using the sample fraction in each training as weights. Marker size is proportional to the weights.

Figure B.7: Full Matrix of Returns and Sample Fraction

Notes: The figure plots training-occupation cell returns against the within-training sample fraction of workers in the relevant occupation for each off-diagonal training-occupation pair. Marker size is proportional to the fraction of workers in each cell.

Figure B.8: Normalized On- versus Off-Diagonal Returns by Work Experience - Robustness



Notes: The figure plots regression coefficient estimates for τ^{exp} in a version of equation (1) with $\tau_{jk} = \delta_k + \tau^{exp} D_{j=k}$, where experience levels have been binned into yearly categories. All coefficient estimates are normalized to zero at zero years of work experience. In contrast to Figure B.5 in Section 6.2, the non-parametric control function estimator also includes the on-diagonal probability $p_{i(k=j|j)rt}$.

Figure B.8 provides a comparison between the slope estimates τ^{exp} under no selection control, the parametric selection and the non-parametric selection control where, in contrast to Figure B.5, the probability of selection into one's training occupation $p_{i(k=j|j)rt}$ has been added as an additional term in the non-parametric control function (see Sections 5.1 and 6.2). It can be seen that the slope estimates when using the additional probability in the non-parametric control function are very similar to those in Figure B.5, closely mapping the slope estimates obtained when using the parametric selection control. This result provides further support to the distributional assumptions made.

Figure B.9: Training-Occupation Distance and Sample Fraction

Notes: The figure plots training-occupation distances against the within-training sample fraction of workers in the relevant occupation for each off-diagonal training-occupation pair. The fitted line corresponds to a weighted OLS regression where each training-occupation pair is weighted by the fraction of total workers in that cell. Marker size is proportional to the weights.

Appendix C. Proof

Proof for Section 4.1:

$E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})>0]-E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})<0]\geq 0.$

Given $\epsilon_{i1}\sim N(0,\sigma_{\epsilon_1})$, $\epsilon_{i2}\sim N(0,\sigma_{\epsilon_2})$ with $\sigma_{\epsilon_1}=\sigma_{\epsilon_2}$,

$$\begin{aligned}E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})=\nu>z] &= \frac{\sigma_{\epsilon_1}\sigma_{\epsilon_2}}{\sigma_\nu}(\frac{\sigma_{\epsilon_1}}{\sigma_{\epsilon_2}}-\rho_{\epsilon_1\epsilon_2})(\frac{\phi(z)}{1-\Phi(z)})\\ &= \frac{\sigma_{\epsilon_1}\sigma_{\epsilon_2}}{\sigma_\nu}(1-\rho_{\epsilon_1\epsilon_2})(\frac{\phi(z)}{1-\Phi(z)})\geq 0,\end{aligned}$$

where $\rho_{\epsilon_1\epsilon_2}=\frac{\sigma_{\epsilon_1\epsilon_2}}{\sigma_{\epsilon_1}\sigma_{\epsilon_2}}\leq 1$.

It follows that $E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2)})>0]-E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})<0]\geq 0$.

Defining $(\frac{\phi(z)}{1-\Phi(z)})=\kappa(z)$, $\kappa'(z)>0$ from the assumption of normality. It follows that $E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})>-\tau]-E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})>\tau]\leq 0$.

Appendix D. Estimation Details

Splitting Vacancies into Expectation and Shock

In order to obtain the training and occupation instruments defined in equations (**??**) and (**??**), vacancies need to be split into expectations and shocks. To do so, I estimate separate linear time trend models in each region-time cell, where log vacancies for each occupation are explained using five years of previous data. Note that using five years of past data to predict future vacancies implies that predictions and shocks will not be available during the first five years of the sample, 1978-1981. Moreover, due to regional classification changes following German reunification, data on predictions and shocks will also not be available for four regions between 1994-1997. This will reduce the number of observations in the baseline sample used in the estimation. I estimate the following model:

$$vac_{krt} = \kappa_{kr} + \pi_{krt} \times t + \varepsilon_{krt}, \quad \forall rt. \tag{D.1}$$

