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

Dita Eckardt[*]

December 2, 2019

Job market paper

Please click here for the latest version.

Abstract

Individuals are often trained in a specific field but work in another. This paper analyses the returns to different training-occupation combinations. To this end, I use an administrative employment panel which contains the apprenticeship training for a large sample of workers in Germany. In this context, 70% of individuals with at least upper-secondary education hold apprenticeships, and 40% of these work in occupations they were not trained for. I combine the administrative data with historical data on occupation-specific vacancies to causally identify the returns. To implement the identification strategy, I set up an augmented Roy model and extend existing control function approaches to deal with selection in a two-stage, high-dimensional setting. The results suggest that workers trained in their current occupation earn $10 - 12\%$ more than workers trained outside their occupation, and that not controlling for selection leads to substantial negative bias. I find considerable heterogeneity in the estimated returns and use task content data to show that returns to training-occupation combinations are decreasing in the task distance between training and occupation. Finally, I argue that ex-ante imperfect information may lead to training choices that are suboptimal ex-post and find that, as a result, $4 - 6\%$ of wages are foregone for the average worker. Back-of-the-envelope calculations suggest that retraining programmes could be very effective in addressing this 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, London School of Economics, Email: D.Eckardt@lse.ac.uk.
I am extremely grateful to my advisors Alan Manning and Johannes Spinnewijn for their continued support and guidance. 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, and using different approaches to overcome biases typically interpreted as resulting from an omitted ability variable. More recently, a smaller but 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 new insights on the value of 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 largely survey-based. Field-occupation matches are therefore 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 of education 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 data on German apprentices. This data is a sample of German social security records 1975-2010 provided by the German Federal Employment Agency, containing information on all (un)employment spells for a randomly selected sample of workers. Importantly, since German apprentices are trained in firms for three days a week 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 row choice and occupation as column choice, with the same number of rows and columns. I estimate the returns in these cells.

[1] Examples of studies that find such correlations using subjective classifcations of field-occupation matches include Robst (2007), Nordin *et al.* (2010), Lemieux (2014), Kinsler & Pavan (2015) and Ransom (2016).

The German setting is attractive for the present analysis for two other reasons. Firstly, apprenticeships are the main form of upper- and post-secondary education in Germany, held by roughly 70% of those who obtained this level of education. Importantly, an average of 40% of those are employed in an occupation different from the one they were trained in. Secondly, I observe both the training and later occupations coded within a former German occupational classification. In contrast to many international systems, this classification has the advantage of being field-based such that promotions do not imply occupation changes.

In order to address the second challenge of causal identification, I combine the employment panel with historical data on occupation-specific apprenticeship vacancies. These vacancies are posted by firms looking to train an apprentice in a specific occupation. The data contains the universe of apprenticeship vacancies by occupation, year and region which were posted through local employment agencies between 1978-2010. I show that these vacancies are a close proxy for occupation-specific labour demand in the economy, and thereby affect earnings. For a particular training choice, I then use *expected* vacancies posted in occupations *other* than the chosen training occupation as instruments. For an occupation choice, I use subsequent shocks to these expectations in occupations *other* than the chosen one as instruments. An important advantage of using apprenticeship vacancies instead of actual job vacancies is that the majority of training firms use the agency channel and coverage is large. 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 recently documented by Arcidiacono *et al.* (2019).

To put structure on the selection problem, I set up an augmented Roy (1951) model. While the model provides a behavioural justification for the identification strategy, it does not impose any assumptions for identification, and the empirical strategy is robust to alternative selection mechanisms. In the given 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 maximise *expected* payoff including *expected* wages, occupations are chosen to maximise *current* payoff including *current* wages. Importantly, individuals have imperfect information about future labour demand and own occupation-specific abilities when choosing a training. As a result, unexpected changes in labour demand or new information about occupation-specific 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 generalisation 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. I extend their approach to two selection stages and implement this using a machine learning technique (random forests) where I predict training and occupation choices with the instruments.

The returns to different training-occupation combinations naturally depend on the granularity of the underlying occupation classification. Based on the availability of the historical vacancy data, I restrict my classification to 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 $10-12\%$ more than workers employed in occupations different from their training. The effect is highly significant and comparable in magnitude to empirical 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 a couple of percentage points before stabilising after 12 years of experience. Off-diagonal workers thus only partially catch up with their co-workers who were trained in their current occupation.

Secondly, not controlling for selection leads to substantial negative bias in the estimated returns to working on versus off the diagonal such that, observationally, on-diagonal workers have lower earnings. 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 early in their careers (Lentz *et al.* (2018)). However, I also find evidence that the bias becomes stronger over a career, 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 considerable 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 estimated returns and the fraction of workers with the relevant training observed working on the diagonal. In line with the proposed selection 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 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)). Studies using the task approach consider occupations as task vectors and argue that the transferability of human capital across occupations depends on how similar occupations are in terms of their mix of required tasks.[2] Based on this theory, one may

[2] A number of empirical papers confirm this conjecture, showing that wage drops after displacement are

expect workers in the present context to incur larger wage penalties, the more distant the occupation is from the original 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 $4pp$. Overall, these 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 key friction in the proposed framework is the lack of information about the labour market and own occupation-specific abilities at the time of training choice. Two groups of workers are affected by this friction. The first group are off-diagonal workers. These workers miss out on the return to working on the diagonal. 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 any on- versus off-diagonal return. My findings suggest that 60% of workers either work off the diagonal or are locked in, i.e. only 40% of workers hold the optimal training ex-post. Using these shares, I find that the welfare loss due to the friction is $4 - 6\%$ of wages for an average worker. To address the friction, I consider retraining programmes as a policy instrument. Back-of-the-envelope calculations suggest that these could effectively address the friction 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 labour market entry by providing workers with specialised 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 current occupation.

This paper contributes to four strands of literature. Firstly, it builds on the literature on the returns to college majors. Arcidiacono (2004) estimates a dynamic structural model of college and major choice and concludes that there are large relative earnings premiums for certain majors. Altonji *et al.* (2012) and Altonji *et al.* (2016) provide a survey 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 analysing the important heterogeneity that returns to fields display across occupations.

larger for workers who move to occupations which are less related to the previous one (see e.g. Poletaev & Robinson (2008), Gathmann & Schönberg (2010)).

In doing so, I consider an additional selection stage, the selection into occupations. To address the challenges arising from the selection into both trainings and occupations, this paper contributes to a second strand of literature on high-dimensional selection models. Heckman & Robb (1985) show that, in single-index self-selection models, control functions may be written as functions of the propensity to self-select.[3] In multiple-index models, Lee (1983) shows that the selection problem may be written as a single-index model using highest order statistics, and develops a parametric control function estimator. Using a similar idea, Dahl (2002) imposes an index sufficiency assumption to develop a non-parametric control function approach where the control function becomes a function of a small set of selection probabilities only. I contribute to this literature by extending the Lee/Dahl approach to a two-stage selection setting, and implementing it using machine learning techniques.

Thirdly, and more generally, this paper relates to the literature on human capital specificity. The idea that human capital is specific was first proposed by Becker (1964) in the context of the firm, and has subsequently 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.* (forthcoming)). 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 strand of 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)). 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 organised as follows. Section 2 outlines the setting of the German apprenticeship system and discusses the data. Section 3 sets up the augmented Roy model. Section 4 discusses the identification strategy. 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 the welfare and policy implications of my results. 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 length of an apprenticeship 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 standardisation across states. 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 workers into the labour market. While enrolment rates have been declining in recent years, still about 60% of high-school graduates took up an apprenticeship in the dual system in 2011.[5]

Most non-university occupations are included in the dual system, with a few exceptions in the medical and care occupations.[6] In order 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. These schools are run and financed by the state so that the financial burden of apprenticeships in the dual system is split between the private and public sectors. The curriculum is determined centrally by each state for each apprenticeship occupation (*Rahmenlehrplan*) and consists of general and specialised subjects which may vary depending on the apprenticeship occupation.

All apprenticeships in the dual system are completed through a final examination which is organised and monitored by the chambers (*Kammern*) and consists of a theoretical and

[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: *Statistisches Bundesamt, Berufsbildung auf einen Blick, 2013.*

[6] For example, training as a physiotherapist is provided in specialised schools outside the dual system.

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).[7]

2.2 Data

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

2.2.1 Employment Panel

The employement panel dataset consists of a 2% sample of all German social security records between 1975-2010 (*Stichprobe der Integrierten Arbeitsmarktbiografien*) provided by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). 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.[8] 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 to four days a week, they have to pay social security contributions and their apprenticeship spells are contained in the employment panel dataset. I therefore observe the occupation registered for an apprentice while completing the apprenticeship and define this as the worker's training.

2.2.2 Vacancy Data

I use a second dataset for my analysis which contains the universe of all apprenticeship vacancies posted through any local German employment agency between 1978-2010.[9] Between

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

[8] During the sampling period, Germany did *not* have a general minimum wage. Instead, the majority of firms were bound by minimum wages set through industry-specific tariff contracts which varied by region and occupation.

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

1978-1992, the data only covers West Germany. From 1993, vacancies are recorded for both West and East Germany. In addition to firm-based vacancies, the data includes external apprenticeship vacancies which are not posted by a firm (e.g. by nursing schools), but these make up a small fraction of the data ($\sim$ 12%).[10] Recorded vacancies include those which are filled and those which are not filled after a year and aggregate information is available by year, training and location. Yearly data is measured as a flow of vacancies between 1. October and 30. September.[11]

Occupation-specific apprenticeship vacancies are a close proxy for occupation-specific labour demand in the economy and, in 2010, the correlation coefficient between occupation-specific apprenticeship and non-apprenticeship vacancies posted through local employment agencies was 0.82. At the same time, 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.[12] Moreover, in contrast to most available data on general job vacancies, the degree of involvement of local employment agencies for apprenticeship vacancies (*Einschaltungsgrad*) is high. In 2013, 71% of firms publicised their apprenticeship vacancies through an agency, while the same figure only amounted to 45% for all non-apprenticeship vacancies.[13]

2.3 Field-Based Occupational Classification

Occupations in the employment panel and the vacancy dataset are coded based on the same occupational classification called the *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 thesis, the *KldB88* has a key advantage over the newer national system and other internationally used systems such as the *International Standard Classification of Occupations (ISCO)* or the *Standard Occupational Classification (SOC)* in that it is *field-based.* The other systems contain a broad category for *Managers* and as a

[10] Figures based on 2010. Source: *Arbeitsmark in Zahlen - Bewerber und Berufsausbildungsstellen Deutschland, September 2010.*

[11] For example, the data for 2010 contains all vacancies which are posted between 1. October 2009 and 30. September 2010. In practice, given the high-school graduation date in summer, most vacancies ($\sim$ 60%) are posted between April and September (Source: *Statistik zum Ausbildungsstellenmarkt - Bewerber für Berufsausbildungsstellen und Berufsausbildungsstellen - Zeitreihe).*

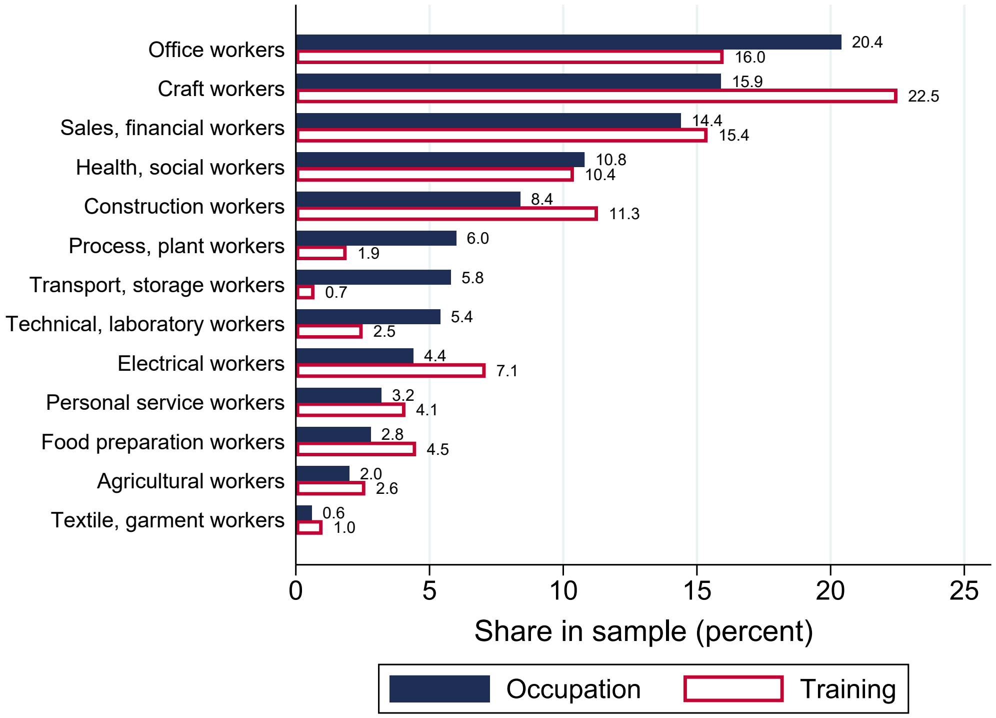
[12] As an example, job vacancies advertised as health care occupations may refer to doctors and nurses, but only those as nurses would be relevant for a workers who does not hold a medical degree.

[13] 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.

result, being promoted might 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.

In order to make the estimation both feasible and tractable, I will use a level of the *KldB88* which classifies occupations into 13 distinct categories, implying 169 distinct cells in the training-occupation matrix. The chosen level of categorisation is the finest for which the historical vacancy dataset is available for a sufficiently long time period. 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.

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 Appendix A.1.

2.4 Sample Selection

The full Sample of Integrated Employment Biographies contains 1,594,466 individuals and 41,390,318 observations. Since individuals can have more than one employment relationship at a time, some of the spells are overlapping. I define the main employment spell as the highest wage spell and drop all secondary spells (18.79%) from the sample. Of the remaining individuals, I only keep those who were enrolled in exactly one apprenticeship in the dual system and whose apprenticeship period falls into the sampling period (26.19% of individuals). I thus drop all spells belonging to individuals without any apprenticeship training during the sampling period (69.2%), individuals who were enrolled in two apprenticeship with distinct training occupations (4.02%)[14], and individuals enrolled in three or more apprenticeships (< 1%). In order to ensure that the apprenticeship was completed, I further drop all 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.

I restrict the remaining spells to full-time spells (86.17%) and exclude all spells with missing location, occupation or missing (or zero) wage.[15] 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, 7.41%) and higher education levels (additional university or technical college degree, 5.43%) 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 all individuals work outside their training occupation for at least one spell, and 38% work in more than one occupation throughout their career. The spell length varies depending on the reason why a new spell is registered with the Federal Employment Agency. Most commonly, a new spell is registered as part of the compulsory annual notification of employment.[16]

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

[14] Included in this figure are those individuals with missing occupation entry for part of their apprenticeship training. Since the missing spells could refer to training in a different occupation, I exlude these individuals.

[15] Zero wages are used in the data to indicate interrupted employment spells. This could be due to illness after wage continuation, maternity leave or sabbaticals.

[16] Spells can have a shorter length if a worker leaves the employer or a notification is required e.g. due to an occupation change with the same employer or switch from full-time to part-time employment.

Table 1: Summary Statistics

	Mean	Min	Max	P^{10}	P^{50}	P^{90}
Individuals ever off diagonal (%)	47.6					
Occupation switchers (%)	37.7					
Occ. switches per individual	0.7	0	38	0	0	2
Distinct occ. per individual	1.5	1	10	1	1	2.5
Spell length (days)	255	1	365	31	365	365
Daily wages in 2010 Euros	81	1	186	39	77	130
Age	30.6	17	62	20.5	28.5	42.5
Female (% of individuals)	45.4					
Female (% of spells)	37.3					
N of observations/spells	4,012,034					
N of individuals	291,098					

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

that less than 3% of wages exceed the upper earnings limit in the statutory pension scheme and are capped in the sample. I provide robustness checks for my results excluding these capped wages from the sample (see Section 6.5). About 45% of all individuals are female.

To give a sense of the distribution of individuals across training-occupation cells, Table 2 reports the number of spells in each training-occupation cell for the five largest trainings as a percentage of all training spells. Table 3 reports the same figures for the five largest occupations as a percentage of all occupation spells.[17] Both tables are restricted to spells with ten years of full-time work experience. It can be seen that, while the majority of individuals work on the diagonal where the training is the same as the occupation, the fraction of individuals working off the diagonal is large. This may seem surprising in the German context where the apprenticeship system is widely believed to facilitate entry into worklife by providing apprentices with the necessary skills for a specific occupation. Tables 2 and 3 also document considerable variation in the share working on the diagonal across trainings and occupations. For instance, while only 55.2% of individuals with ten years of work experience who were trained as *craft workers* also work in this occupation (see Table 2), 84.2% of those working as *craft workers* with the same level of experience were also trained in this occupation (see Table 3). On the other hand, these same figures amount to 80.6% and 59.5% for *office workers*.

[17] Since only selected categories are reported, the row percentages in Table 2 and column percentages in Table 3 do *not* sum to 100. Equivalent tables containing all 13 occupations are provided in Appendix A.2.

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	12.5	1.6	0.1
	Craft workers	4.8	55.2	3.9	2.4	2.5
	Sales, financ. w.	26.5	1.6	60.5	2.2	0.3
	Health, social w.	12.2	0.7	4.3	79.1	0.2
	Construction w.	3.5	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. Occupations are classified using the 13 category baseline classification. Only the five most common trainings and occupations are reported. As a result, row percentages do not sum to 100. A full table containing all 13 occupation categories is available in Appendix A.2.

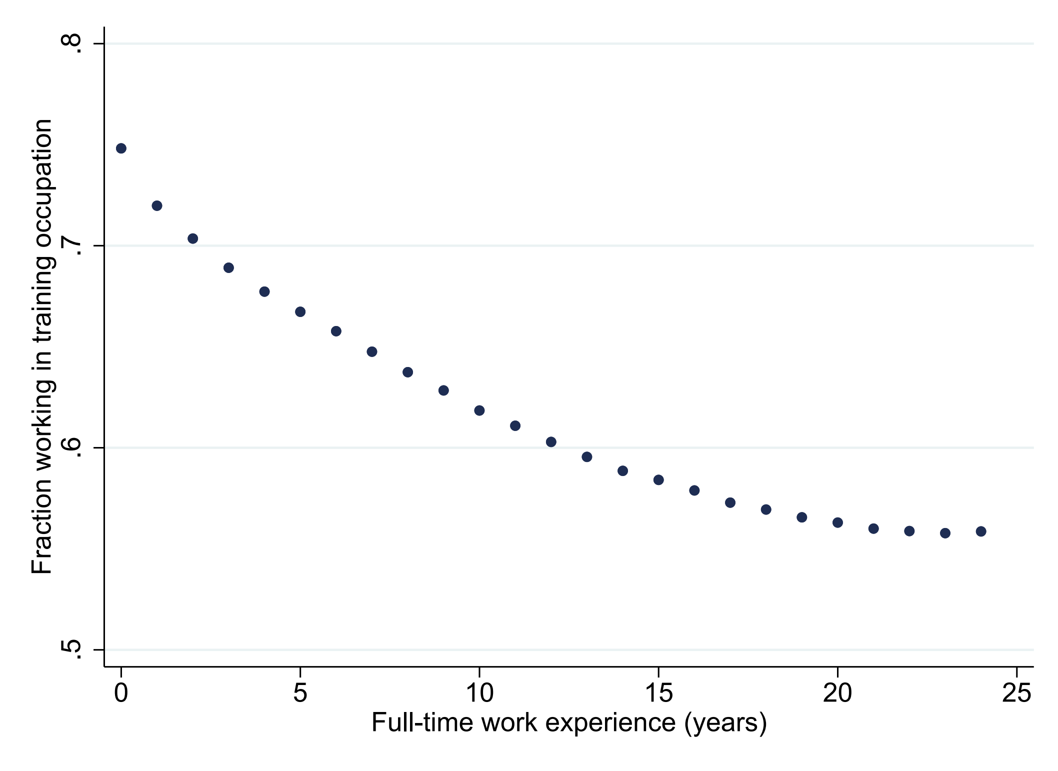
Table 3: Spells as Percentage of Occupations - Selected Categories

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	59.5	0.7	14.4	2.8	0.3
	Craft workers	5.0	84.2	6.2	5.7	7.6
	Sales, financ. w.	18.2	1.7	64.8	3.4	0.6
	Health, social w.	5.0	0.4	2.7	76.3	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. Occupations are classified using the 13 category baseline classification. Only the five most common trainings and occupations are reported. As a result, column percentages do not sum to 100. A full table containing all 13 occupation categories is available in Appendix A.2.