Note that I allow both the intercepts and slopes to be occupation-specific. Based on the region and time when first starting the apprenticeship, r_0 and t_0, 30-year ahead predictions for vacancies in each occupation are then computed for each individual as conditional expectations using equation (D.1):

$$E_{t_0}[vac_{kr(t_0+\tau)}|\Omega_{r_0t_0}] = \hat{\kappa}_{kr_0} + \hat{\pi}_{kr_0t_0} \times (t_0 + \tau), \quad \forall \tau = 0, ..., 30. \tag{D.2}$$

For any $t = t_0 + \tau$, individual-specific shocks to vacancies are then defined as residuals relative to the expectation formed at the time of training choice t_0 in region r_0:

$$vac_{krt} - E_{t_0}[vac_{kr(t_0+\tau)}|\Omega_{r_0t_0}], \quad \forall \tau = 0, ..., 30. \tag{D.3}$$

While the conditional expectations derived using equation (D.2) will serve as training instruments IV_{train_j}, the residuals from equation (D.3) will serve as occupation instruments IV_{occ_k}. Note that, using this definition, expectations and shocks will be orthogonal by construction.

Reduction of Dimensionality

Define the joint cumulative distribution of the outcome and selection error terms as $F_{jk}(...)$, and the joint cumulative distribution of the outcome error and the two maximum order statistics as $G_{jk}(...)$. Evaluating $F_{jk}(...)$ at the observed value function and utility differences, the equivalence between $F_{jk}(...)$ and $G_{jk}(...)$ can be established with the following steps:

$$
\begin{aligned}
&F_{jk}(z_0, \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt}) \\
&= Pr(\epsilon_{ikrt} \leq z_0, e_{i1r_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\qquad\qquad e_{i1rt} - e_{ikrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., e_{iKrt} - e_{ikrt} \leq \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= Pr(\epsilon_{ikrt} \leq z_0, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}) \leq 0, \\
&\qquad\qquad \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ik'rt} - e_{ikrt}) \leq 0| \\
&\qquad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= G_{jk}(z_0, 0, 0|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.4}
$$

This equivalence may also be written in terms of density functions:

$$
\begin{aligned}
&f_{jk}(\epsilon_{ikrt}, e_{i1r_0t_0} - e_{ijr_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0}, e_{i1rt} - e_{ikrt}, ..., e_{iKrt} - e_{ikrt}| \\
&\qquad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= g_{jk}(\epsilon_{ikrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ik'rt} - e_{ikrt})| \\
&\qquad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.5}
$$

Given the one-to-one mapping between the selection probabilities and the observed utility and value function differences, the joint distribution $g_{jk}(...)$ may be conditioned on $(p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt})$, where $p_{ijr_0t_0}$ is the probability of selcting into training j at time t_0, and $p_{i(k|j)rt}$ is the probability of selecting into occupation k conditional on training j at time t:

$$
\begin{aligned}
&= g_{jk}(\epsilon_{ikrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ik'rt} - e_{ikrt})| \\
&\qquad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= g_{jk}(\epsilon_{ikrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ik'rt} - e_{ikrt})| \\
&\qquad p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt}).
\end{aligned} \tag{D.6}
$$

Rewriting the joint distribution $g_{jk}(...)$ in this way captures the fact that the vector of selection probabilities contains the same information as the observed utility and value function differences.

Lee's (1983) Parametric Control Function

Define $M_{ijkrt} = train_{ij} \times occ_{i(k|j)rt}$ and recall that the selection problem is given by

$$M_{ijkrt} = 1 \quad \text{iff} \quad \max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq 0 \ \text{ and } \ \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq 0. \tag{D.7}$$

Lee (1983) points out that it is possible to create new random variables based on the distribution of the maximum order statistics (see Appendix A in Dahl (2002) for details). I use Dahl's (2002) notation and adapt Lee's (1983) approach to the present selection problem. To do so, define the marginal distribution of the selection errors as $L_{jk}(...)$, and the marginal distribution of the two maximum order statistics as $H_{jk}(...)$. Denote the corresponding density functions by $l_{jk}(...)$ and $h_{jk}(...)$, respectively. Using Lee's (1983) insight on maximum order statistics, and evaluating $L_{jk}(...)$ at the observed utility and value function differences, the distribution may be written as

$$\begin{aligned}
&L_{jk}(\tilde{V}_{ijrt_0} - \tilde{V}_{i1rt_0} + z_1, ..., \tilde{V}_{ijrt_0} - \tilde{V}_{iJrt_0} + z_1, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt} + z_2, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt} + z_2) \\
&\quad = Pr(e_{i1r_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijrt_0} - \tilde{V}_{i1r_0t_0} + z_1, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijrt_0} - \tilde{V}_{iJrt_0} + z_1, \\
&\qquad\quad e_{i1rt} - e_{ikrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt} + z_2, ..., e_{iKrt} - e_{ikrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt} + z_2) \\
&\quad = Pr(\max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq z_1, \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq z_2| \\
&\qquad\quad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&\quad = H_{jk}(z_1, z_2|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.8}$$

Now define the random variables ζ_{ijkrt} as

$$\begin{aligned}
\zeta_{ijkrt} = \Gamma_{jk}^{-1}\{&H_{jk}(0, 0|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \\
&\tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt})\},
\end{aligned} \tag{D.9}$$

where Γ_{jk} is any continuous cumulative distribution function. Based on the above transformation, the selection problem may be written as

$$\begin{aligned}
M_{ijkrt} = 1 \quad \text{iff} \quad \zeta_{ijkrt} \leq \Gamma_{jk}^{-1}\{&L_{jk}(\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt})\},
\end{aligned} \tag{D.10}$$

where $L_{jk}(...)$ is evaluated at the observed value function and utility differences.

The key step in Lee's (1983) approach is then to assume that the vector $(\epsilon_{ikrt}, \zeta_{ijkrt})$ is independent and identically distributed with joint cumulative distribution function $I_{jk}(...)$,

thereby specifying the joint distribution of outcome and selection errors $F_{jk}(...)$. Importantly, the distribution function $I_{jk}(..)$ is not allowed to vary with the observed utility and value function differences, i.e. the same transformation is applied to maximum order statistics regardless of the specific values for $\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ...$ and $\tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ...$. Dahl (2002) shows that this simplification is equivalent to the index sufficiency assumption from Section 5.1. Using this assumption, the joint distribution of outcome and selection errors $F_{jk}(...)$ may be written as

$$
\begin{aligned}
&F_{jk}(z_0, \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(K|j)rt}) \\
&= Pr(\epsilon_{ikrt} \leq z_0, e_{i1r_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\qquad\qquad e_{i1rt} - e_{ikrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., e_{iKrt} - e_{ikrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt}) \\
&= Pr(\epsilon_{ikrt} \leq z_0, \zeta_{ijkrt} \leq \Gamma_{jk}^{-1}\{L_{jk}(\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\qquad\qquad\qquad\qquad \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt})\}) \\
&= I_{jk}(z_0, \Gamma_{jk}^{-1}\{L_{jk}(\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\qquad\qquad\qquad \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt})\}). \qquad \text{(D.11)}
\end{aligned}
$$

The final step involves making parametric assumptions on the distributions $\Gamma_{jk}(...)$ and $I_{jk}(...)$. As described in Section 5.2, I follow Lee (1983) and assume that $\Gamma_{jk}(...)$ is a univariate standard normal cdf and $I_{jk}(...)$ is a bivariate standard normal cdf.

Random Forest Algorithm

Leo Breiman's and Adele Cutler's random forest algorithm belongs to the class of supervised machine learning algorithms and is commonly used in prediction problems with categorical dependent variables. Random forests operate by constructing a large number of decision trees based on different samples of observations which are combined to give as an outcome the average prediction of all trees. In doing so, random forests avoid problems of overfitting.

Individual trees are grown using an optimal splitting algorithm where explanatory variables are first selected and then split according to the algorithm, resulting in new branches starting from an original node. This process is repeated until no explanatory variable meets the selection criteria (see Hastie *et al.* (2009) for details).