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 occupation by full-time work experience. While, within the 13 occupation classification system, around 75% of all workers start their career after the apprenticeship working in their training occupation, this fraction drops to around 65% after 6 years, and around 55% after 25 years of full-time work experience. To alleviate concerns that this relationship could be confounded by changes in the task content of occupations which are not picked up by the classification system, Figure A.1 in Appendix A.3 shows the fraction of individuals working on the diagonal over time for the three different career stages. It can be seen that, over the thirty year sampling period, these fractions have remained remarkably stable.

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. Occupations are classified using the 13 category baseline classification.

3 Selection Model

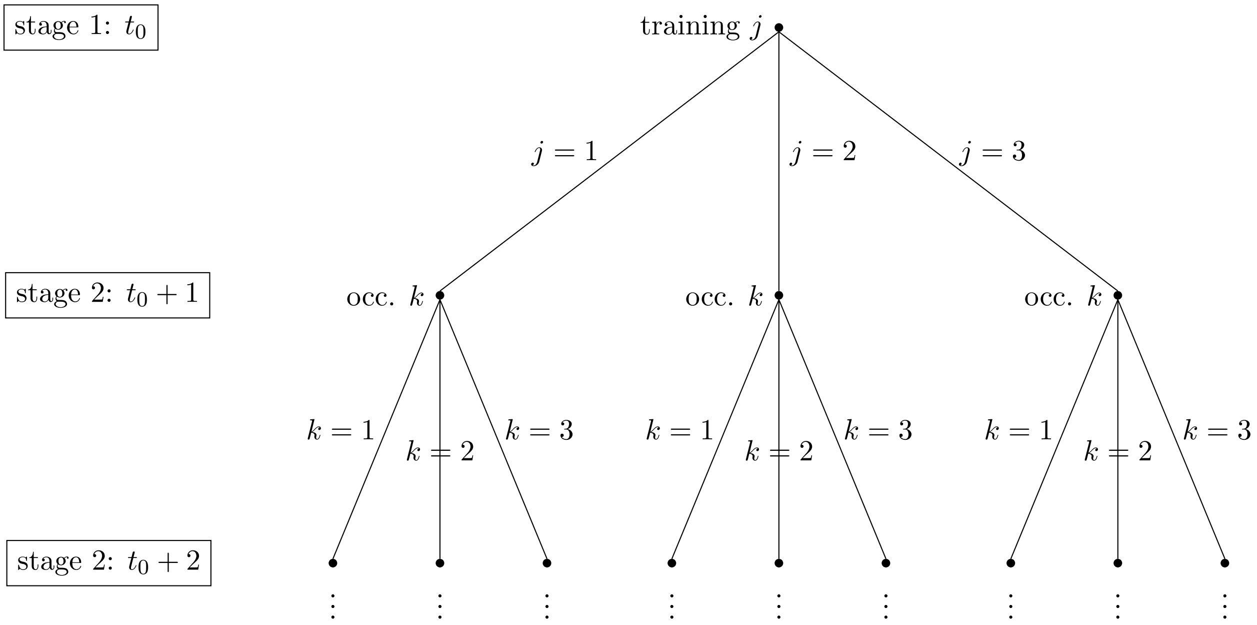
I model the selection into training and occupations using a generalised two-stage Roy (1951) model. The model will provide a behavioural justification for the identification strategy proposed in Section 4, and put sufficient structure on the choice problem to implement the identification strategy using a control function approach. It does not, however, impose any assumptions required for identification, all of which will be discussed in Section 4 below.

The threshold-crossing nature of the proposed model will be key to the estimation approach described in Section 5. While I suggest a specific model that includes this feature, note that alternative threshold-crossing models could have been used.[18] Importantly, this implies that, as long as the identification assumptions from Section 4 hold, my empirical approach is robust to alternative selection models.

3.1 General Setup

The structure of the sequential choice problem is depicted in Figure 3.

Figure 3: Model Structure



Notes: The figure illustrates the structure of the selection model described in Section 3.

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 $k = 1, ..., K$.

[18] An example would be a search model where workers pay a search cost in order to enter a specific occupation which depends negatively on the number of available vacancies in that occupation.

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. To outline the selection into training and occupations, I start by specifying wages in Section 3.2. Given the sequential nature of the selection problem, I will then solve the model backward, starting with occupation choices in Section 3.3 before discussing training choice in Section 3.4.

3.2 Wages

I assume a standard human capital model where wages w_{ijkrt} of individual i, trained in occupation j, working in occupation k in region r at time t are given by

$$w_{ijkrt} = \bar{W}_{krt} H_{ijkrt} \tag{1}$$

$$ln(w_{ijkrt}) = \bar{w}_{krt} + h_{ijkrt}. \tag{2}$$

H_{ijkrt} denotes human capital of individual i and $\bar{W}_{krt}$ denotes the skill price per unit of human capital. Lower case letters $\bar{w}_{krt}$ and h_{ijkrt} denote the corresponding log variables.

The log skill price $\bar{w}_{krt}$ is assumed to take the following form:

$$\bar{w}_{krt} = \delta_r + \delta_t + f(vac_{krt}), \tag{3}$$

where δ_r and δ_t are region and time fixed effects, respectively, and $f(vac_{krt})$ denotes a flexible function in log vacancies vac_{krt} posted for occupation k in region r at time t. While $f(vac_{krt})$ captures the impact of changes in labour demand on log skill price, the fixed effects summarise any exogenous effect through labour supply.

Following Griliches (1977), I assume an exponential human capital production function

$$H_{ijkrt} = e^{(\delta_i + \tau_{jk} + \beta' X_{it} + \iota_{ik})} e^{(\nu_{ijkrt})} \tag{4}$$

$$h_{ijkrt} = \delta_i + \tau_{jk} + \beta' X_{it} + \epsilon_{ijkrt}, \tag{5}$$

where δ_i denotes an individual-invariant fixed effect, X_{it} is a vector of controls, and $\epsilon_{ijkrt} = \iota_{ik} + \nu_{ijkrt}$ is the sum of non-random and random productivity effects, respectively. The non-random productivity effects ι_{ik} may be interpreted as occupation-specific abilitiy of individual i. The $(J \times K)$ fixed effects τ_{jk} are the parameters of interest. These parameters capture the producitivity effect from a particular combination of training j and occupation k. For ease of interpretation, I start the empirical analysis by parameterising the effect τ_{jk} to capture the average effect of working on versus off the diagonal, before going on to estimate the effect of all training-occupation combinations non-parametrically. Combining equations (3) and (5), I will estimate the following specifications for log wages:

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \tau D_{j=k} + \beta' X_{it} + \epsilon_{ijkrt}, \tag{i}$$

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \tau^{exp} D_{j=k} + \beta' X_{it} + \epsilon_{ijkrt}, \tag{ii}$$

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \tau_j D_{j=k} + \beta' X_{it} + \epsilon_{ijkrt}, \tag{iii}$$

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

where X_{it} may include full-time work experience exp and its square, and occupation-specific work experience exp_k and its square, depending on the specification.[19] Model (i) sets $\tau_{jk} = \delta_k + \tau D_{j=k}$, where $D_{j=k}$ is a dummy variable equal to one if training j is the same as occupation k, and parameter τ captures the average effect of working on versus off the diagonal. Model (ii) explores the heterogeneity in τ, and estimates separate effects τ^{exp} for each yearly full-time work experience bin.[20] Model (iii) estimates separate effects τ_j for each training j. Finally, model (iv) estimates the fixed effects for each training-occupation combination. Note that, since individuals in the sample complete exactly one training, the inclusion of individual fixed effects implies that parameters τ_{jk} correspond to the within-training returns.[21] In the following, the parameters of interest from models (i) to (iv) will jointly be referred to as $\boldsymbol{\tau} = (\tau, \tau^{exp}, \tau_j, \tau_{jk})$.

3.3 Occupation Choice

Define the utility when working in occupation k at time $t > t_0$ as an additively separable function of log wages and tastes t_{ijkrt}:

$$U_{i(k|j)rt} = ln(w_{ijkrt}) + t_{ijkrt}, \tag{6}$$

where the subscripts are chosen to emphasise the conditioning on training choice j. Assume that individual i observes $U_{i(k|j)rt}$, including current vacancies vac_{krt}.

Conditional on training choice j in $t = t_0$, individual i chooses occupation k in $t =$

[19] The inclusion of exp_k as control makes the choice model dynamic and can lead to an additional source of endogeneity since past occupation-specific work experience is equivalent to past selection into occupations. See Appendix E for details on the inclusion of exp_k and the implications for estimation in the current context.

[20] More specifically, τ^{exp} denote the parameters on the interaction between yearly experience bins and the dummy variable $D_{j=k}$. I also estimate a model to explore the heterogeneity in parameter τ by *occupation-specific* work experience exp_k. Details on the model and estimation can be found in Appendix E. The results are presented in Appendix F.1.1

[21] Estimating within-occupation returns requires identifying the training fixed effects separately from the individual fixed effects. See Appendix C.2 for details. Results are presented in Appendices F.1.4 and F.1.5.

$t_0+1,...,t_0+T$ to maximise the current period-t utility. The occupational choice problem will therefore be static, but this depends on the specification of log wages outlined in models (i) to (iv) rather than a restriction on individual i's foresight. For instance, the inclusion of occupation-specific experience exp_k in the wage regressions leads to current choices affecting future utilities, thereby making the occupational choice problem dynamic from the individual's perspective. Note that this only affects the empirical approach in that it adds a potentially endogenous regressor (exp_k) to the wage equation. As noted above, fully specifying the selection mechanism is *not* key to the empirical approach as long as the identification assumptions from Section 4 hold. I therfore focus on the static model in the main body of the paper, and present a dynamic version of the occupation choice problem in Appendix E.1.

Assume that tastes t_{ijkrt} are a function of observed characteristics X_{ijkrt}, and define the error component ψ_{ijkrt} as

$$\psi_{ijkrt} = t_{ijkrt} - E[t_{ijkrt}|X_{ijkrt}]. \tag{7}$$

The utility of choosing occupation k conditional on training j may then be written as

$$U_{i(k|j)rt} = \tilde{U}_{i(k|j)rt} + e_{ijkrt}, \tag{8}$$

where $\tilde{U}_{i(k|j)rt} = E[ln(w_{ijkrt})|X_{ijkrt}] + E[t_{ijkrt}|X_{ijkrt}]$ is the component of utility observed to the researcher, the sub-utility function, and $e_{ijkrt} = \epsilon_{ijkrt} + \psi_{ijkrt}$ is the unobserved utility component. Individual i chooses occupation k in period t if and only if

$$(e_{ijkrt} - e_{ijk'rt}) > (\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt}), \quad \forall k' \neq k. \tag{9}$$

Using equations (8) and (9), an occupation dummy variable $occ_{i(k|j)rt}$ may be defined as

$$\begin{aligned} occ_{i(k|j)rt} &= \begin{cases} 1 & \text{iff } U_{i(k|j)rt} \geq U_{i(k'|j)rt}, \quad \forall k', \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{iff individual } i \text{ observed in occupation } k \text{ conditional on training } j \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

3.4 Training Choice

For ease of exposition, assume that the current-period utility of training choice j in $t = t_0$ takes the same form as the utility of choosing occupation $k = j$.[22] Denote this utility by $U_{ijr_0t_0}$. Define the value function of training choice j in $t = t_0$ and denote this by $V_{ijr_0t_0}(\Omega_{r_0t_0})$, where $\Omega_{r_0t_0}$ is the information available in region $r = r_0$ at time $t = t_0$. For any training choice j, $V_{ijr_0t_0}(\Omega_{r_0t_0})$ may be written as

$$V_{ijr_0t_0}(\Omega_{r_0t_0}) = \tilde{U}_{ijr_0t_0} + e_{ijr_0t_0} + \beta E_{t_0}[V_{ijr(t_0+1)}|\Omega_{r_0t_0}], \tag{11}$$

where β denotes a discount factor and $E_{t_0}[V_{ijr(t_0+1)}|\Omega_{r_0t_0})]$ is the maximal expected reward in $t = t_0 + 1$, conditional on training choice j in $t = t_0$. Note that while current vacancies vac_{krt} are observed to individual i there is uncertainty about future vacancies $vac_{krt'}$, $t' > t$. Moreover, in $t = t_0$, there is uncertainty regarding own occupation-specific ability ι_{ik}.[23]

Individual i maximises expected utility in $t = t_0$ and chooses training j if and only if

$$\begin{aligned}(e_{ijr_0t_0} - e_{ij'r_0t_0}) >& (\tilde{U}_{ij'r_0t_0} + \beta E_{t_0}[V_{ij'r(t_0+1)}|\Omega_{r_0t_0}]) - (\tilde{U}_{ijr_0t_0} + \beta E_{t_0}[V_{ijr(t_0+1)}|\Omega_{r_0t_0}]) \\ &= \tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0}, \quad \forall j' \neq j, \end{aligned} \tag{12}$$

where $\tilde{V}_{ij'r_0t_0}$ denotes the conditional value function $V_{ijr_0t_0}(\Omega_{r_0t_0}) - e_{ijr_0t_0}$. Using equations (11) and (12), a training dummy variable $train_{ij}$ may be defined as follows:[24]

$$\begin{aligned} train_{ij} &= \begin{cases} 1 & \text{iff } V_{ijr_0t_0}(\Omega_{r_0t_0}) \geq V_{ij'r_0t_0}(\Omega_{r_0t_0}), \quad \forall j', \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{iff individual } i \text{ observed in training } j \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{13}$$

Note that, unlike the occupation selection problem, the training selection problem will be dynamic unless today's training choice does not affect future wages across occupations, i.e. unless the parameters of interest $\boldsymbol{\tau}$ are equal to zero.

[22] Choosing training j is therefore contemporaneously equivalent to choosing occupation $k = j$ in $t = t_0$.

[23] This may be modelled as additional noise parameter in $\psi_{ijkr_0t_0}$.

[24] For ease of exposition, the region and time subscripts have been omitted for the training dummy as training is only chosen once at the start of a career. This is in contrast to the occupation dummy which can vary across time as individuals change occupations.

4 Identification

This section explains the potential biases resulting from workers' self-selection into a training and an occupation, provides intuition for these biases using two hypothetical experiments, and discusses the assumptions required for the proposed instrumental variable strategy.

4.1 Selection Biases

Wages given in models (i) to (iv) will only be observed for a sample of individuals who selected into training j and occupation k, and the non-random allocation of individuals to training-occupation cells may lead to sample selection biases.

Based on the definition of the training and occupation dummies in equations (13) and (10), the selection problem may be written as

$$\begin{aligned} E[\epsilon_{ijkrt}|train_{ij} = 1, occ_{ij(k|j)rt} = 1] \neq 0, \\ E[\epsilon_{ijkrt}|M_{ijkrt} = 1] \neq 0, \end{aligned} \tag{14}$$

where $M_{ijkrt} = train_{ij} \times occ_{i(k|j)rt}$ is an indicator taking value one if individual i is observed in training-occupation cell jk. As a result, naive estimation of models (i) to (iv) will likely result in biased estimates of the parameters of interest $\boldsymbol{\tau}$.

The key insight from the hypothetical experiments discussed in Section 4.2 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 explained since off-diagonal workers must be especially good in their chosen occupation to compensate for their lack of training.

4.2 The Ideal Experiment

Given the two-stage selection in the present context, identifying the effect of a particular training-occupation combination on wages requires randomising 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 randomisation at the training or the occupation stage alone will not be sufficient to identify the parameters of interest, consider the following example of two hypothetical experiments which look at the two selection biases into training and into occupations independently. For ease of illustration, focus on a stylised version of model

(i) with only two possible trainings and occupations, $j, k \in \{1, 2\}$, and two time periods, $t \in \{0, 1\}$. The parameter of interest $\tau > 0$ captures the average log wage effect of working on versus off the diagonal. 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/working in occupation k. Further denote by $\epsilon_{i1}, \epsilon_{i2}$ the error terms in $t = 1$ in occupations 1 and 2.[25] For simplicity, assume that the error terms are known in $t = 0$. Finally, assume that individuals self-select into a training and occupation based on a simplified version of the Roy (1951) model above where the training is chosen in $t = 0$ to maximise *expected* log wages, $E_{t=0}[ln(w_{it=1})]$, and the occupation is chosen in $t = 1$ to maximise current log wages $ln(w_{it=1})$.

4.2.1 Selection into Training

In order to analyse the 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 they will be randomly allocated to occupations when making their training choice in $t = 0$. Focusing on occupation $k = 1$, the resulting selection bias when estimating parameter τ may be written as

$$\begin{aligned} &E[\epsilon_{i1}|train_{i1} = 1, occ_{i1} = 1] - E[\epsilon_{i1}|train_{i1} = 0, occ_{i1} = 1] \\ &\quad = E[\epsilon_{i1}|train_{i1} = 1] - E[\epsilon_{i1}|train_{i1} = 0] \\ &\quad = E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) > 0}_{\text{chose training 1}}] - E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) < 0}_{\text{chose training 2}}] \geq 0, \end{aligned} \tag{15}$$

where the difference in the observed component of expected log wages has been normalised to zero.[26] The final inequality follows from the assumptions made on the error terms (see Appendix B.1 for a proof of this result). Intuitively, comparing individuals working in occupation 1 who previously selected into training 1 to individuals working in occupation 1 who previously selected into training 2 will result in estimates of τ which are upward biased as those with higher ability in occupation 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.[27]

[25] 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.

[26] Alternatively, one could consider on- versus off-diagonal workers in the same *training*. Under the given assumptions, the qualitative conclusions would be unchanged.

[27] Note that, unless specific assumptions are made on the distribution and correlation structure of the error terms, the sign of the bias is ambiguous, i.e. the selection into *training* could lead to positive or negative bias in parameter τ.

4.2.2 Selection into Occupations

Now consider a second hypothetical experiment where individuals are randomly allocated to a training in $t = 0$, and can subsequently choose their occupation in $t = 1$. Again focusing on occupation $k = 1$, the selection bias when estimating parameter τ may be written as

$$\begin{aligned} & E[\epsilon_{i1}|train_{i1} = 1, occ_{i1} = 1] - E[\epsilon_{i1}|train_{i1} = 0, occ_{i1} = 1] \\ &\quad = E[\epsilon_{i1}|(occ_{i1} = 1|train_{i1} = 1)] - E[\epsilon_{i1}|(occ_{i1} = 1|train_{i1} = 0)] \\ &\quad = E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) > -\tau}_{\substack{\text{choose occupation 1} \\ \text{cond. on training 1}}}] - E[\epsilon_{i1}|\underbrace{(\epsilon_{i1} - \epsilon_{i2}) > \tau}_{\substack{\text{choose occupation 1} \\ \text{cond. on training 2}}}] \leq 0, \end{aligned} \tag{16}$$

where the difference in observed log wages net of τ has been normalised to zero.[28] As before, the final inequality follows from the assumptions made on the error terms (see Appendix B.1 for a proof of this result). Workers who choose occupation 1 in $t = 1$ conditional on having been randomly allocated to training 1 in $t = 0$ are positively selected. These workers realise the benefit τ from working on the diagonal where $j = k$ by choosing occupation 1. On the other hand, workers who have previously been allocated to training 2 would have realised 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 received training in occupation 1. Intuitively, workers working off the diagonal after random allocation to their 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.[29]

4.3 Instrumental Variables

The high dimensionality of the selection problem poses a challenge to identification in the present setting. An instrumental variables strategy needs to randomly allocate individuals to one of J trainings, and one of K occupations.

To address this challenge, I use the historical occupation-specific apprenticeship vacancy data described in Section 2.2.2. The general idea behind this approach is to use changes in occupation-specific labour demand to generate random variation in training and occupation choices. As discussed in Section 2.2.2, using data on apprenticeship vacancies as

[28] As before, one could consider on- versus off-diagonal workers in the same *training*. Under the given assumptions, the qualitative conclusions would be unchanged.