In order to account for sampling variation due to the estimation of the selection probabilities when conducting inference in the outcome equations, I randomly select 50% of the individuals as training dataset. I then use the training dataset to grow separate random forests for the training and occupation choices using the explanatory variables described in Section 5.3. Both random forests are based on 500 trees, where 1000 randomly selected

observations from the training dataset are used to grow each tree. While training choice is predicted using a single observation for each individual, occupation choices are predicted using all employment spells of the selected individuals. In a second step, the resulting forests are applied to the remaining 50% of the sample, the test dataset. Probability predictions for each training or occupation option in the test dataset are computed as the proportion of counts for that option across all trees in the final nodes.

Appendix E. Task Content

Task Distance Measure

Define a task vector for each occupation k, $q_k = (q_{1k}, ..., q_{Sk})$, where q_{sk} is the fraction of workers performing task s in occupation k. Similarly, define a task vector for each training j, $q_j = (q_{1j}, ..., q_{Sj})$, where q_{sj} is the fraction of workers performing task s when being trained in training j. Assume that the composition of tasks when being trained in j is equivalent to the composition of tasks performed when working in occupation $k = j$.

Following Gathmann & Schönberg (2010), I define the angular separation between training j and occupation k as a measure of similarity using task vectors q_j and q_k:

$$AngSim_{jk} = \frac{\sum_{s=1}^{S}(q_{sj} \times q_{sk})}{[(\sum_{s=1}^{S} q_{sj}^2) \times (\sum_{s=1}^{S} q_{sk}^2)]^{1/2}}. \tag{E.1}$$

The angular separation is equivalent to the uncentered correlation or the cosine difference between two vectors and is a symmetric, purely directional measure, i.e. it is unaffected by the length of two skill vectors q_j and q_k. In contrast to that, the Euclidean distance between two vectors q_j, q_k measures the length of the vector connecting q_j and q_k and is therefore sensitive to their length. As a result, two occupations with relatively short vector lengths could be classified as similiar even when they are orthogonal. $AngSim_{jk}$ ranges between zero and one, with two orthogonal task vectors having similarity zero, and is increasing in the degree of overlap between two task vectors q_j and q_k. Following Gathmann & Schönberg (2010), I define $(1 - AngSim_{jk})$ as the *distance* between training j and occupation k:

$$Dist_{jk} = (1 - AngSim_{jk}). \tag{E.2}$$

Appendix F. Welfare and Policy

Locked-in Workers

Table F.1 summarizes the calculations to estimate the share of locked-in workers. Figure 3 in Section 6.2 shows that the return to working on versus off the diagonal falls by about $2.5pp$ between $3-12$ years of work experience. At the same time, the fraction of on-diagonal workers falls from about 70% to 60% (see Figure 2 in Section 2.5). This latter change may in part be induced by a reduction in the lock-in effect caused by the fall in returns to working on versus off the diagonal. However, other factors such as newly revealed information about own occupation-specific abilities may have contributed to the decline. Assume that other factors causing a decline in the fraction of on-diagonal workers are stable throughout a career and consider the change in the fraction of workers on the diagonal after returns have stabilized, i.e. after 12 years of work experience. Figure 2 shows that between $12-21$ years of work experience, this fraction drops by about $5pp$. This implies that about half of the drop between $3-12$ years of work experience may be associated with the fall in the returns to working on versus off the diagonal. These simple calculations therefore suggest that a $1pp$ reduction in the return to working on the diagonal leads to a $2pp$ drop in the fraction of individuals working on the diagonal. Note that this is likely going to be an upper bound on the lock-in effect since information on occupation-specific abilities is expected to be revealed at a higher rate early on in a career. In a hypothetical world without a 15% return on the diagonal, a world without lock-in effects, the fraction of individuals working on the diagonal would thus be $30pp$ lower.