[29] Note that, unless specific assumptions are made on the distribution and correlation structure of the error terms, the sign of the bias is ambiguous, i.e. the selection into *occupations* could lead to positive or negative bias in parameter τ.

opposed to non-apprenticeship vacancies has the advantage of reducing measurement error and maximising data coverage in the present context. At the same time, as shown in Section 2.2.2, occupation-specific apprenticeship vacancies (henceforth *vacancies*) are a close proxy for occupation-specific labour demand more generally.

Recall that r_0 and t_0 denote the region and time period in which an individual starts their apprenticeship. While future expected vacancies in occupations *other* than the chosen training serve as instruments for the training choice, subsequent shocks to these vacancies in occupations *other* than the chosen one serve as instruments for the occupation choice.[30] Denote these sets of instruments by IV_{train_j} and IV_{occ_k}, and define them as follows:

$$IV_{train_j} = E_{t_0}[vac_{j'r(t_0+\tau)}|\Omega_{r_0t_0}], \qquad \forall j' \neq j, \quad \forall \tau = 0, ..., 30, \tag{17}$$

$$IV_{occ_k} = (vac_{k'rt} - E_{t_0}[vac_{k'rt}|\Omega_{r_0t_0}]), \qquad \forall k' \neq k, \tag{18}$$

where, as in Section 3, $\Omega_{r_0t_0}$ denotes the information available at the time of training choice. Under the standard IV assumptions discussed below (see Section 4.4), the interaction of all instruments will provide sufficient variation to identify the $(J \times K) - 1$ parameters in the training-occupation matrix. A behavioural justification for this identification strategy is given by the two-stage Roy (1951) model presented in Section 3. Intuitively, if individuals consider their comparative advantage when making their training choice, *expected* labour demand *outside* the chosen training j will affect the choice of training. Similarly, *shocks* to labour demand *other* than the chosen option will affect the choice of occupation.

4.4 Identification Assumptions

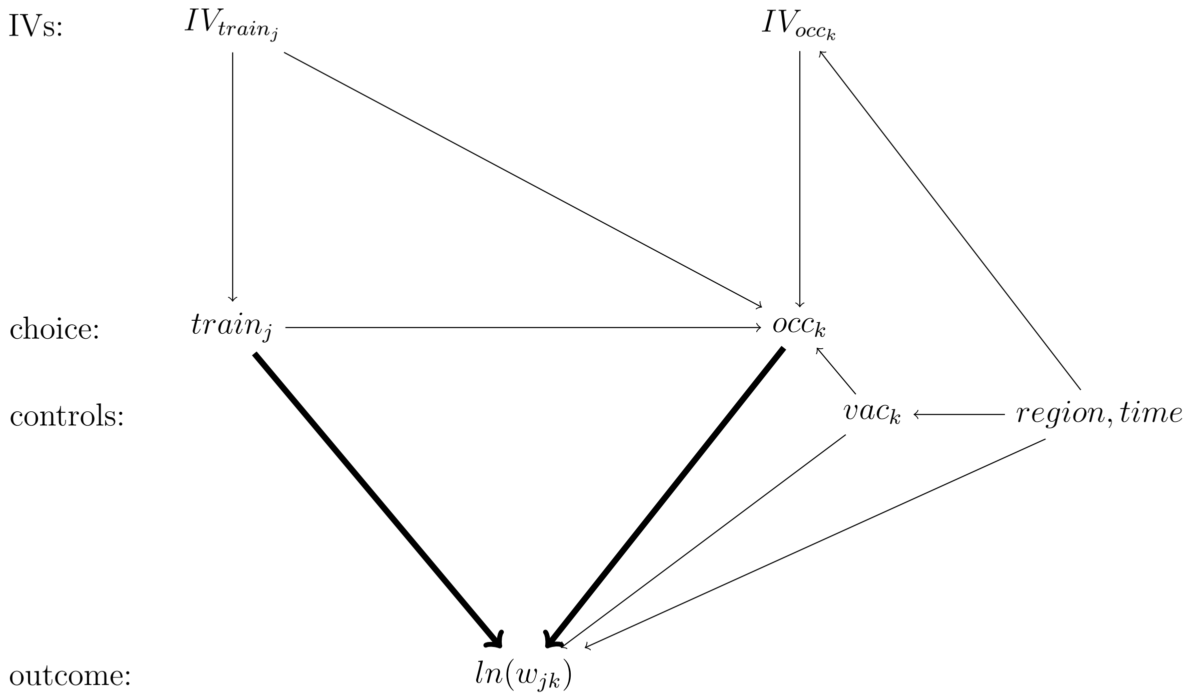
The identification assumptions for the training and occupation instruments are summarised in Figure 4. All arrows represent direct causal effects of one variable on another. The bold arrows are the causal parameters of interest. For ease of exposition, individual, time and region subscripts have been omitted.

4.4.1 Conditional Independence and Exclusion

The exclusion restriction and conditional independence assumption state that the instruments do not affect the outcome other than through the selection into a particular jk-cell, and are as good as randomly assigned relative to the outcome variable. In Figure 4, this is visualised by the absence of further arrows that start at the instruments and point to the outcome variable, or arrows connecting the instruments with the outcome variable.

[30] See Section 5.4 for details on how I split vacancies into expectations and shocks.

Figure 4: Identification Assumptions



Notes: The figure illustrates the identification assumptions described in Section 4.

A potential threat to the exclusion restriction is given by general equilibrium feedback effects through supply. Even though vacancies *outside* the chosen option do not directly affect wages, they may lead to changes in occupation-specific labour supply which may affect wages over time. Within a time period, such feedback effects are unlikely and therefore do not pose a concern for the occupation instruments. In the long-run, there may be feedback effects from the training instruments as changes in the expected labour demand affect current labour supply. The main related concern would be trends in relative occupation size over time. To address this concern, I show that my results are robust to including occupation *times* time fixed effects in the baseline model (see Table F.5 in Appendix F.2).

The key assumption underlying random assignment is that, within a given region and time period, changes in labour demand are caused by random productivity shocks to an occupation. To rule out strategic vacancy setting motives, each firm needs to be small relative to the market. This is true empirically where around three quarters of apprentices are trained in small and medium-sized firms.[31] Moreover, random assignment requires that, within region and time period, own vacancies effectively summarise all information that is relevant for occupation-specific wages. If this assumption fails, correlated shocks across occupations

[31] Around 50% are trained in small firms with less than 50 employees, a further 23% are trained in medium-sized firms with more than 50 and less than 250 employees (see Figure A.2 in Appendix A.4). Source: *Bundesagentur für Arbeit.*

can be problematic as they may imply confounding correlations between the instruments and the outcome.[32] An obvious concern would be industry-specific shocks that systematically affect vacancies across occupations. In a robustness check, I therefore explicitly control for industry *times* time fixed effects in the estimation (see Table F.5 in Appendix F.2). My results are unchanged, giving strong support to the identification assumptions made.[33]

Since occupation shocks may display serial correlation, a further potential threat to conditional random assignment would be any impact of past shocks on current wages arising from feedback effects through labour supply. Insofar as those are related to trends in relative occupation sizes, they should be accounted for by the robustness check discussed above.

Finally, conditional random assignment rules out systematic relocation of individuals as a result of labour 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.[34] 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.

4.4.2 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 Figure 4, this is visualised by arrows leading from the instruments to the choice variables. In the context of categorical endogenous variables, first stage F-statistics are not available, and 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 5.5 for details on the estimation of the selection probabilitities). Histograms of the selection probabilities into the five largest trainings and occupations are shown in Appendix C.1. It can be seen that, for both trainings and occupations, there is considerable variation in the selection probabilities and, on average, the variation generated by the instruments is above 25*pp*. Similar ranges have been found by Hull (2018), who concludes that there is sufficient first stage variation.

[32] Similarly, the exclusion restriction rules out any spillover effects from vacancies in occupation k' on vacancies in occupation k. Given that information on posted vacancies is not made publicly available immediately, such spillovers seem unlikely within any given year.

[33] 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.

[34] Source: *Datenreport zum Berufsbildungsbericht 2016*. Figure represents the population-weighted average across states in 2016.

5 Estimation

The implementation of identification strategies in high-dimensional selection models poses a number of challenges. Given the categorical nature of the selection variable, simple linear 2SLS estimation is not feasible. Instead, I will adopt a control function approach to implement the identification strategy proposed in Section 4. The general idea behind control function estimators is to model the endogenous component of the regression error term and control for it in the estimation. In the present context, define $\lambda_{jk}(...)$ as the appropriate control function using equation (14):

$$\lambda_{jk}(...) = E[\epsilon_{ijkrt}|M_{ijkrt} = 1], \tag{19}$$

where $\lambda_{jk}(...)$ depends on a set of variables further defined below. Note that it is important to allow the control function to vary across jk-cells.

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 the outcome and selection error terms. My approach in the present setting will be to use both a non-parametric and a parametric control function estimator which are implemented using assumptions on the joint distribution of the outcome and selection errors to reduce the dimensionality of the problem. This 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

To reduce the dimensionality of the problem, first note that it is possible to write the selection problem in terms of the vector of utility and value function differences:

$$train_{ij} = 1 \quad \text{iff} \quad (V_{i1r_0t_0} - V_{ijr_0t_0}, ..., V_{iJr_0t_0} - V_{ijr_0t_0}) \leq \mathbf{0}, \tag{20}$$

$$occ_{i(k|j)rt} = 1 \quad \text{iff} \quad (U_{i(1|j)rt} - U_{i(k|j)rt}, ..., U_{i(K|j)rt} - U_{i(k|j)rt}) \leq \mathbf{0}, \tag{21}$$

where $\mathbf{0}$ is a (J/K)-dimensional vector. Lee (1983) shows that the selection rules may be reframed in terms of maximum order statistics:

$$train_{ij} = 1 \quad \text{iff} \quad \max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq 0, \tag{22}$$

$$occ_{i(k|j)rt} = 1 \quad \text{iff} \quad \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq 0. \tag{23}$$

Note that rewriting equations (20) and (21) in this way does not impose any assumptions on the underlying selection rules. Define the joint cumulative distribution of the outcome and selection error terms as $F_{jk}(...)$, and the joint cumulative distribution of the outcome error terms and the two maximum order statistics as $G_{jk}(...)$. Denote the corresponding probability density functions by $f_{jk}(...)$ and $g_{jk}(...)$. Note that, in line with the control function defined in equation (19), I allow the distributions to vary by jk-cell. Evaluating $F_{jk}(...)$ at the observed value function and utility differences and using Lee's (1983) insight on maximum order statistics, the following equality holds:

$$\begin{aligned} &F_{jk}(z_0, \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt}) \\ &= G_{jk}(z_0, 0, 0|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}). \end{aligned} \tag{24}$$

The equivalence of the above distribution functions may also be written in terms of density functions $f_{jk}(...)$ and $g_{jk}(...)$. Moreover, there exists a one-to-one mapping between selection probabilities, and utility and value function differences, so that the joint distribution $g_{jk}(...)$ may be conditioned on the vector of selection probabilities instead of the observed utility and value function differences (see Appendix D.1 for details). 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)). Applying a similar result to the present multiple-index framework, model (iv) may be written as

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \tau_{jk} + \lambda_{jk}(p_{i1r_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) + u_{ijkrt}, \tag{25}$$

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_{ijkrt} is a mean zero error term.

Note that, given the sequential nature of the selection problem, the control functions $\lambda_{jk}(...)$ depend only on those occupation probabilities $p_{i(k|j)rt}$ that condition on the observed training choice j. The sequential nature of the selection problem therefore reduces the dimensionality of the control function. Nonetheless, estimating this equation non-parametrically would require a flexible function in $(J + K)$ probabilities to be included in $(J \times K)$ different control functions which will be infeasible in the present high-dimensional context. I follow Lee (1983) and Dahl (2002) and impose a distributional assumption to further reduce the dimensionality of the problem.

Dahl (2002) imposes the following index sufficiency assumption which states that the distribution $g_{jk}(...)$ depends on the set of selection probabilities only through the probabilities of selection into the observed jk-cell:

$$\begin{aligned}
&g_{jk}(\epsilon_{jkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\qquad p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt}) \\
&= g_{jk}(\epsilon_{ijkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\qquad p_{ijr_0t_0}, p_{i(k|j)rt}). \qquad \text{(A1)}
\end{aligned}$$

Intuitively, assumption (A1) states that all information about the joint distribution $g_{jk}(...)$ is summarised by the indices $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$. This assumption may be relaxed by assuming that $g_{jk}(...)$ also depends on a (small) set of selection probabilities such as the second or third best alternative. These probabilites are however not observed.[35] In the present sequential context, a natural extension would be to also include the probability of working on the diagonal, $p_{i(k=j|j)rt}$. I show that my results are robust to relaxing assumption (A1) and also allowing $g_{jk}(...)$ to depend on $p_{i(k=j|j)rt}$ (see Appendix F.2).

Implicit in Lee's (1983) parametric approach is a stronger assumption which states that the joint distribution of outcome errors and maximum order statistics $g_{jk}(...)$ does not depend on the vector of selection probabilities. Dahl (2002) makes this assumption explicit:

$$\begin{aligned}
&g_{jk}(\epsilon_{jkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\qquad p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt}) \\
&\text{does } \textit{not} \text{ depend on } (p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt}). \qquad \text{(A2)}
\end{aligned}$$

In the following, I will use assumptions (A1) and (A2) to estimate the parameters of interest τ. As further discussed in Sections 5.2 and 5.3, since jk-specific intercepts are only identified at infinity, their estimation relies on assumption (A2) and requires a specific distributional assumption on $g_{jk}(...)$. On the other hand, the estimation of slope parameters may proceed using a non-parametric approach under the weaker assumption (A1).[36]

I use model (ii) to explain both estimation approaches as the parameters of interest τ^{exp} contain both intercepts (the experience-invariant components) and slopes (the change in τ^{exp} over the experience schedule). Since both the non-parametric and the parametric approach may be used to estimate the slope parameters, comparing the estimated experience slopes under both approaches will allow me to assess the distributional assumption required for the estimation of the experience-invariant intercepts (see Sections 5.3 and 6.2).

[35] In his application on the selection of individuals into US states, Dahl (2002) also includes the probability of staying in the state of birth in the estimation.

[36] Slope parameters are effects of variables that vary conditional on selection choice and probabilities.

5.2 Non-parametric Control Function

Under the index sufficiency assumption (A1), model (ii) may be written as

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \tau^{exp} D_{j=k} + \lambda_{jk}(p_{ijr_0t_0}, p_{i(k|j)rt}) + u_{ijkrt}, \tag{26}$$

where $E[u_{ijkrt}|M_{ijkrt} = 1] = 0$.

Dahl (2002) provides a proof of this result which I adapt to the present context and present in Appendix D.2. Under the instrumental variable assumptions presented in Section 4.4, the difference in parameters τ^{exp} for any two experience levels, i.e. the slope of the effect by experience, will be identified non-parametrically given consistent estimators for the probabilities $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$. On the other hand, the estimation of the jk-specific intercepts contained in τ^{exp} requires a stronger distributional assumption which will be discussed in Section 5.3 below.

In order to implement the non-parametric approach, the control functions $\lambda_{jk}(...)$ may be approximated using a flexible polynomial in the selection probabilities $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$, where parameters on all terms in the polynomial are allowed to vary by jk-cell. In Section 5.5, I show how I derive the estimates for the selection probabilities, $\hat{p}_{ijr_0t_0}$ and $\hat{p}_{i(k|j)rt}$. With these at hand, I approximate $\lambda_{jk}(...)$ with a second-order polynomial in the estimated probabilities $\hat{p}_{ijr_0t_0}$ and $\hat{p}_{i(k|j)rt}$, where parameters are allowed to vary for a selected number of jk-cells. The selected jk-cells include all jk-cells where $j = k$, and an additional cell for each occupation k where $j \neq k$.

5.3 Parametric Control Function

While Dahl's (2002) flexible non-parametric approach can be used to estimate the slope parameters, it is problematic for any jk-specific intercepts since identifying these separately from the intercepts in the control function relies on strong support requirements on the instruments.[37] 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 outcome and selection errors. Intuitively, this method extrapolates the selection probabilities by imposing a functional form assumption in order to separately identify the parameters

[37] In particular, semi-parametric identification of average treatment effects (ATEs) in Roy models is often based on "identification at infinity" (IAI) arguments (Chamberlain (1986), Heckman (1990)). Intuitively, the IAI approach relies on the instruments 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 consistenly estimates the ATE.

of interest from the intercepts of the control function. However, parameterising the distribution of the error terms in high-dimensional settings may make estimation infeasible. Lee (1983) addresses this problem by constructing new random variables from the maximum order statistics that greatly simplify the estimation. The details of his approach in the present setting are described in Appendix D.3. His basic idea is that even though the joint distribution of maximum order statistics may not be identically distributed, it is possible to construct a new random variable from the maximum order statistics that is identically distributed. Lee (1983) then assumes that the joint distribution of outcome errors and the newly constructed random variable does not vary with the observed utility or value function differences. Dahl (2002) shows that this assumption is equivalent to assumption (A2).

The final step involves making parametric assumptions on the distributions. I follow Lee (1983) and impose standard normality assumptions for the relevant distributions. The control function is then given by the well-known function of the inverse Mill's ratio (Heckman (1976, 1979)):

$$\lambda_{jk}(p_{i1r_0t_0}, ...p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) = -\rho_{jk}\frac{\phi[\Phi^{-1}(p_{ijkrt})]}{p_{ijkrt}}, \tag{27}$$

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 newly constructed random variable, 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.[38] Log wages in model (ii) may then be written as

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \tau^{exp}D_{j=k} - \rho_{jk}\frac{\phi[\Phi^{-1}(p_{ijkrt})]}{p_{ijkrt}} + u_{ijkrt}, \tag{28}$$

where $E[u_{ijkrt}|M_{ijkrt} = 1] = 0$.

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. In a seminal contribution, Vytlacil (2002) establishes the equivalence between the standard Local Average Treatment (LATE) model first proposed by

[38] See Section 5.2 for the selection of jk-cells.

Imbens & Angrist (1994), and a non-parametric latent threshold crossing model. Based on this result, 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. However, the same does not necessarily hold for the estimation of ATEs which may be more sensitive to the choice of distributional assumptions. To justify the distributional assumptions required for the identification of intercept parameters in the present setting, I use the fact that the slope parameters in model (ii) may be estimated using a non-parametric control function estimator. The idea is to compare the estimated slope parameters from the parametric approach described in Section 5.3 to those obtained using the non-parametric approach from Section 5.2. The more similar the parametric estimates are to the non-parametric ones, the more likely the distrubtional assumptions are to hold in practice. Figure 6 in Section 6.2 provides the results for this comparison. It can be seen that the slope estimates from both approaches coincide almost exactly, lending support to the distributional assumptions made.

5.4 Splitting Vacancies into Expectation and Shock

In order to obtain the training and occupation instruments defined in equations (17) and (18), 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:[39]

$$ln(vac_{krt}) = \kappa_{kr} + \pi_{krt} \times t + \varepsilon_{krt}, \quad \forall rt. \tag{29}$$

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 (29):

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

[39] 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.

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:

$$vac_{krt} - E_{t_0}[ln(vac_{kr(t_0+\tau)})|\Omega_{r_0t_0}], \quad \forall\tau = 0, ..., 30. \tag{31}$$

While the conditional expectations derived using equation (30) will serve as training instruments IV_{train_j}, the residuals from equation (31) will serve as occupation instruments IV_{occ_k}. Note that, using this definition, expectations and shocks will be orthogonal by construction.

5.5 Estimating the Selection Probabilities

Implementing the control function approach requires consistent estimators for the selection probabilities $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$. Using the definition of the instrumental variables in equations (17) and (18), the selection probabilities may be written as

$$\begin{aligned} p_{ijr_0t_0} &= Pr(train_{ij} = 1|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}) \\ &= Pr(train_{ij} = 1|IV_{train_j}, E_{t_0}[vac_{jr(t_0+\tau)}|\Omega_{r_0t_0}], X_{ijkrt}),\ \forall\tau = 0, ..., 30, \end{aligned} \tag{32}$$

$$\begin{aligned} p_{i(k|j)rt} &= Pr(occ_{i(k|j)rt} = 1|\tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\ &= Pr(occ_{i(k|j)rt} = 1|train_{ij} = 1, IV_{train_j}, IV_{occ_k}, vac_{krt}, X_{ijkrt}), \end{aligned} \tag{33}$$

where X_{ijkrt} is a set of controls including gender and full-time work experience. Writing the probabilites in this way makes clear that, as shown in Figure 4, the training choice will only depend on the first set of instruments and the occupation choice will depend on both sets of instruments as well as the previous training choice.

The most common approach to estimating choice probabilities in high-dimensional settings is the conditional logit model (McFadden (1974)). While this method provides convenient closed-form expressions for the choice probabilities, it requires specific functional form assumptions. This can be particularly problematic in the presence of a large number of independent variables. In the present context, there are over 400 independent variables to predict both training and occupation choices. Parameterising choices using a sufficiently flexible functional form in these variables would quickly make the estimation infeasible.

I therefore employ an alternative approach and estimate selection probabilities non-parametrically using a machine learning algorithm, random forests. This algorithm uses the large set of explanatory variables described in equations (32) and (33) to predict the choice variables $train_{ij}$ and $occ_{i(k|j)rt}$ by using optimal splitting rules on the explanatory variables. Details on the random forest algorithm as well as the implementation of the algorithm in the present context can be found in Appendix D.4. With the estimated probabilities at hand, the control function estimation proceeds by replacing the selection probabilties from Sections 5.2 and 5.3 with their estimates $\hat{p}_{ijr_0t_0}$ and $\hat{p}_{i(k|j)rt}$.

6 Results

This section discusses the results for models (i) to (iv) presented in Section 3.2. In all baseline estimations, $f(vac_{krt})$ will be approximated using a fourth order polynomial in vac_{krt}. Robustness using a tenth 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 model (i), where parameter τ captures the average return to working on versus off the diagonal. Table 4 shows the regression results. The main variable of interest is the dummy $D_{j=k}$ 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, exp_k, columns (3) and (4) add exp_k and its square as control variables. Both specifications are first estimated without controlling for selection (columns (1) and (3)), then using the control function estimator (columns (2) and (4)).[40]

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)), a result which is in line with workers on the diagonal having more occupation-specific experience. When accounting for selection using the parametric control function estimator, the effect of $D_{j=k}$ becomes *positive* and significant (columns (2) and (4)), implying that not contolling for selection into training and occupations leads to a sizeable *negative* selection bias of around 10 percentage points. The coefficients on the parametric control function are highly significant in all regressions, confirming the importance of the selection bias. Intuitively, workers working on the diagonal may be negatively selected relative to workers off the diagonal as the latter must compensate for the lack of training with increased occupation-specific ability (see Section 4.2).

Results from column (2) suggest that working on the diagonal leads to a significant wage increase of about 12%. This figure may be interpreted as the full effect of having received training in the current occupation, including potentially higher experience in that occupation

[40] Note that, since occupation-specific experience corresponds to past selection into occupations, an additional instrument is required and the control function estimator needs to be adapted for the results in column (4) (see Appendix E.2 and E.3 for details). Empirically, also instrumenting for exp_k as in Table 4 column (4) gives a very similar result compared to treating exp_k as exogenous.

which was accumulated as a result of the training.[41] As before, controlling for occupation-specific experience lowers the effect of $D_{j=k}$ to about 10% (column (4)). Albeit smaller than the full effect of 12%, 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.

In the heterogeneity and full-matrix analysis that follows, the full-effect specification corresponding to columns (1) and (2) will be used as the default, i.e. regressions will control for total work experience exp but not for occupation-specific work experience exp_k. Appendix F.1.1 explicitly considers heterogeneity by occupation-specific work experience.

Table 4: Average On- versus Off-Diagonal Returns

	(1)	(2)	(3)	(4)
$D_{j=k} = 1$	-0.0076	0.1226^{***}	-0.0281^{***}	0.1006^{***}
	(0.0061)	(0.0311)	(0.0070)	(0.0278)
exp	0.0602^{***}	0.0600^{***}	0.0442^{***}	0.0437^{***}
	(0.0022)	(0.0026)	(0.0026)	(0.0046)
exp^2	-0.0010^{***}	-0.0010^{***}	-0.0004^{***}	-0.0004^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
exp_k			0.0176^{***}	0.0176^{***}
			(0.0015)	(0.0049)
exp_k^2			-0.0008^{***}	-0.0007^{***}
			(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	1,143,782	1,143,782	1,146,854	1,146,854

Notes: The table reports regression results for model (i). Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Standard errors (in parentheses) are clustered at the region and time level. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

[41] Alternatively, one can think of the causal effect of "going back in time" and re-choosing the training.

6.2 Heterogeneity by Experience

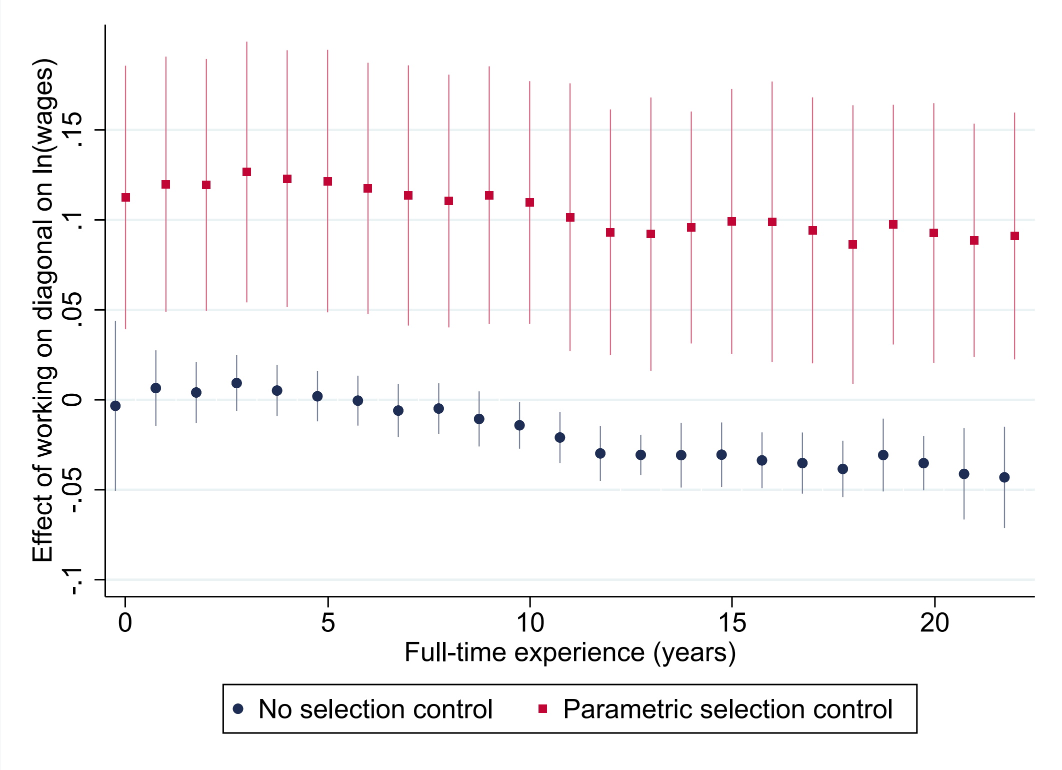
This section reports and discusses the results for model (ii), which looks at the heterogeneity in on- versus off-diagonal returns across different levels of full-time work experience. Figure 5 plots separate coefficient estimates for τ^{exp}, where experience levels have been binned into yearly categories. Each coefficient compares workers with a specific level of full-time work experience who were trained in the occupation they are working in to workers with the same level of experience who were *not* trained in the occupation they are working in. Similar to columns (1) and (2) from Table 4, the effect shown therefore represents the full effect of having received training, including any occupation-specific work experience accumulated in the current occupation as a result of having received the training. Results by occupation-specific experience are presented in Appendix F.1.1.

Figure 5 plots coefficient estimates from an estimation without selection control, and using the parametric control function estimator. In line with the results from Table 4, it can be seen that not controlling for selection leads to a sizeable negative bias in the coefficient estimates. The set of coefficients estimated using the parametric control function suggests that the effect of $D_{j=k}$ first increases slightly from around 11% to 12.5%, then falls to just under 10% after 12 years of work experience where it stabilises. While the increase over the first few years 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 are able to partly catch up with their co-workers who received the relevant training. However, there is no full catch-up and sizeable differences remain after 20 years of work experience.

Figure 6 plots the same set of coefficients estimated with the parametric control function estimator as Figure 5, adding the coefficients estimated without selection control, and those estimated using the non-parametric control function estimator. Since the non-parametric control function approach identifies the slope but not the level of the parameters of interest (see Sections 5.3 and 5.2), all coefficients are normalised to zero at zero years of work experience. Comparing the set of coefficients estimated without selection and with the parametric control function estimator, it can be seen 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.

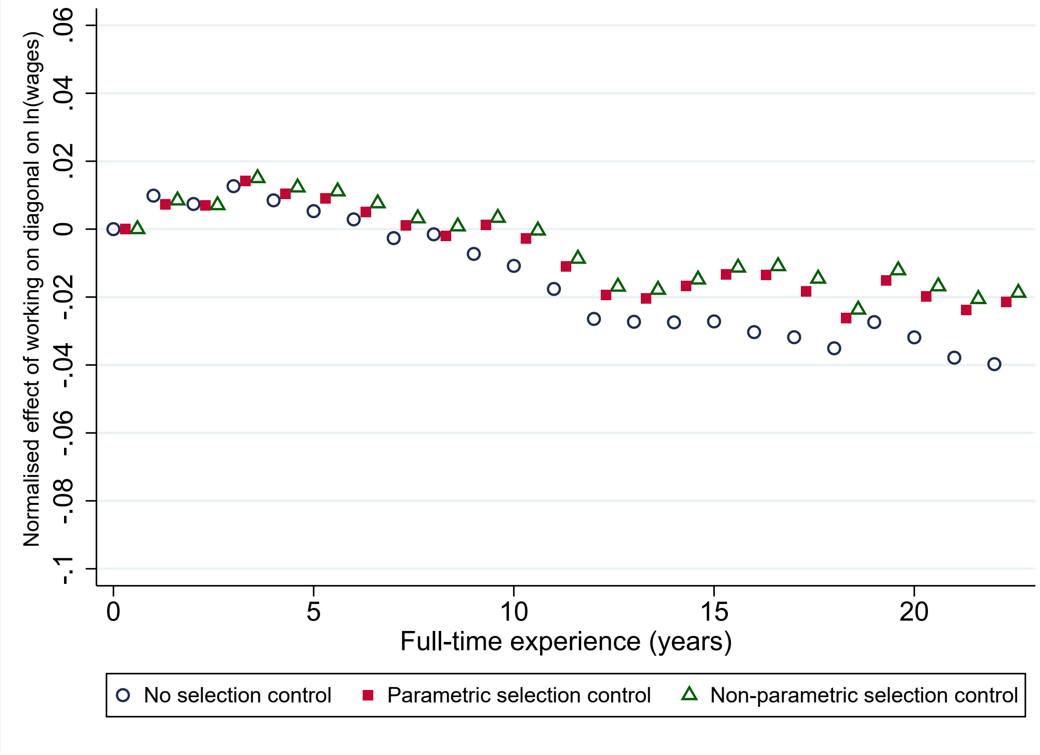
Moreover, comparing the set of coefficients estimated with the parametric and the non-parametric control function estimator shows that the slope estimated using the non-parametric selection control is almost identical to that of the parametric selection control. This lends support to the distributional assumptions made to implement the parametric approach (see Section 5.3). As a robustness check, Appendix F.2 reports a similar comparison, adding on-diagonal probability terms to the non-parametric control function (see Section 5.1). The results provide further support to the distributional assumptions made.

Figure 5: On- versus Off-Diagonal Returns by Experience



Notes: The figure plots regression coefficient estimates for τ^{exp} in model (ii), where experience levels have been binned into yearly categories. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Standard errors are clustered at the region and time level. 95% confidence intervals are shown.

Figure 6: Normalised On- versus Off-Diagonal Returns by Experience



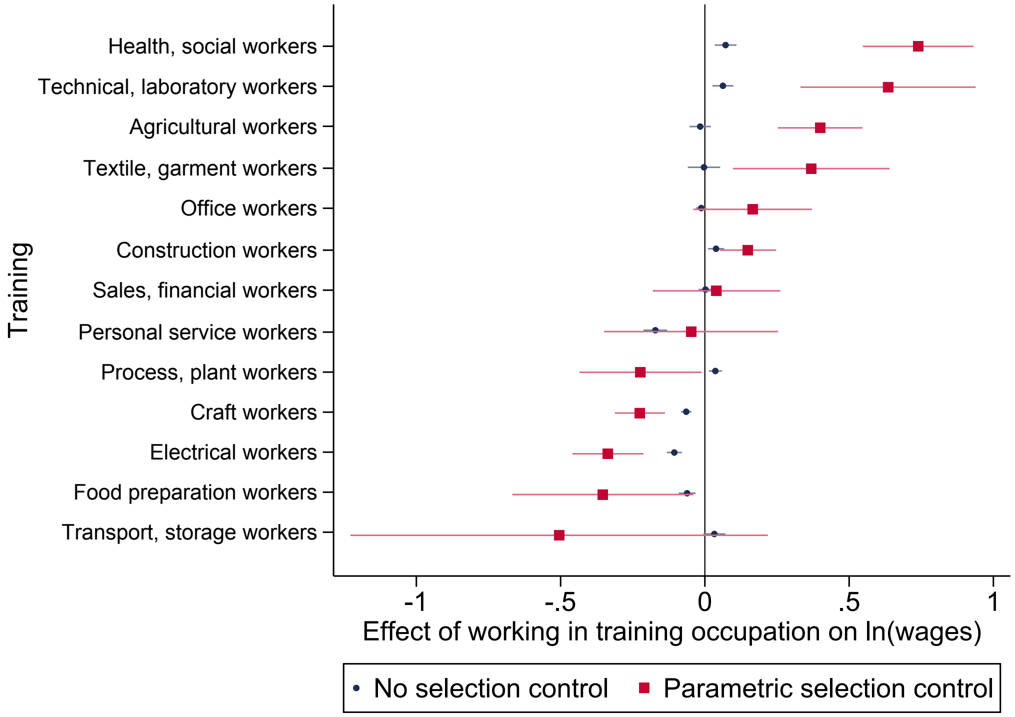
Notes: The figure plots regression coefficient estimates for τ^{exp} in model (ii), where experience levels have been binned into yearly categories. All coefficient estimates are normalised to zero at zero years of work experience. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010.

6.3 Heterogeneity by Training

This section presents and discusses the results for model (iii) which explores the heterogeneity in on- versus off-diagonal returns across trainings. Figure 7 plots the coefficients τ_j estimated with and without the parametric control function estimator. There appears to be an inverse relationship between both sets of coefficients which will be further discussed below. The coefficient estimates are highly heterogeneous and, out of the five largest trainings, *health and social workers* have the highest and *craft workers* the lowest return to working in their training occupation. Note that model (iii) does not contain occupation fixed effects. As a result, negative coefficient estimates $\hat{\tau}_j$ may be explained 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

Figure 7: Average On- versus Off-Diagonal Returns by Training

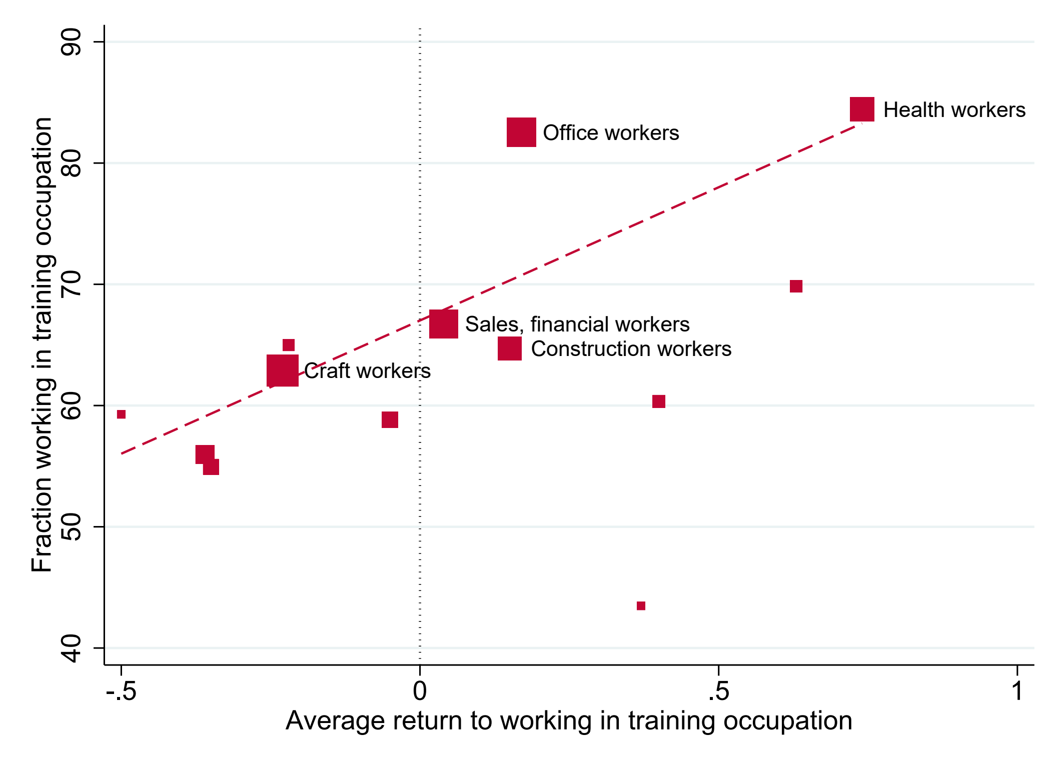


Notes: The figure shows regression coefficient estimates for τ_j in model (iii) for each training. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Standard errors are clustered at the region and time level. 95% confidence intervals shown.

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 the on- versus off-diagonal return in that training. Figure 8 explores this relationship by plotting the average return to working *on* the diagonal (from Figure 7) and the fraction working *on* the diagonal for each training. The positive slope is consistent with the Roy model predictions outlined above.[42]

[42] Note that the model is also consistent with alternative explanations for the positive correlation in Figure 8 which are based on the *ex-ante* selection into training rather than the *ex-post* occupation choice conditional on training. For instance, conditional on the expected average returns to a training, higher returns to working on the diagonal may lead to lower expected *value* of that training if individuals are risk averse. As a result, workers choosing that training may be more positively selected. This will in turn lead to fewer individuals with that training choosing to work *off* the diagonal ex-post. Note that, in contrast to the direct effect of the returns on ex-post occupation choice outlined in the main body of the paper, alternative explanations based on the selection into training rely more heavily on the exact specification of preferences in the model.

Figure 8: Average Return and Fraction Working in Training Occupation



Notes: The figure plots average on- versus off-diagonal returns for each training from Figure 7 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.

A further implication of the Roy model is that heterogeneity in the average return to working on versus off the diagonal across trainings affects the size of the selection bias. As outlined in Section 6.1, the strong negative bias in average returns suggests that *on*-diagonal workers are negatively selected relative to *off*-diagonal workers, as the latter need to compensate for their lack of training through 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.[43] Figure F.2 in Appendix F.1.2 confirms this by showing a negative correlation between the estimated return and the selection bias across trainings.

[43] Note that selection into training may instead lead to a positive correlation between the on-diagonal return and the estimated selection bias across trainings. Conditional on expected average returns across trainings, higher on-diagonal returns may lead to lower expected *value* of that training if individuals are risk averse. As a result, workers choosing that training are more positively selected, i.e. they have higher ability in their training relative to other trainings. This may lead to returns being more highly *overestimated*.

6.4 Full Training-Occupation Matrix

This section reports and discusses the results for model (iv) which contains separate parameters τ_{jk} for all cells in the training-occupation matrix. Table 5 reports the coefficient estimates $\hat{\tau}_{jk}$ using the parametric control function estimator. For expositional clarity, results are shown for the five largest occupations only. The full tables of coefficients, estimated with and without selection control, can be found in Appendix F.1.3. As outlined above, the inclusion of individual fixed effects in the estimation implies that all coefficients should be interpreted relative to the diagonal within the same training. Coefficients relative to the diagonal in the same occupation are presented in Appendix F.1.5.