Table F.1: Estimating the Share of Locked-in Workers

work exp. (years)	Δ return on diag.	Δ fraction on diag.		implied Δ fraction *not* due to Δ return	Δ fraction on diag. per Δ return on diag.
$3-12$	$-2.5pp$	$-10pp$	}	$-5pp$	$\frac{-10pp-(-5pp)}{-2.5pp}=2$
$12-21$	0	$-5pp$	}	–	–

Notes: The table summarizes the estimation of the share of locked-in workers.

Retraining Calculations

Total costs in Euros

In 2010, the average annual cost per apprentice in the dual system was around $5,280$ Euros for firms and $6,620$ Euros for all government bodies (Source: *Finanzierung der beruflichen Ausbildung in Deutschland, BWP 2/2016, BiBB.* Figures are 2012/13 cpi adjusted). In terms of private cost, the average yearly difference in earnings between an apprentice and a trained worker with less than 15 years of work experience was about $20,060$ Euros in 2010 (Source: *BiBB press release, 01/2011* and author's own calculations).

Net benefits in Euros

My estimates suggest that the annual average gain of retraining of τ corresponds to 15% of wages for the average worker. The cost of a year of foregone work experience is about 6%. Assuming that the effective foregone work experience of two years of retraining is one year (apprentices spend about two thirds of their time working in firms), the net gain of retraining is therefore equal to 9%. Based on average annual earnings of $39,000$ Euros in 2010, this amounts to $3,510$ Euros in 2010.

Cost-benefit calculations

Assuming a discount factor of 0.96, retraining costs would be recouped after about 39 years of subsequent work in the new occupation:

$$31,960 + \beta \times 31,960 = \beta^2 \times 3,510 \times \frac{1-\beta^{t+1}}{1-\beta} \tag{F.1}$$

$$t \approx 35. \tag{F.2}$$

Based on an average training completion age of 20, and a retirement age of 67, off-diagonal workers would therefore need to switch out of their training occupation with at most six years of work experience for retraining to be profitable ($67 - 20 - 2 - 35 = 10$ years). In addition, the return to working on versus off the diagonal only drops by at most $2pp$ from its peak of 15% by 10 years of experience (see Figure 3 in Section 6.2). Using the calculations for locked-in workers in this appendix, this suggests that 26% of all workers are still locked in at 10 years of experience. Based on a final share of 30%, this corresponds to a fraction of over 85% of locked-in workers.

Information Provision

Section 8.2 considers retraining as a potential policy instrument. This section briefly discusses the provision of ex-ante information as an alternative policy intervention. Note that ex-ante information provision may be a perfect substitute for costless retraining, at least in a model with only a single second-stage occupation choice. Differences arise with multiple occupation choices since workers would need to take into account the average payoffs across all occupation stages when choosing their training in a perfect foresight environment. On the other hand, with costless retraining, training choices can be readjusted each period. Given that the vast majority of workers works in at most two occupations, this distinction is unlikely to matter in the given context. With perfect information at the time of training choice, retraining costs do not impact wages. Similarly, in the absence of any retraining cost, imperfect information at the time of training choice does not affect wage outcomes.

Albeit harder to quantify, the *ex-ante* provision of information at the time of training choice is likely to be more cost-effective than *ex-post* retraining programs. In particular, my findings suggest that government programs causing high-school graduates to start training in instead of outside the occupation they will ultimately work in would generate a net benefit up to a cost of $6,300$ Euros in 2010 for every year participants will subsequently spend working on instead of off the diagonal (14% of an average of $45,000$ Euros in 2010). Moreover, using the 2010 empirical distribution of workers across training-occupation cells, I find that over 50% of off-diagonal workers could be trained in their ex-post optimal occupation without making changes to the total number of apprentices trained in each occupation. In other words, a substantial fraction of workers could have been trained in their current occupation without changing occupation-specific training capacities. Policies could include internships to provide information on own occupation-specific abilities or workshops indicating occupations that may be in high demand in the foreseeable future. While it is hard to know exactly how much *additional* information may be provided through such initiatives, the figures suggest that only a very small percentage of apprentices would need to make a better training choice for these programs to be cost-effective.