In line with the *positive* on- versus off-diagonal returns for four out of five of the largest trainings (see Figure 7), Table 5 shows that, with the exception of training as a *craft worker*, most coefficients are significant and *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* who work in other occupations, with estimates ranging between -0.93 and -0.43 log points.[44] Table 5 also shows that returns are highly asymmetric. While trained *office workers* incur sizeable penalities 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 generally worse opportunities (see Section 6.3). Similar to Figure 8, Figure F.3 in Appendix F.1.3 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 8, the positive correlation in the full training-occupation matrix confirms the importance of returns in determining the selection into occupations.

While hard to interpret individually, the estimates from the full training-occupation matrix provide an opportunity to study the mechanisms underlying the results presented in this paper. In Section 7, I use data on the task content of trainings and occupations to explore both the heterogeneity and asymmetry in estimated returns, thereby providing a microfoundation for the estimates presented in this study.

[44] Note this is in line with the higher on- versus off-diagonal return estimated for *health and social workers* in Section 6.3.

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	-0.37^{*}	0.01	-0.30^{**}	0.13
			(0.19)	(0.08)	(0.13)	(0.25)
	Craft workers	0.20^{**}	0	0.39^{***}	0.05	0.43^{**}
		(0.07)		(0.06)	(0.12)	(0.15)
	Sales, financ. w.	-0.04	-0.10	0	-0.13	0.36^{**}
		(0.10)	(0.13)		(0.17)	(0.15)
	Health, social w.	-0.77^{***}	-0.93^{***}	-0.49^{***}	0	-0.43^{*}
		(0.10)	(0.16)	(0.09)		(0.23)
	Construction w.	-0.15	-0.24^{**}	0.08	-0.27^{*}	0
		(0.10)	(0.10)	(0.08)	(0.13)	

Notes: The table shows regression coefficient estimates for τ_{jk} in model (iv), estimated using the parametric control function estimator. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Results are shown for the five largest occupations. The full matrix of coefficients is presented in Appendix F.1.3. 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.[45] Column (1) excludes employment spells where workers change the region in which their employer is located (see Section 4.4.1); 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 have been 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. As in column (2) Table 4, results are obtained using the parametric control function estimator.

Table 6 shows that the effect of working on versus off the diagonal is significantly positive in all columns, with most results being quantitatively very similar to the main sample estimate of 12.3%. Columns (1), (2) and (3) all report coefficient estimates of around 12%. While column (1) alleviates potential concerns regarding the conditional random assignment

[45] Further robustness results focusing on the estimation method can be found in Appendix F.2.

assumption (see Section 4.4.1), columns (2) and (3) suggest that differences in the length across apprenticeships or the presense of institutional wage caps in the data are not driving the main result. The point estimate is slightly higher at 14.4% in column (4), and slightly lower at 7.2% in column (5). While the result in column (5) is partly driven by the fact that the main coefficient is lower at higher levels of experience (see Figure 5 Section 6.2), and that spells in apprenticeship firms are concentrated early in a worker's career, these results also point to potential complementarities of working both in the *occupation* and the *firm* one has been trained for. However, column (5) shows that the coefficient remains sizeable even when controlling for such complementarities by excluding spells in the firm workers got trained in.

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
$D_{j=k} = 1$	0.1216^{***}	0.1147^{**}	0.1213^{***}	0.1436^{***}	0.0718^{**}
	(0.0300)	(0.0383)	(0.0297)	(0.0326)	(0.0266)
exp	0.0601^{***}	0.0570^{***}	0.0605^{***}	0.0601^{***}	0.0573^{***}
	(0.0028)	(0.0033)	(0.0025)	(0.0028)	(0.0026)
exp^2	-0.0010^{***}	-0.0010^{***}	-0.0011^{***}	-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	1,121,091	281,571	1,121,626	1,027,817	827,396

Notes: The table reports regression results for model (i). Each column restricts the baseline sample as indicated in the column header, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Standard errors (in parentheses) are clustered at the region and time level. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

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. While these tasks are defined as units of work activity, skills refer to the human capital required to carry out the tasks (Autor (2013)). Occupations, as discrete classification units, can thus be viewed as vectors of tasks to be carried out by workers.[46]

Measures of task content have been used to analyse 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 (2017)). Altonji *et al.* (2014) use task content measures to study changes in earnings inequality across college majors. A different strand of the literature uses the concept of occupations as task vectors to construct measures of *distance* between occupations. Poletaev & Robinson (2008) and Gathmann & Schönberg (2010) argue that, if human capital is task-specific, it should be more easily transferable across occupations that require a similar mix of tasks. Using samples of displaced workers, they find that wage penalties are larger the more distant the occupational switch after displacement.[47]

When applied to the present context, these findings suggest a potential 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 might 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 Data

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 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 three survey waves that fall within the time period used for the estimation of returns in

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

[47] Yamaguchi (2012) sets up a structural model to formalise 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.

the training-occupation matrix (1985/86, 1991/92, 1998/99).[48]

The survey records information about workers' occupations and asks them to pick from a list of tasks the ones they perform in their current occupation. I follow Gathmann & Schönberg (2010) and combine all survey waves to create a list of 19 tasks. A summary table of the tasks together with the percentage of individuals working in the two largest occupations (*office workers* and *craft workers*) who indicated that they perform these tasks is presented in Figure 7 below.[49]

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.[50] 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).

7.2 Measuring Task Distance

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:

$$AngSim_{jk} = \frac{\sum_{s=1}^{S}(q_{sj} \times q_{sk})}{[(\sum_{s=1}^{S} q_{sj}^2) \times (\sum_{s=1}^{S} q_{sk}^2)]^{1/2}}. \tag{34}$$

Measuring similarity between two vectors by the angular separation has first been proposed by Jaffe (1986, 1989a) in the context of estimating R&D spillovers across technologically similar firms.[51] The angular separation is equivalent to the uncentered correlation or the

[48] All three surveys were conducted in collaboration with the German Institute for Employment Research (IAB). Survey wave 2006 was carried out jointly with the German Federal Institute for Occupational Safety and Health (*Bundesanstalt für Arbeitsschutz und Arbeitsmedizin - BAuA*) and is not used since, unlike the other three waves, occupations were coded using an updated classification system.

[49] To construct averages, sample observations within each wave are weighted using provided survey weights and subsequently combined giving equal weight to each survey wave.

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

[51] Subsequently, a number of other studies have used the measure in a variety of different 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)).

Table 7: List of Tasks and Fraction Performing

Task	Office workers	Craft workers
Cultivate	0%	1%
Serve or accommodate	2%	0%
Clean	2%	4%
Manufacture, install or construct	3%	41%
Secure	3%	3%
Publish, present or entertain others	5%	0%
Repair, renovate, reconstruct	6%	68%
Equip or operate machines	11%	61%
Nurse or treat others	16%	9%
Pack, ship or transport	17%	20%
Execute laws or interpret rules	25%	3%
Design, plan, sketch	28%	19%
Research, evaluate or measure	32%	35%
Program	33%	2%
Teach or train others	36%	25%
Employ, manage personnel, organise, coordinate	38%	14%
Calculate or do bookkeeping	41%	6%
Sell, buy or advertise	43%	17%
Correct texts or data	74%	9%

Notes: The table shows the average fraction of individuals indicating they perform the given task. Fractions are based on survey waves 1985/86, 1991/92 and 1998/99.

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.[52] $AngSim_{jk}$ 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:

$$Dist_{jk} = (1 - AngSim_{jk}). \tag{35}$$

Excluding on-diagonal training-occupation cells where $Dist_{jk} = 0$, the distance measure varies between 0.01 and 0.59 with a mean of 0.34. When weighting off-diagonal cells by their sample fractions, the mean distance drops to 0.28 suggesting that, on average, workers who leave their training occupation work in occupations which are more similar to their

[52] In contrast to that, the Euclidean distance between two vectors q_j and q_k measures the length of the vector connecting q_j and q_k and is therefore sensitive to the length of q_j and q_k. As a result, two occupations with relatively short vector lengths could be classified as similiar even when they are orthogonal.

training than the average occupation. Figure G.1 in Appendix G.1 confirms this by showing a negative correlation between training-occupation distance and the fraction of workers in the relevant occupation. Tables G.1 and G.2 in Appendix G.1 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.

7.3 A Simple Model of Match Returns and Tasks

I model the returns to a match between a training and an occupation combining three elements from the task approach literature. (1) Tasks are inputs to production and therefore occupation attributes. (2) Skills refer to the human capital that is required to carry out tasks and are therefore worker attributes. (3) Wage penalties after displacement tend to be larger the more distant the new occupation is from the previous one. While the first two elements are conceptual issues emphasising the importance of a clear distinction between tasks and skills (Autor (2013), Autor & Handel (2013)), element three relates to an empirical finding that has emerged in the task content literature (see e.g. Gathmann & Schönberg (2010))

I combine elements (1) to (3) to model the estimated returns τ_{jk} from model (iv) in Section 3.2 as functions of skills that workers with training j have at performing tasks in occupation k (elements (1) and (2)). These skills are in turn a function of the task distance between j and k (element (3)). Returns τ_{jk} may therefore be written as

$$\begin{aligned} \tau_{jk} &= f[skill_{jk}] + \eta_{jk} \\ &= f[g(Dist_{jk}, X_{jk})] + \eta_{jk}, \end{aligned} \tag{36}$$

where $skill_{jk}$ are the skills workers with training j have at performing tasks in occupation k, $Dist_{jk}$ is the distance between j and k as defined in Sections 7.2, X_{jk} is a control variable measuring the direction of a training-occupation move (see Section 7.4 for details), and η_{jk} is a match-specific error term. Assuming a linear specification and replacing off-diagonal returns τ_{jk} with their estimates from Section 6.4, $\hat{\tau}_{jk}$ may be written as

$$\hat{\tau}_{jk} = \alpha_j + \gamma_1 Dist_{jk} + \beta' X_{jk} + \eta_{jk}. \tag{37}$$

Note that, since returns $\hat{\tau}_{jk}$ are relative to the diagonal in the same training, the model contains training-specific fixed effects. If skills are more easily transferable the closer the tasks in occupation k are to those acquired in training j, one would expect a negative coefficient on $Dist_{jk}$.

7.4 Results

This section provides results on the relationship between returns and task distance based on the simple model presented in Section 7.3. To this end, I estimate equation (37) using the estimated on- versus off-diagonal returns $\hat{\tau}_{jk}$ from Section 6.4 as dependent variable. Table 8 reports the regression results where on-diagonal ($\hat{\tau}_{jk} = 0$) observations have been excluded.[53]

Column (1) shows that the average effect of task distance on returns is negative and significant, suggesting that higher task distance between occupation and original training leads to lower returns relative to working in one's training occupation. $Dist_{jk}$ has been standardised to have mean zero and standard deviation equal to one. The results therefore suggest that a one-standard-deviation increase in task distance is associated with a reduction in the return to the match of around $4pp.$, or around 50% of the average $\hat{\tau}_{jk}$.

Recent empirical findings by Robinson (2018) suggest that the effect of task distance may depend on the direction of an occupational move as measured by the difference in the overall skill levels required by the new and the old occupation. I test for such an effect by constructing a dummy variable, $Down_{jk}$, which is equal to one if the estimated on-diagonal return in occupation k is lower than that of training j (see Appendix F.1.5 for results on these returns). Since on-diagonal task distance is zero by construction, on-diagonal returns may be used as proxy for the overall level of required skills. Columns (2) and (3) report the regression results. As expected, the coefficient on $Down_{jk}$ is negative and significant. Moreover, in line with the results by Robinson (2018), column (3) shows that after allowing for differential effects of distances, the effect of $Dist_{jk}$ is entirely driven by *downward* moves. While difficult to motivate theoretically, this empirical heterogeneity in the effect of distance can help explain the important asymmetries in estimated returns in the training-occupation matrix.

When estimating the same regressions as in Table 8 with left-hand-side returns τ_{jk} estimated *without* selection control, the effect of $Dist_{jk}$ is smaller in magnitude and only marginally significant (see Appendix G.2, Table G.3). The above findings are robust to including the on-diagonal observations where $Dist_{jk} = 0$ (see Appendix G.3, Table G.4), and restricting the sample to the five largest trainings and occupations (see Appendix G.3, Table G.5). Overall, the results are in line with the proposed hypothesis that apprentices are trained in a specific mix of tasks and their skill at performing tasks in a different occupation is lower, the less applicable the acquired skills are to the current occupation. Moreover, the findings show that this mechanism primarily works through training-occupation matches where the occupation requires *lower* overall skill levels than the original training.

[53] Appendix G.3 provides a robustness result including these observations.

Table 8: Match Returns and Task Distance

	(1)	(2)	(3)
$Dist_{jk}$	-0.0394^{**}	-0.0389^{***}	0.0216
	(0.0149)	(0.0106)	(0.0278)
$Dist_{jk} \times Down_{jk}$			-0.1275^{**}
			(0.0503)
$Down_{jk}$		-0.1140^{**}	-0.1053^{**}
		(0.0456)	(0.0462)
Mean of $\hat{\tau}_{jk}$	-0.0786	-0.0786	-0.0786
Train. FE	yes	yes	yes
Adj. R^2	0.6078	0.6151	0.6246
N	156	156	156

Notes: The table reports regression results from equation (37). $Dist_{jk}$ is constructed using survey waves 1985/86, 1991/92 and 1998/99, and scaled by its standard deviation. Diagonal coefficients (where $\hat{\tau}_{jk} = 0$ and $Dist_{jk} = 0$) have *not* been included in the regression. Standard errors are clustered at the training level. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

8 Welfare and Policy

The results from Section 6.1 suggest large average positive returns to working on versus off the diagonal. This section explores the welfare losses from ex-post suboptimal training choices which are implied by these estimates. In line with the proposed model, the key underlying friction is the ex-ante imperfect information at the time of training choice. As a result, new information about the labour market and individual occupation-specific abilities may be revealed over time, causing workers to seek work in an occupation different from the one they got trained for. In addition to these off-diagonal workers, I argue that a second group of workers is affected by the friction. These are workers who are locked into their training despite having higher occupation-specific ability in a different occupation, since those skills are not sufficient to compensate for the lack of training.

Section 8.1 quantifies the welfare loss associated with the friction for both off-diagonal and locked-in workers. Section 8.2 considers a specific potential policy intervention, retraining programmes.[54] Back-of-the-envelope calculations suggest that such programmes could be very effective in addressing the friction in the present context.

8.1 Welfare Losses

This section looks at the welfare losses associated with the informational frictions at the time of training choice. All calculations are based on the simple model of homogenous on- versus off-diagonal returns, model (i). For both off-diagonal and locked-in workers, I compute losses as the product of the loss per worker and the share of affected workers.

Focusing first on off-diagonal workers, it is possible to show based on a revealed preference argument that the loss from the friction is equal to the on-diagonal benefit τ.[55] My findings suggest that τ is 10% (see column (4) in Table 4 in Section 6.1).[56] Figure 2 in Section 2.5 shows that the average share of off-diagonal workers is about 40%. The welfare loss from off-diagonal workers is therefore given by 4% of wages for an average worker in the system.

[54] Importantly, retraining programmes target ex-post wage outcomes and do not require any assumptions on preferences beyond those given in Section 3. Policies targeting welfare *conditional* on wage outcomes such as insurance against a "lost training investment" would necessitate strong assumptions on the distribution of preferences and beliefs at the time of training choice.

[55] An underlying assumption is that non-monetary within occupation does not depend on the training received. Note also that this estimate will likely be a lower bound on the gains from retraining since retraining in the current occupation may not be the first best outcome when taking into account (1) differences in δ_j, or (2) heterogeneous returns across the training-occupation matrix (model (iv)). Moreover, given the distribution of off-diagonal workers across occupations, the average change in estimated on-diagonal returns across trainings, δ_j, resulting from retraining workers in their current occupation is small and positive. This implies that the welfare gains estimated in this section would be even larger.

[56] Note that the result controlling for occupation-specific experience should be used here as a hypothetical retraining scenario would *not* lead to higher accumulated work experience in the newly chosen occupation.

Consider now the workers who are locked into their training. These are defined as individuals who are working on the diagonal, but who would choose a different occupation in the absence of any on- versus off-diagonal return. Again using a revealed preference argument, the welfare loss per locked-in worker is bounded from above by the on- versus off-diagonal return τ. The intuition behind this is that locked-in workers do not choose to move given the on- versus off-diagonal return.[57] 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 the fraction of on-diagonal workers over a career. Table 9 summarises the calculations. Figure 5 in Section 6.2 shows that the return to working on versus off the diagonal falls by about 2.5*pp* between $3-12$ years of work experience.[58] At the same time, the fraction of on-diagonal workers falls from about 70% to 60% (see Figure 2 in Section 2.5).[59] 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 once returns have stabilised, i.e. after 12 years of work experience. Figure 2 shows that between $12-21$ years of work experience, this fraction dropped by about 5*pp*. 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 1*pp* reduction in the return to working on the diagonal leads to a 2*pp* drop in the fraction of individuals working on the diagonal.[60] In a hypothetical world without a 10% return on the diagonal, a world without lock-in effects, the fraction of individuals working on the diagonal would thus be 20*pp* lower. Combining the estimated share of 20% with the upper bound on losses per locked-in worker, these results suggest that the welfare loss from locked-in workers is given by at most 2% of wages for an average worker in the apprenticeship system.

[57] An underlying assumption is that non-monetary utility within occupation is independent of the training.

[58] The same is true for returns by *occupation-specific* work experience (see Appendix F.1.1).

[59] This 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.

[60] 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.

Table 9: Estimating the Share of Locked-in Workers

work exp.	Δ return	Δ fraction		implied Δ fraction	$\frac{\Delta \text{ fraction on diag.}}{\Delta \text{ return on diag.}}$
(years)	on diag.	on diag.		if Δ return = 0	
$3-12$	$-2.5pp$	$-10pp$	$\}$	$-5pp$	$\frac{-5pp}{-2.5pp}=2$
$12-21$	0	$-5pp$		$-$	$-$

Notes: The table summarises the estimation of the share of locked-in workers (see text for details).

Combining the results for off-diagonal and locked-in workers, the total loss is given by

$$\begin{aligned} L_{total} &= L_{off-diagonal} + L_{locked-in} \\ &= (loss_{off-diagonal} \times share_{off-diagonal}) + (loss_{locked-in} \times share_{locked-in}) \\ &\leq (\tau \times share_{off-diagonal}) + (\tau \times share_{locked-in}), \\ L_{total} &\approx 4-6\%. \end{aligned} \tag{38}$$

8.2 Retraining Programmes

The calculations in Section 8.1 show that the welfare losses from the informational frictions at the time of training choice are large. This section considers retraining programmes as a potential policy intervention to recover these losses.[61] In light of the results from Section 7, such programmes may be viewed as moving workers' skills at performing certain tasks closer to those required in their current occupation. Since workers are trained in special subjects relating to their training occupation for two thirds of their training, I assume that retraining programmes would last two thirds of the original training time.

Retraining programmes will be costly to both the government and training firms. In addition, workers will face private costs in the form of foregone earnings while retraining. In 2010, total costs to train an apprentice, including training and schooling costs as well as foregone earnings amounted to 29,460 Euros.[62] As outlined in Section 8.1, the per worker yearly welfare gain from retraining is τ.[63] My empirical results suggest that this implies net benefits including foregone work experience of around $1,800$ Euros per year in 2010.[64]

Note that, while the costs need to be paid at the time of retraining, the benefits will

[61] 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 H.2 for a discussion.

[62] See Appendix H.1 for details on the calculations.

[63] This is equivalent to the per worker yearly welfare loss due to the friction.

[64] See Appendix H.1 for details on the calculations.

subsequently accrue for every year spent working after retraining. Whether or not retraining has a net benefit therefore depends on the career stage of a worker. Using a discount factor of 0.99, retraining costs would be recovered for workers with at most six years of work experience. Since the majority of off-diagonal workers leave their training occupation in the first few years after completing the apprenticeship (see Figure 2 in Section 2.5) and only switch occupations once (see Table 1 in Section 2.5), these findings suggest that retraining could pass a cost-benefit test for 35% of all workers, or over three quarters of workers ever working off the diagonal. Moreover, 18% of all workers are still locked in after six years (over 90% of workers ever locked in). These workers would equally benefit from retraining.[65]

Given the above figures, an imminent question is why, in practice, few individuals choose to retrain.[66] In theory, firms are not allowed to discriminate against older applicants and trained workers could apply for training in a different occupation. Moreover, a number of firms offer special training programmes for career-switchers.[67] Anecdotal evidence from internet forums and newspaper articles suggests that the barriers to retraining fall into two broad categories: liquidity constraints and other (perceived) costs such as the "fear of starting over again".[68] In light of this evidence, facilitating entry into retraining programmes by providing easily accessible loans, or reducing the (perceived) barriers of entry would lead to substantial welfare gains.

9 Conclusion

This paper combines a large employment panel dataset with data on historical occupation-specific vacancies to identify and estimate the returns to different training-occupation combinations. To this end, I extend previous work on control function 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 behavioural justification for the identification strategy, and implement the estimation approach by setting up a generalised two-stage Roy (1951) model where individuals seek relative advantage when choosing their training

[65] See Appendix H.1 for details on the calculations.

[66] The sample fraction of individuals enrolled in two apprenticeships in distinct occupations is 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.

[67] See e.g. https://www.faz.net/aktuell/beruf-chance/ue-40-ausbildung-neubeginn-in-der-lebensmitte-11968937.html for a newspaper article on this phenomenon.

[68] The former may be particularly important given retraining becomes relevant when many workers start a family. Regarding the latter, workers repeatedly state that they fear starting over again, being isolated amongst their peers due to their age or struggling to keep up with the study load. Moreover, firms offering programmes for carrer-switchers often ask applicants to pass a set of tests confirming their motivation and ability to retrain in a different field.

and occupation. To the best of my knowledge, this work is the first to present results on the returns to different training-occupation combinations which are well identified.

The results suggest significant returns of about $10 - 12\%$ to working in the occupation one has been trained for, with considerable heterogeneity across trainings and occupations. The estimated selection bias is sizeable and negative, suggesting that occupation-specific ability needs to compensate for the lack of training in the chosen occupation.

I find considerable heterogeneity in returns across the full training-occupation matrix. Combining these 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 in that cell. These findings provide an explanation for the heterogeneity in 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 the imperfect information available at the time of training choice leads to important welfare losses of around $4 - 6\%$ of wages for the average worker. These losses are economically meaningful and may seem surprising in the German apprenticeship system which has been repeatedly termed a role model for other economies in Europe, the US, China and India.[69] My findings show that there are frictions within this system that should be addressed by policy makers, and I discuss a specific policy instrument, ex-post retraining programmes. Back-of-the-envelope calculations suggest that such programmes are likely to generate sizeable net welfare gains. Througout my policy analysis, I took the existing training system as given, and looked at improvements in the allocation of workers to training choices. Future work could consider optimal training programmes by building on my estimates of the effects of task distance on the returns to training-occupation combinations.

[69] See e.g. https://www.ft.com/content/1a82e8e0-04cf-11e7-aa5b-6bb07f5c8e12 and https://www.bbc.co.uk/news/business-16159943.

References

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. 2019. *Ex Ante* Returns and Occupational Choice. *Unpublished manuscript.*

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

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.

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.

Dustmann, Christian, & Meghir, Costas. 2005. Wages, Experience and Seniority. *The Review of Economic Studies*, **72**(1), 77–108.

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. forthcoming. Multidimensional Skill Mismatch. *American Economic Journal: Macroeconomics.*

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.

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.

Imbens, Guido W., & Wooldridge, Jeffrey M. 2007. Control Functions and Related Methods. *NBER Summer Institute Methods Lectures.*

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.

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.

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. 2018. On Worker and Firm Heterogeneity in Wages and Employment Mobility: Evidence form Danish Register Data. *Unpublished manuscript.*

McFadden, Daniel L. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. *Chap. 4, pages 105–142 of:* Zarembka, Paul (ed), *Frontiers in Econometrics.* Academic Press, New York.

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. 2016. Selective Migration, Occupation Choice, and the Wage Returns to College Major. *Unpublished manuscript.*

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.

Soskice, David. 1994. Reconciling Markets and Institutions: The German Apprenticeship System. *Chap. 1, pages 25–60 of:* Lynch, Lisa M. (ed), *Training and the Private Sector: International Comparisons.* University of Chicago Press.

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, Lukas. 2017. Within Occupation Wage Dispersion and the Task Content of Jobs. *GRAPE Working Paper 22.*

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. Descriptives

A.1 List of Occupations

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 labourers	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
79-89	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%.

A.2 Distribution of Spells

Table A.2: Spells as Percentage of Trainings

		Occupation												
		01-09	10-18, 52-54	19-30, 32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	51.9	6.4	4.9	0.8	0.2	0.4	4.8	3.3	5.3	9.6	6.5	4.6	1.4
	10-18, 52-54	0.7	57.4	4.2	0.6	0.1	0.2	1.5	12.0	4.7	5.7	8.6	3.6	0.6
	19-30, 32	0.9	9.5	55.2	1.6	0.2	0.4	2.5	8.9	3.9	9.2	4.8	2.4	0.6
	31	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	0.5	9.6	8.0	0.5	35.5	1.5	3.6	7.1	8.3	5.6	12.7	4.8	2.5
	39-43	1.1	8.6	6.2	0.7	0.3	43.2	3.6	1.7	7.8	13.2	6.7	3.6	3.4
	44-51	1.1	7.6	5.7	0.5	0.3	0.4	60.2	4.3	3.1	9.4	3.5	2.9	0.9
	60-63	0.3	2.7	2.5	3.2	0.0	0.1	0.7	68.6	4.1	1.6	12.8	2.8	0.5
	68-70	0.2	2.3	1.6	0.2	1.2	0.6	0.3	0.8	60.5	3.4	26.5	2.1	1.1
	71-74	0.1	5.3	3.5	0.9	0.0	0.3	2.5	2.1	7.6	55.4	18.8	2.8	0.7
	75-78	0.1	0.8	0.6	0.1	0.0	0.0	0.1	1.1	12.5	2.0	80.6	1.6	0.4
	79-89	0.1	0.8	0.7	0.1	0.0	0.2	0.2	0.8	4.3	1.0	12.2	79.1	0.7
	90-93	0.4	5.2	3.6	0.5	0.3	3.0	0.6	1.0	10.6	3.8	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* occupation for the baseline sample. Results are restricted to individuals with ten years of work experience.

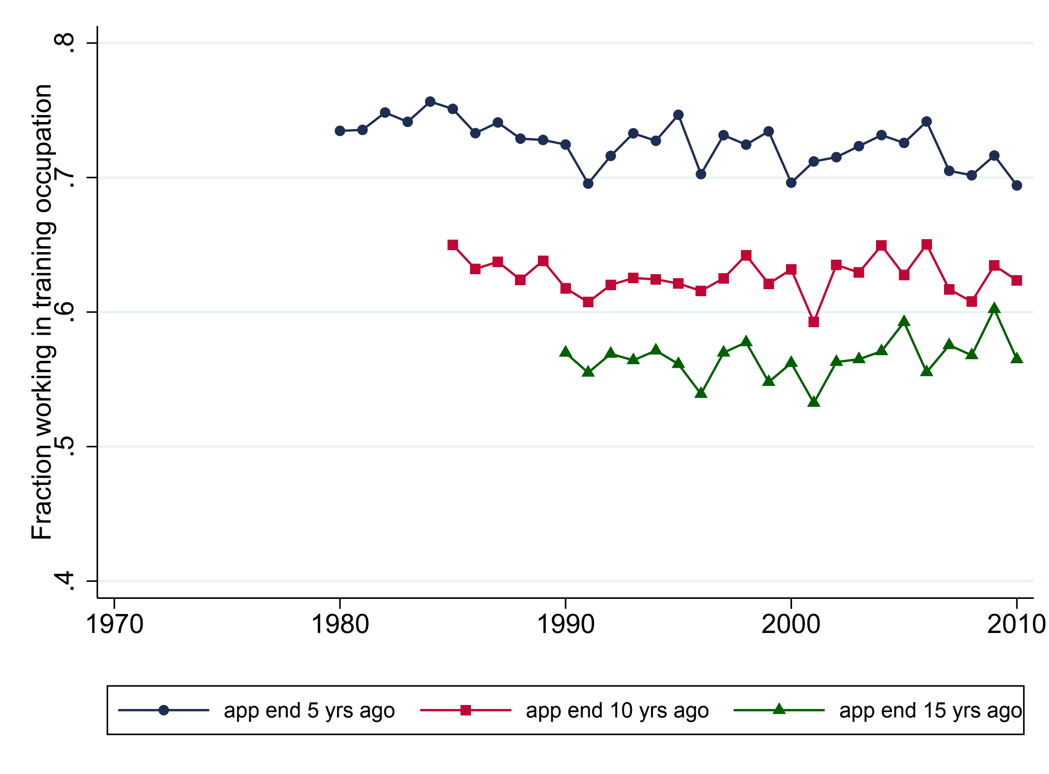
Table A.3: Spells as Percentage of Occupations

		Occupation												
		01-09	10-18, 52-54	19-30, 32	31	33-37	39-43	44-51	60-63	68-70	71-74	75-78	79-89	90-93
Training	01-09	70.7	2.6	0.8	0.5	0.9	0.4	1.5	1.2	0.9	3.9	0.7	1.2	1.5
	10-18, 52-54	0.8	19.2	0.6	0.3	0.3	0.2	0.4	3.8	0.7	1.9	0.8	0.8	0.6
	19-30, 32	11.4	36.8	84.2	8.9	8.8	3.6	7.6	32.3	6.2	36.3	5.0	5.7	6.7
	31	2.6	6.4	4.1	84.2	0.9	0.7	1.2	19.3	1.9	5.7	2.6	2.1	1.6
	33-37	0.2	1.5	0.5	0.1	70.0	0.6	0.4	1.1	0.5	0.9	0.5	0.5	1.1
	39-43	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	7.0	13.6	4.0	1.2	6.3	2.1	85.1	7.2	2.3	17.1	1.7	3.2	4.3
	60-63	0.4	1.2	0.4	2.1	0.2	0.1	0.2	27.6	0.7	0.7	1.5	0.7	0.6
	68-70	1.8	5.9	1.7	0.9	6.2	4.4	0.6	1.9	64.8	9.1	18.2	3.4	8.0
	71-74	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	1.0	2.3	0.7	0.3	1.0	0.3	0.3	2.8	14.4	5.5	59.5	2.8	3.1
	79-89	0.8	1.2	0.4	0.3	0.6	0.7	0.3	1.1	2.7	1.5	5.0	76.3	2.7
	90-93	0.8	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.

A.3 Fraction On Diagonal Over Time

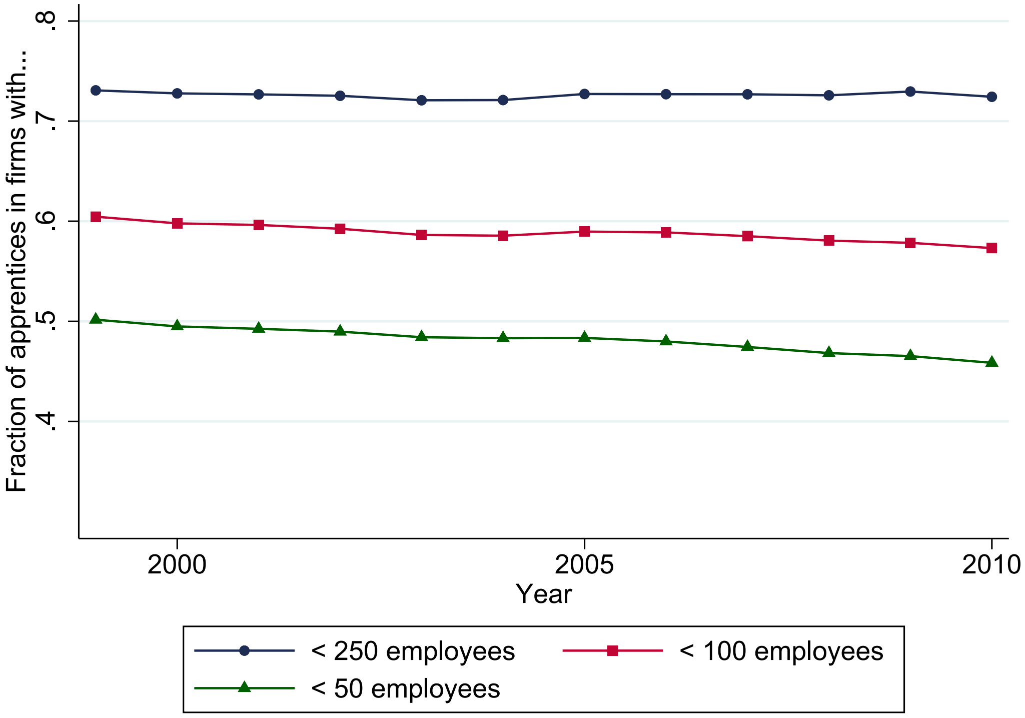
Figure A.1: Fraction On Diagonal over Time



Notes: The figure plots the fraction of individuals working in an occupation equal to their training occupation over time for the baseline sample. Each line plots this fraction for individuals who finished their apprenticeship either 5, 10, or 15 years prior to the date shown on the x-axis. Occupations are classified using the 13 category baseline classification.

A.4 Training Firm Statistics

Figure A.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*.

Appendix B. Proof

B.1 Proof for Section 4.2 :

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

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

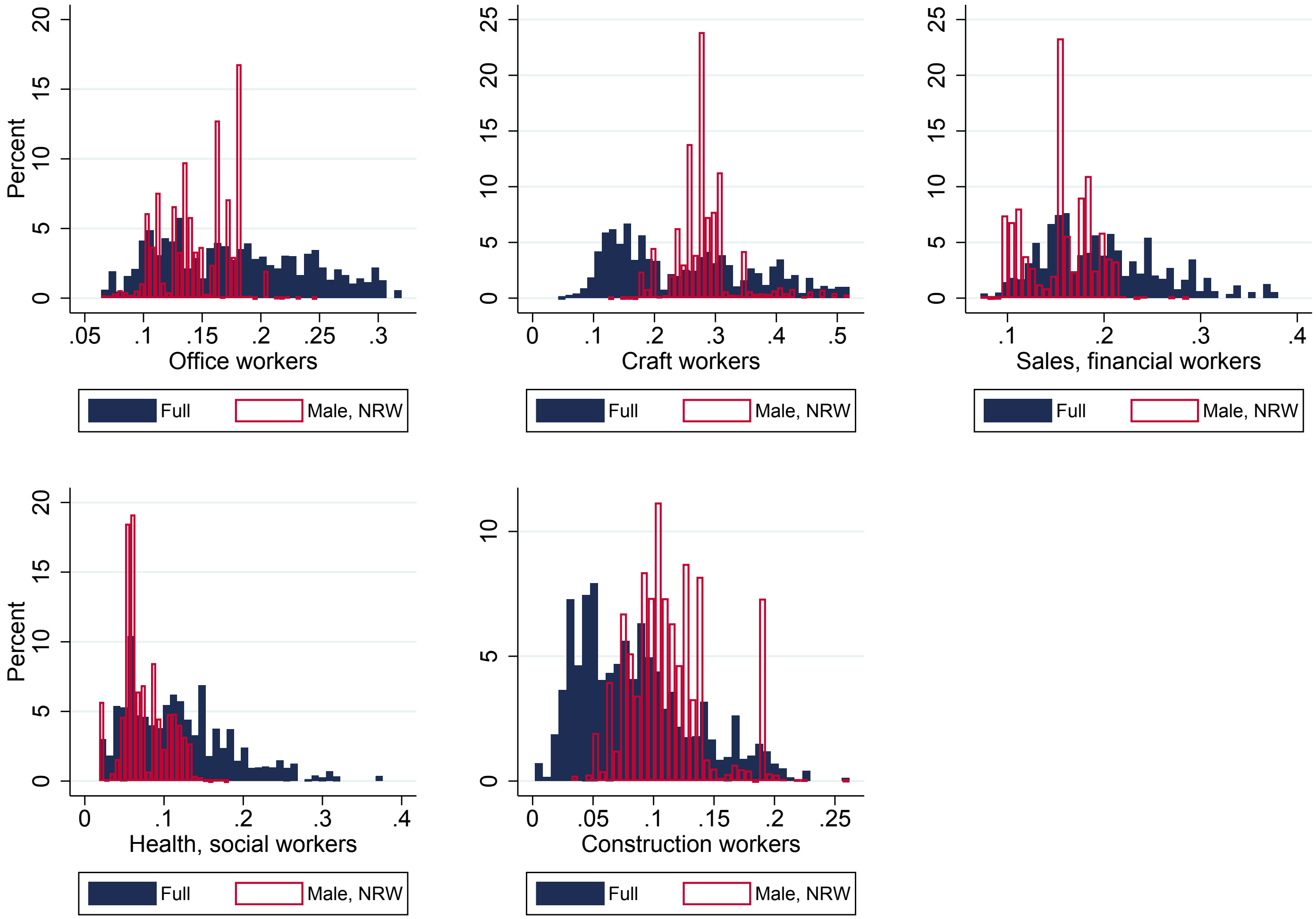
It follows that $E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2)})>0]-E[\epsilon_{i1}|(\epsilon_{i1}-\epsilon_{i2})<0]>0.$

Now 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 C. Identification

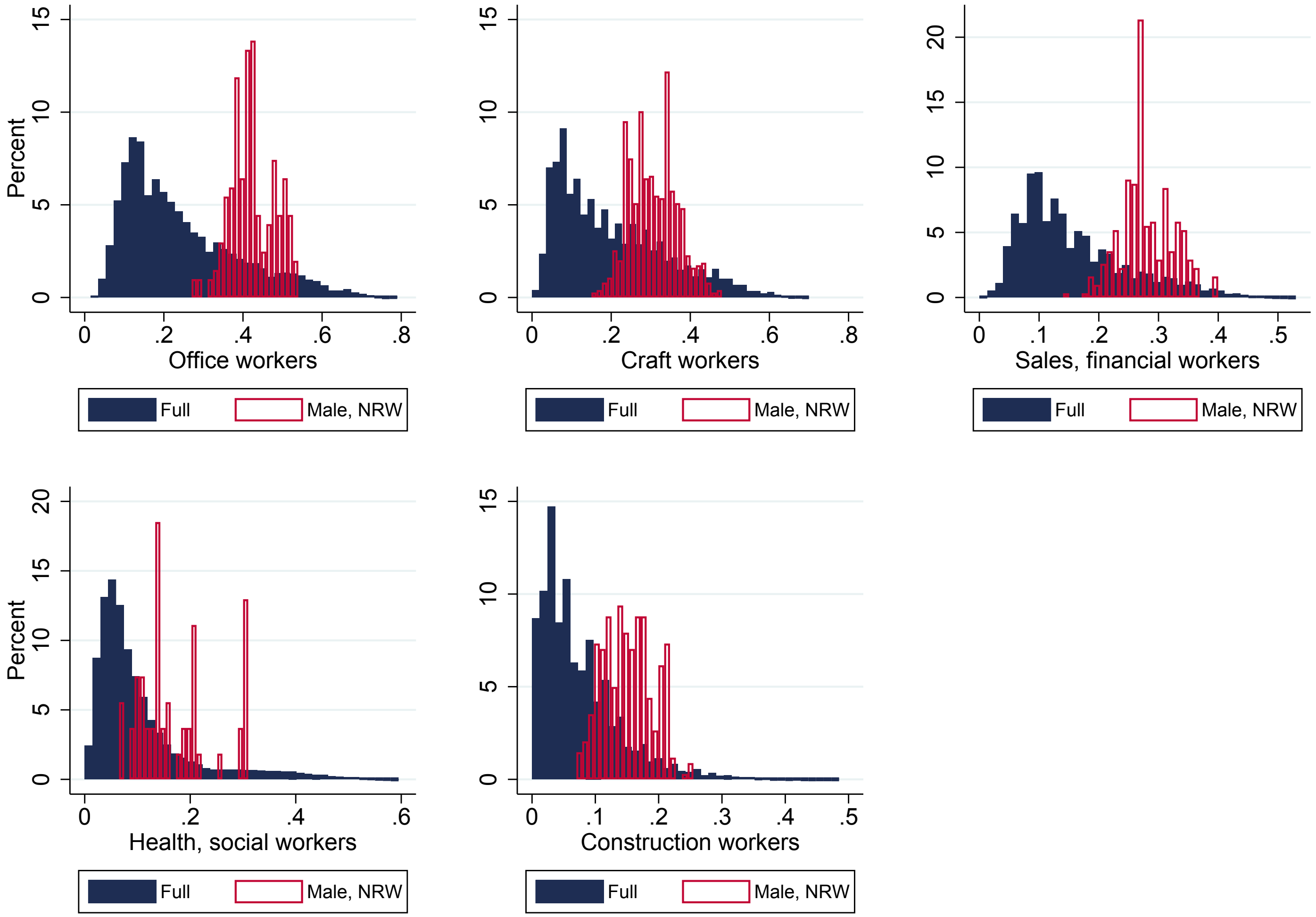
C.1 First Stage

Figure C.1: First Stage Variation in Selection Probabilities - Training



Notes: The figure shows a set of histograms of the selection probabilities for the five largest trainings. Histograms in blue show the full variability in estimated selection probabilities. To give a sense of the variation used in final log wage regressions, histograms in red restrict the sample to males in North Rhine-Westphalia (NRW), the largest state in Germany.

Figure C.2: First Stage Variation in Selection Probabilities - Occupation



Notes: The figure shows a set of histograms of the selection probabilities for the five largest occupations. Histograms in blue show the full variability in estimated selection probabilities. To give a sense of the variation used in final log wage regressions, histograms in red restrict the sample to males in North Rhine-Westphalia (NRW), the largest state in Germany, with ten years of work experience who were *trained* in the respective occupation.

C.2 Training Fixed Effects

As outlined in Section 3.2, workers in the sample completed exactly one apprenticeship and the inclusion of individual fixed effects therefore absorbs any training fixed effects. As a result, one coefficient in each training row (chosen to be the diagonal) is not identified, and the coefficients τ_{jk} in model (iv) from Section 3.2 correspond to returns relative to the diagonal within the same training. They may only be interpreted relative to the diagonal in the same *occupation* if there are no differences in the on-diagonal returns across trainings, i.e. the training fixed effects are zero. While the duration of apprenticeships is relatively homogenous in the German context, there may be systematic quality differences across trainings which could lead to sizeable differences in these training fixed effects (e.g. Soskice (1994)).

The aim of this section is to separately identify the on-diagonal returns across different trainings from permanent individual heterogeneity. The challenge for such an exercise is the systematic self-selection of individuals into different trainings. While the control function estimator should account for these selection effects, and models (i) to (iv) could have therefore been estimated without individual fixed effects, the approach taken in this thesis is a more conservative one where the estimated individual fixed effects $\hat{\delta}_i$ are used to distinguish between training effects δ_j, and permanent individual heterogeneity α_i ex-post.

Assume that the individual fixed effects δ_i in models (i) to (iv) are additively separable in an individual permant component α_i, and a training effect δ_j which captures the quality across different trainings j:

$$\delta_i = \delta_j + \alpha_i. \tag{C.2.1}$$

If individuals self-select into trainings, a simple regression of $\hat{\delta}_i$ on a set of training dummies will not identify the training effects δ_j, since $E[\alpha_i | train_{ij} = 1] \neq 0$.

Instead, the approach taken here is to average the estimated individual fixed effects for a set of 322 districts, and use the arguably exogenous variation in shares of individuals with different trainings *across* these districts to identify the training effects δ_j. Using equation (C.2.1), the average of δ_i in district d is given by

$$\begin{aligned}
\bar{\delta}_i^d &= s_{j=1}^d \delta_{j=1} + ... + s_{j=13}^d \delta_{j=13} + \bar{\alpha}_i^d \\
&= (1 - s_{j=2}^d - s_{j=3}^d - ...)\delta_{j=1} + ... + s_{j=13}^d \delta_{j=13} + \bar{\alpha}_i^d \\
&= \delta_{j=1} + s_{j=2}^d(\delta_{j=2} - \delta_{j=1}) + ... + s_{j=13}^d(\delta_{j=13} - \delta_{j=1}) + \bar{\alpha}_i^d,
\end{aligned} \tag{C.2.2}$$

where s_j^d denotes the share of individuals trained in j in district d.

Based on equation (C.2.2), I run the following regression:

$$\hat{\bar{\delta}}_i^d = \delta_r + \delta_1 + \delta_2 s^d_{j=2} + ... + \delta_{13} s^d_{j=13} + \beta' X_d + \bar{\alpha}^d_i + \epsilon^d, \tag{C.2.3}$$

where $\hat{\bar{\delta}}_i^d$ is the empirical counterpart of $\bar{\delta}_i^d$, X_d includes district-level control variables for population and population density, δ_r is a fixed effect for the region district d belongs to, $\bar{\alpha}^d_i$ is the average of individual fixed effects in district d, and ϵ^d is a mean zero error term. Results from the estimation of equation (C.2.3) are presented in Appendix F.1.4.

The key assumption for this regression to identify the training effects δ_j is that differences in shares *across* districts are exogenous and therefore $E[\bar{\alpha}^d_i | s^d_j] = 0$. This is equivalent to assuming that average ability levels are the same across districts, but structural differences generate exogenous variation in s^d_j.

Appendix D. Estimation Details

D.1 Reduction of Dimensionality

As in the Section 5.1, 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 $G_{jk}(...)$ in equation (24) can be established with the following steps:

$$\begin{aligned}
&F_{jk}(z_0, \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt}) \\
&= Pr(\epsilon_{ijkrt} \leq z_0, e_{i1r_0t_0} - e_{iJr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\qquad e_{ij1rt} - e_{ijkrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., e_{ijKrt} - e_{ijkrt} \leq \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= Pr(\epsilon_{ijkrt} \leq z_0, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}) \leq 0, \\
&\qquad \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt}) \leq 0| \\
&\quad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= G_{jk}(z_0, 0, 0|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.1.1}$$

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

$$\begin{aligned}
&f_{jk}(\epsilon_{ijkrt}, e_{i1r_0t_0} - e_{ijr_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0}, e_{ij1rt} - e_{ijkrt}, ..., e_{ij1rt} - e_{ijKrt}| \\
&\quad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&= g_{jk}(\epsilon_{jkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\quad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.1.2}$$

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:

$$\begin{aligned}
&= g_{jk}(\epsilon_{jkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\qquad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}). \\
&= g_{jk}(\epsilon_{jkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt})| \\
&\qquad p_{i1r_0t_0}, ..., p_{ijr_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(k|j)rt}, ..., p_{i(K|j)rt}). \qquad \text{(D.1.3)}
\end{aligned}$$

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.

D.2 Non-Parametric Control Function

Proof for Section 5.2 : $\lambda_{jk}(p_{i1r_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) = \lambda_{jk}(p_{ijr_0t_0}, p_{i(k|j)rt})$.

For notational convenience, denote $\vec{V} = (\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0})$ and $\vec{U} = (\tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt})$. Using the definition of $g_{jk}(...)$ as joint distribution of the outcome error and maximum order statistics, the control function may be written as

$$\begin{aligned}
&\lambda_{jk}(p_{i1r_0t_0}, ..., p_{iJr_0t_0}, p_{i(1|j)rt}, ..., p_{i(K|j)rt}) \\
&\quad = E[\epsilon_{ijkrt}|M_{ijkrt} = 1] \\
&\quad = \int_{-\infty}^{\infty} t_1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_{jk}(t_1, t_2, t_3 | M_{ijkrt} = 1, \vec{V}, \vec{U}) dt_2 dt_3 dt_1 \\
&\quad = \frac{\int_{-\infty}^{\infty} t_1 \int_{-\infty}^{0} \int_{-\infty}^{0} g_{jk}(t_1, t_2, t_3 | \vec{V}, \vec{U}) dt_2 dt_3 dt_1}{Pr(M_{ijkrt} = 1 | \vec{V}, \vec{U})},
\end{aligned}$$

where the final equality follows from Bayes' Theorem, and the denominator $Pr(M_{ijkrt}|\vec{V}, \vec{U}) = p_{ijr_0t_0} \times p_{i(k|j)rt}$. Under the index sufficiency assumption (A1),

$$\begin{aligned}
&g_{jk}(\epsilon_{ijkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt}|\vec{V}, \vec{U}) \\
&\quad = g_{jk}(\epsilon_{ijkrt}, \max_{j'}(\tilde{V}_{ij'r_0t_0} - \tilde{V}_{ijr_0t_0} + e_{ij'r_0t_0} - e_{ijr_0t_0}), \max_{k'}(\tilde{U}_{i(k'|j)rt} - \tilde{U}_{i(k|j)rt} + e_{ijk'rt} - e_{ijkrt}| \\
&\qquad p_{ijr_0t_0}, p_{i(k|j)rt}).
\end{aligned}$$

As a result, the control function $\lambda_{jk}(...)$ only depends on probabilities $p_{ijr_0t_0}$ and $p_{i(k|j)rt}$.

D.3 Parametric Control Function

Recall that the selection problem is given by

$$M_{ijkrt} = 1 \quad \text{iff} \quad \max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq 0 \;\text{ and }\; \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq 0. \tag{D.3.1}$$

Lee (1983) points out that it is possible to create new random variables based on the distribution of the maximum order statistics.[70] 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

$$\begin{aligned}
&L_{jk}(\tilde{V}_{ijrt_0} - \tilde{V}_{i1rt_0} + z_1, ..., \tilde{V}_{ijrt_0} - \tilde{V}_{iJrt_0} + z_1, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt} + z_2, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt} + z_2) \\
&\quad = Pr(e_{i1r_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijrt_0} - \tilde{V}_{i1r_0t_0} + z_1, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijrt_0} - \tilde{V}_{iJrt_0} + z_1, \\
&\qquad\quad e_{ij1rt} - e_{ijkrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt} + z_2, ..., e_{ijKrt} - e_{ijkrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt} + z_2) \\
&\quad = Pr(\max_{j'}(V_{ij'r_0t_0} - V_{ijr_0t_0}) \leq z_1, \max_{k'}(U_{i(k'|j)rt} - U_{i(k|j)rt}) \leq z_2| \\
&\qquad\quad \tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}) \\
&\quad = H_{jk}(z_1, z_2|\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt}).
\end{aligned} \tag{D.3.2}$$

Now define the random variables ζ_{ijkrt} as

$$\begin{aligned}
\zeta_{ijkrt} = \Gamma_{jk}^{-1}\{H_{jk}(0, 0|&\tilde{V}_{i1r_0t_0} - \tilde{V}_{ijr_0t_0}, ..., \tilde{V}_{iJr_0t_0} - \tilde{V}_{ijr_0t_0}, \\
&\tilde{U}_{i(1|j)rt} - \tilde{U}_{i(k|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(k|j)rt})\},
\end{aligned} \tag{D.3.3}$$

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

$$\begin{aligned}
M_{ijkrt} = 1 \quad \text{iff} \quad \zeta_{ijkrt} \leq \Gamma_{jk}^{-1}\{L_{jk}(&\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
&\tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt})\},
\end{aligned} \tag{D.3.4}$$

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 $(\epsilon_{ijkrt}, \zeta_{ijkrt})$ is independent and identically distributed with joint cumulative distribution function $G_{jk}(...)$,

[70] See Appendix A in Dahl (2002) for details.

thereby specifying the joint distribution of outcome and selection errors $F_{jk}(...)$. Importantly, the distribution function $G_{jk}(..)$ 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 assumption (A2). Using this assumption, the joint distrbution of outcome and selection errors $F_{jk}(...)$ may be written as

$$\begin{aligned}
& F_{jk}(z_0, \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(1|j)rt} - \tilde{U}_{i(K|j)rt}) \\
& = Pr(\epsilon_{ijkrt} \leq z_0, e_{i1r_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., e_{iJr_0t_0} - e_{ijr_0t_0} \leq \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
& \qquad\qquad e_{ij1rt} - e_{ijkrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., e_{ijKrt} - e_{ijkrt} \leq \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt}) \\
& = Pr(\epsilon_{ijkrt} \leq z_0, \zeta_{ijkrt} \leq \Gamma_{jk}^{-1}\{L_{jk}(\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
& \qquad\qquad\qquad \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(K|j)rt})\}) \\
& = G_{jk}(z_0, \Gamma_{jk}^{-1}\{L_{jk}(\tilde{V}_{ijr_0t_0} - \tilde{V}_{i1r_0t_0}, ..., \tilde{V}_{ijr_0t_0} - \tilde{V}_{iJr_0t_0}, \\
& \qquad\qquad \tilde{U}_{i(k|j)rt} - \tilde{U}_{i(1|j)rt}, ..., \tilde{U}_{i(K|j)rt} - \tilde{U}_{i(1|j)rt})\}). \qquad \text{(D.3.5)}
\end{aligned}$$

The final step involves making parametric assumptions on the distributions $\Gamma_{jk}(...)$ and $G_{jk}(...)$. As described in Section 5.3, I follow Lee (1983) and assume that $\Gamma_{jk}(...)$ is a univariate standard normal cdf and $G_{jk}(...)$ is a bivariate standard normal cdf.

D.4 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.[71]

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 choice using the explanatory variables described in

[71] For details, see Hastie *et al.* (2009).

Section 5.5.[72] 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 are then computed as the proportion of votes for that option across all trees in the final nodes.

[72] Both random forests are based on 500 trees, where 1000 randomly selected observations from the training dataset are used to grow each tree.

Appendix E. Occupation-Specific Experience

E.1 Dynamic Model

When occupation-specific experience exp_k affects productivity, the occupational choice problem becomes dynamic. Consider the following two specifications for log wages:

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \boldsymbol{\tau} D_{j=k} + \beta exp_k + \epsilon_{ijkrt}, \tag{v}$$

$$ln(w_{ijkrt}) = \delta_r + \delta_t + f(vac_{krt}) + \delta_i + \delta_k + \boldsymbol{\tau^{exp_k}} D_{j=k} + \epsilon_{ijkrt}, \tag{vi}$$

where, for simplicity, other controls have been omitted. As before, $D_{j=k}$ is a dummy variable equal to one if training j is the same as occupation k. Model (v) controls for occupation-specific experience exp_k, model (vi) estimates different effects of working on versus off the diagonal for each occupation-specific experience level. Including exp_k in this way captures the intuition that even if one is not trained in the current occupation, there may be occupation-specific learning while working in an occupation k, which could affect productivity directly or change the wage gap relative to co-workers who received the relevant training.

Assume that that the period-t utility when working in occupation k is given by equation (6) in Section 3.3. Given the wage structure from models (v) and (vi), current period occupation choices affect future utility and the choice problem becomes dynamic. Define the value function of occupation choice k given information Ω_{rt} available at time t, and denote this by $V_{i(k|j)rt}(\Omega_{rt})$. For any occupation choice k, $V_{i(k|j)rt}(\Omega_{rt})$ may be written as

$$V_{i(k|j)rt}(\Omega_{rt}) = \tilde{U}_{i(k|j)rt} + e_{ijkrt} + \beta E_t[V_{i(k|j)r(t+1)}|\Omega_{rt}], \tag{E.1.1}$$

where β denotes a discount factor and $E_t[V_{i(k|j)r(t+1)}|\Omega_{rt}]$ is the maximal expected reward in $t+1$ given today's occupation choice.

Individual i maximises expected utility and chooses occupation k if and only if

$$\begin{aligned}(e_{ijkrt} - e_{ijk'rt}) >&(\tilde{U}_{i(k|j)rt} + \beta E_t[V_{i(k|j)r(t+1)}|\Omega_{rt}]) - (\tilde{U}_{i(k'|j)rt} + \beta E_t[V_{i(k'|j)r(t+1)}|\Omega_{rt}]) \\ &= \tilde{V}_{i(k'|j)rt} - \tilde{V}_{i(k|j)rt}, \quad \forall k' \neq k,\end{aligned} \tag{E.1.2}$$

where $\tilde{V}_{i(k|j)rt}$ denotes the conditional value function $V_{i(k|j)rt} - e_{ijkrt}$. The occupation dummy variable may then be defined by replacing $U_{i(k|j)rt}$ with $V_{i(k|j)rt}$ in equation (10).

The training choice problem will be analoguous to Section 3.4.

E.2 Identification

Controlling for occupation-specific experience exp_k in the outcome equation as in model (v) and (vi) may lead to an additional source of bias resulting from the endogeneity of *past* selection. The selection problem from equation (14) in Section 4 may then be written as

$$E[\epsilon_{ijkrt}|M_{ijkrt} = 1, exp_k] \neq 0. \tag{E.2.1}$$

In order to see why randomising individuals into a specific training and occupation will not be enough to overcome this bias, consider again the simplified example from Section 4.2, but focus now on the estimation of model (vi) from Section E.1. Consider now combining the experiments from Sections 4.2.1 and 4.2.2 such that individuals are first randomly allocated to a training j, and subsequently randomly allocated to an occupation k. Given a specific level of occupation-specific experience in occupation $k = 1$, the selection bias in estimating paramter τ^{exp_k} may be written as

$$\begin{aligned} &E[\epsilon_{i1}|train_{i1} = 1, exp_1] - E[\epsilon_{i1}|train_{i1} = 0, exp_1] \\ &\quad = E[\epsilon_{i1}|\underbrace{(exp_1|train_{i1} = 1)}_{\substack{\text{obtained exp. in occ. 1} \\ \text{cond. on training 1}}}] - E[\epsilon_{i1}|\underbrace{(exp_1|train_{i1} = 0)}_{\substack{\text{obtained exp. in occ. 1} \\ \text{cond. on training 2}}}] \leq 0. \end{aligned} \tag{E.2.2}$$

Intuitively, even though training is randomly allocated, comparing individuals with the same level of experience in occupation $k = 1$ on versus off the diagonal leads to bias as both sets of workers acquired that experience conditional on different trainings $j = 1$ and $j = 2$. As in Sections 4.2.1 and 4.2.2, the final inequality follows from the assumptions made on the error terms. The proof from Appendix B.1 easily extends to equation (E.2.2) as occupation-specific experience corresponds to repeated past selection into occupations.

As a result, an additional instrument is required to account for the endogeneity of exp_k when estimating models (v) and (vi). In line with occupation-specific experience being equivalent to repeated past selection into occupations, I use the average of past shocks to vacancies *outside* the current occupation as an instrument for exp_k:

$$IV_{exp_k} = 1/(t-1) \sum_{s=t_0}^{t-1} (vac_{k'rs} - E_{t_0}[vac_{k'rs}|\Omega_{r_0 t_0}]) \quad \forall k' \neq k. \tag{E.2.3}$$

The instrument is based on the intuition that higher average numbers of past vacancies outside occupation k will lead to lower accumulated experience in occupation k. Given the structure of the sequential selection problem, the identification assumptions discussed in Section 4.4 can be easily extended to IV_{exp_k}.

E.3 Estimation

When controlling for occupation-specific experience in the outcome equation, the control function needs to be adapted to take into account the potential endogeneity of exp_k. Following a common approach in the literature (see e.g. Gathmann & Schönberg (2010) and Dustmann & Meghir (2005)), I assume that the regression error term u_{ijkrt} from equations (26) and (28) which, given the inclusion of exp_k will no longer be mean-zero, may be written as

$$E[u_{ijkrt}|exp_k] = \rho_1 \hat{v}, \tag{E.3.1}$$

where $\hat{v} = exp_k - e\hat{x}p_k$ is the residual after predicting the endogenous variable exp_k with the instruments described in Section E.2. In the present context, this is equivalent to assuming linear separability in the control functions for the training and occupation choice, and the control function for exp_k, such that the full regression error term may be written as

$$E[\epsilon_{ijkrt}|M_{ijkrt} = 1, exp_k] = \lambda_{jk}(p_{ijr_0t_0}, p_{i(k|j)rt}) + \rho_1 \hat{v}. \tag{E.3.2}$$

While common in the literature, the imposed linearity in $\hat{v}$ is a strong assumption (Imbens & Wooldridge (2007)). I will therefore only rely on this modification of the control function approach in a small set of results and the majority of my findings will focus on the full effect of working in a specific jk-cell which includes the effect of any occupation-specific experience accumulated as a result of the cell choice.

E.4 Estimating the Selection Probabilities

In order to implement the full control function approach for models (v) and (vi), the residuals $\hat{v} = exp_k - e\hat{x}p_k$ need to be estimated. To do so, I use the training choice $train_{ij}$, the training instruments IV_{train_j}, the occupation-specific experience instruments IV_{exp_k} as well as past shocks in the current occupation k, and controls X_{ijkrt} as explanatory variables to predict exp_k using a random forest algorithm. Since exp_k is a continuous variable, the regression version of the algorithm is used. Reduced form error terms $\hat{v}$ will then be given by the difference between the actual and predicted values of occupation-specific experience.

In a second step, I also include the occupation-specific experience instruments IV_{exp_k} as explanatory variables to predict $occ_{i(k|j)rt}$ and obtain new estimates for the occupation selection probabilitites $p_{i(k|j)rt}$. With $\hat{v}$ and $\hat{p}_{i(k|j)rt}$ at hand, the control function approach may be implemented as outlined in Sections 5 and E.3.

Appendix F. Empirical Analysis

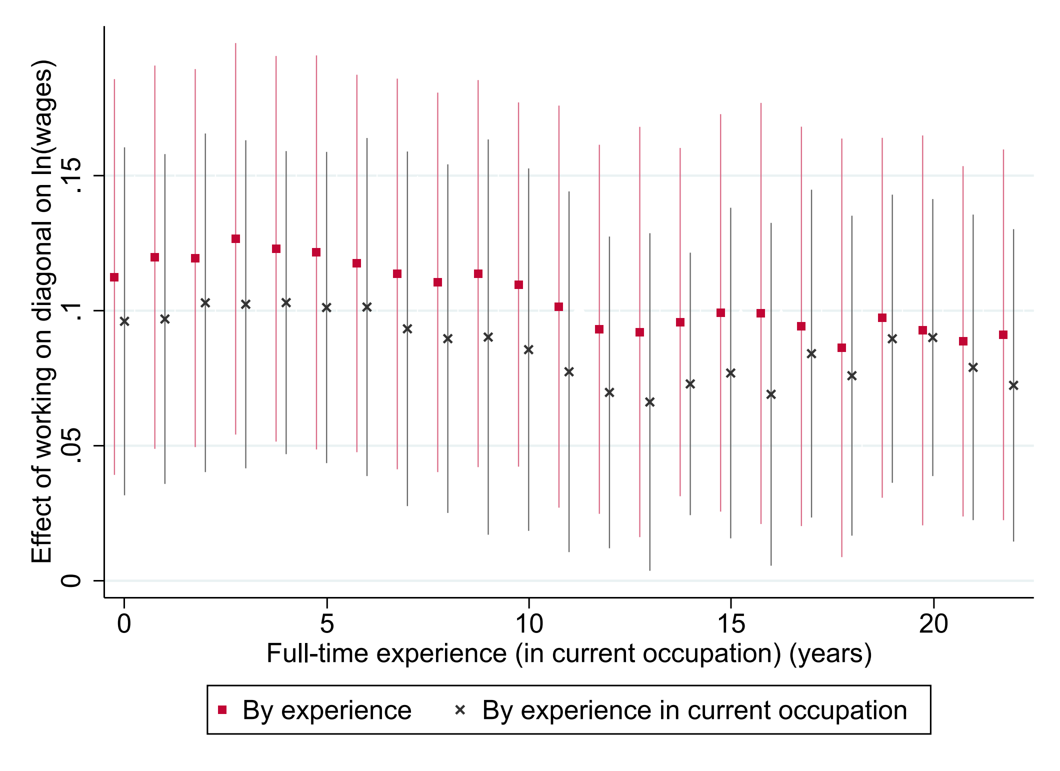
F.1 Further Results

F.1.1 Heterogeneity by Occupation-Specific Experience

The results for model (v) in Section E.1 are reported and discussed in Section 6.1 (see Table 4, columns (3) and (4)). Figure F.5 plots the results for model (vi), where a separate coefficient τ^{exp_k} has been estimated for each yearly occupation-specific experience bin. Each coefficient compares workers with a specific level of experience in their current occupation who were trained in that occupation to workers with the same level of experience in their current occupation who were *not* trained in that occupation.

Coefficient estimates are shown for the parametric control function estimator that takes into account the potential endogeneity of both the selection into a specific jk-cell and the occupation-specific experience parameter exp_k (see Chapter 5 and Appendix E.2 to E.4). In line with the results from column (4) in Table 4 in Section 6.1, the magnitude of the set of coefficients by occupation-specific experience is lower than the corresponding coefficients by experience. Moreover, coefficients are fairly constant in early years which is consistent with the accumulation of occupation-specific experience suggested to explain the patterns in Section 6.2. Given the structure of the dataset, comparisons of the two series become more difficult at higher levels of occupation-specific experience as the samples become increasingly similar. Nonetheless, the pattern suggests that even after accounting for occupation-specific experience, there is no full catch-up for workers who were not trained in their current occupation, and sizeable differences remain after more than 20 years of work experience.

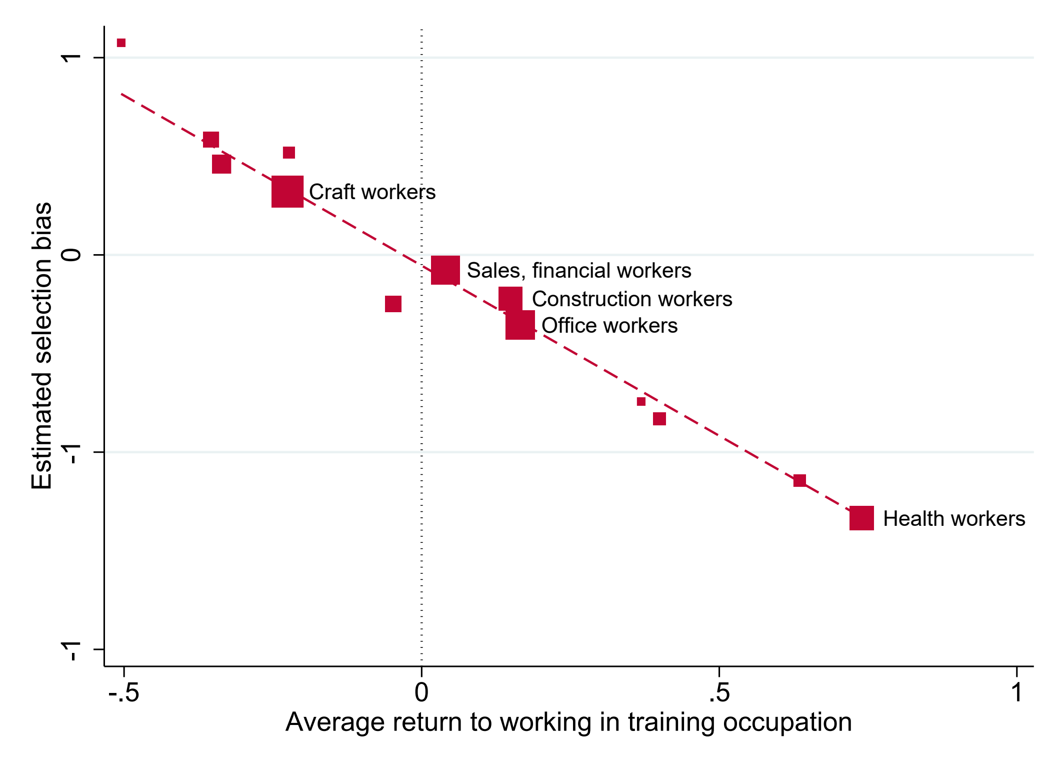
Figure F.1: On- versus Off-Diagonal Returns by Total and Occupation-Specific Experience



Notes: The figure plots regression coefficient estimates for τ^{exp_k} from model (vi), where occupation-specific experience levels have been binned into yearly categories. Coefficients are estimated using the parametric control function estimator which takes into account the potential endogeneity of occupation-specific experience. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. Standard errors are clustered at the region and time level. 95% confidence intervals are shown.

F.1.2 Heterogeneity by Training

Figure F.2: Average Return and Estimated Selection Bias by Training



Notes: The figure plots average on- vs. off-diagonal returns for each training from Figure 7 against the estimated selection bias, i.e. the difference between the returns estimated without selection control and those estimated 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.

F.1.3 Full Training-Occupation Matrix

Table F.1: Full Matrix of Returns - Within-Training Comparisons - 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	0	0.01	0.06	0.05	0.14	-0.00	0.05	0.09	0.06^{**}	-0.00	0.06	0.02	-0.15^{***}
			(0.03)	(0.04)	(0.06)	(0.09)	(0.04)	(0.04)	(0.06)	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)
	10-54	-0.27^{***}	0	-0.08^{*}	-0.11	-0.03	-0.06	-0.15^{**}	0.09^{***}	-0.03	-0.08^{***}	-0.01	-0.06^{**}	-0.51^{***}
		(0.08)		(0.03)	(0.09)	(0.08)	(0.05)	(0.06)	(0.02)	(0.05)	(0.02)	(0.03)	(0.02)	(0.12)
	19-32	-0.07^{*}	0.03	0	0.03	0.06	-0.01	0.01	0.15^{***}	0.07^{***}	-0.00	0.09^{***}	-0.04^{**}	-0.20^{***}
		(0.04)	(0.02)		(0.02)	(0.05)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.05)
	31	-0.04	0.06^{**}	0.08^{***}	0	-0.04	0.08^{*}	-0.02	0.17^{***}	0.13^{***}	0.00	0.19^{***}	0.02	-0.07
		(0.04)	(0.02)	(0.02)		(0.05)	(0.04)	(0.04)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.06)
	33-37	-0.12	0.03	0.11^{**}	-0.11	0	-0.20	0.09	0.10^{*}	0.02	-0.05	0.06	-0.01	-0.30^{***}
		(0.15)	(0.05)	(0.04)	(0.14)		(0.12)	(0.06)	(0.04)	(0.06)	(0.04)	(0.05)	(0.07)	(0.05)
	39-43	0.00	0.02	0.13^{***}	0.07	0.02	0	0.10^{***}	0.19^{***}	0.12^{***}	0.07^{***}	0.13^{***}	0.07	-0.04
		(0.04)	(0.03)	(0.02)	(0.05)	(0.07)		(0.03)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)	(0.03)
	44-51	-0.08^{**}	-0.07^{**}	0.01	-0.05^{**}	-0.16^{**}	-0.05	0	0.07^{***}	-0.03^{*}	-0.07^{***}	0.00	-0.11^{***}	-0.18^{***}
		(0.03)	(0.02)	(0.02)	(0.02)	(0.05)	(0.03)		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
	60-63	-0.03	-0.13^{***}	-0.01	-0.09^{**}	-0.11	-0.00	-0.04	0	0.06	-0.10^{***}	0.00	-0.13^{*}	-0.50^{**}
		(0.06)	(0.03)	(0.04)	(0.03)	(0.11)	(0.10)	(0.05)		(0.04)	(0.03)	(0.03)	(0.06)	(0.17)
	68-70	-0.27^{***}	-0.03	0.08^{**}	0.08	0.00	-0.08	-0.02	0.10^{***}	0	-0.02	0.01	-0.08^{**}	-0.26^{***}
		(0.06)	(0.03)	(0.03)	(0.06)	(0.05)	(0.05)	(0.05)	(0.03)		(0.02)	(0.10)	(0.03)	(0.04)
	71-74	-0.17	-0.05	-0.01	-0.12	-0.08^{*}	-0.17	0.03	0.07	0.00	0	0.01	-0.04	-0.20
		(0.11)	(0.04)	(0.03)	(0.10)	(0.04)	(0.19)	(0.07)	(0.04)	(0.05)		(0.03)	(0.07)	(0.15)
	75-78	-0.36^{***}	-0.11^{***}	-0.03	-0.15^{**}	-0.22	-0.21^{*}	-0.02	0.07^{***}	0.07^{***}	-0.07^{**}	0	-0.05^{*}	-0.37^{***}
		(0.10)	(0.03)	(0.04)	(0.06)	(0.29)	(0.10)	(0.05)	(0.02)	(0.01)	(0.02)		(0.02)	(0.07)
	79-89	-0.21^{***}	-0.06	-0.06^{*}	0.12	-0.24^{**}	-0.23^{**}	-0.10	0.04	0.00	-0.09^{**}	-0.03	0	-0.47^{***}
		(0.05)	(0.04)	(0.03)	(0.10)	(0.09)	(0.08)	(0.09)	(0.03)	(0.02)	(0.03)	(0.02)		(0.06)
	90-93	-0.08	0.23^{***}	0.36^{***}	0.32^{***}	0.32^{***}	0.07^{**}	0.22^{**}	0.23^{***}	0.17^{***}	0.18^{***}	0.18^{***}	0.18^{***}	0
		(0.10)	(0.04)	(0.06)	(0.08)	(0.09)	(0.03)	(0.08)	(0.05)	(0.03)	(0.04)	(0.02)	(0.03)	

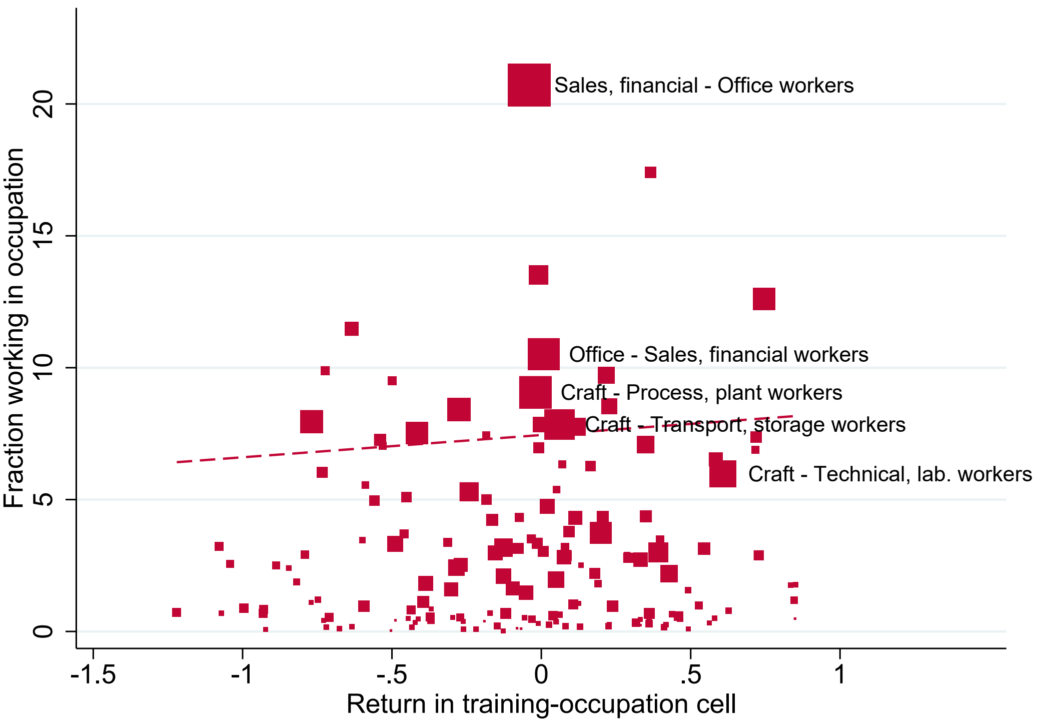
Notes: The table shows coefficient estimates $\hat{\tau}_{jk}$ from model (iv), estimated without selection control. Results are based on the baseline sample, restricting the sample to spells which started after the end of an apprenticeship. Observations are weighted using the empirical training-occupation distribution in 2010. Results are shown for the five largest occupations. 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. $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Table F.2: Full Matrix of Returns - Within-Training Comparisons - 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	0	-0.73^{***}	-0.56^{***}	-0.73^{***}	-0.08	-0.30	-0.03	0.19	-0.18^{*}	-0.54^{***}	-0.45^{***}	-0.46^{*}	-0.75^{***}
			(0.13)	(0.10)	(0.15)	(0.67)	(0.24)	(0.18)	(0.12)	(0.09)	(0.12)	(0.07)	(0.22)	(0.21)
	10-54	0.33	0	-0.07	-0.41	0.42	0.23	0.49^{*}	0.72^{***}	0.40^{**}	-0.01	0.17	0.08	-0.45^{*}
		(0.23)		(0.16)	(0.23)	(0.67)	(0.32)	(0.23)	(0.13)	(0.17)	(0.12)	(0.15)	(0.18)	(0.22)
	19-32	0.36	-0.02	0	-0.05	0.41	0.32	0.43^{**}	0.61^{***}	0.39^{***}	0.06	0.20^{**}	0.05	-0.12
		(0.22)	(0.08)		(0.09)	(0.44)	(0.18)	(0.14)	(0.07)	(0.06)	(0.07)	(0.07)	(0.12)	(0.14)
	31	0.46^{*}	0.02	0.12	0	0.49	0.42	0.53^{**}	0.75^{***}	0.55^{***}	0.11	0.35^{**}	0.18	0.05
		(0.24)	(0.11)	(0.10)		(0.56)	(0.24)	(0.22)	(0.07)	(0.09)	(0.09)	(0.12)	(0.12)	(0.16)
	33-37	-0.49	-0.72^{***}	-0.53^{**}	-0.77^{**}	0	-0.37	0.13	0.07	-0.18	-0.59^{***}	-0.50^{*}	-0.60^{**}	-0.85^{***}
		(0.34)	(0.22)	(0.19)	(0.26)		(0.57)	(0.32)	(0.22)	(0.20)	(0.17)	(0.23)	(0.25)	(0.26)
	39-43	0.62^{*}	-0.00	0.21	-0.06	0.56	0	0.73^{***}	0.85^{***}	0.58^{**}	0.22	0.35^{**}	0.29	0.09
		(0.28)	(0.15)	(0.18)	(0.20)	(0.75)		(0.15)	(0.17)	(0.19)	(0.21)	(0.13)	(0.26)	(0.23)
	44-51	0.11	-0.42^{***}	-0.24^{**}	-0.37^{**}	0.03	-0.03	0	0.33^{***}	0.08	-0.28^{**}	-0.15	-0.27^{*}	-0.44^{**}
		(0.22)	(0.08)	(0.10)	(0.12)	(0.59)	(0.19)		(0.07)	(0.08)	(0.09)	(0.10)	(0.13)	(0.16)
	60-63	-0.26	-1.08^{***}	-0.89^{***}	-1.04^{***}	-0.07	-0.42	-0.17	0	-0.31^{**}	-0.82^{***}	-0.63^{**}	-0.79^{***}	-1.07^{***}
		(0.33)	(0.22)	(0.23)	(0.20)	(0.66)	(0.34)	(0.32)		(0.13)	(0.15)	(0.20)	(0.13)	(0.25)
	68-70	0.08	-0.28^{*}	-0.10	-0.15	0.22	0.04	0.36^{**}	0.46^{***}	0	-0.13	-0.04	-0.13	-0.39^{**}
		(0.30)	(0.15)	(0.13)	(0.16)	(0.53)	(0.25)	(0.15)	(0.14)		(0.13)	(0.10)	(0.17)	(0.16)
	71-74	0.85^{*}	0.05	0.29	-0.19	0	0.33	0.83^{***}	0.85^{**}	0.72^{**}	0	0.37	0.42	0.13
		(0.45)	(0.28)	(0.31)	(0.34)		(0.41)	(0.26)	(0.31)	(0.30)		(0.25)	(0.41)	(0.44)
	75-78	-0.22	-0.59^{***}	-0.37^{*}	-0.68^{***}	-0.13	-0.26	0.13	0.24^{**}	0.01	-0.39^{***}	0	-0.30^{**}	-0.71^{***}
		(0.24)	(0.14)	(0.19)	(0.18)	(0.47)	(0.28)	(0.25)	(0.10)	(0.08)	(0.10)		(0.13)	(0.13)
	79-89	-0.63^{**}	-1.00^{***}	-0.93^{***}	-0.92^{***}	-0.50	-0.72^{**}	-0.43^{*}	-0.27^{*}	-0.49^{***}	-0.93^{***}	-0.77^{***}	0	-1.22^{***}
		(0.27)	(0.16)	(0.16)	(0.19)	(0.55)	(0.28)	(0.23)	(0.13)	(0.09)	(0.11)	(0.10)		(0.13)
	90-93	0.06	-0.16	0.01	-0.01	0.36	0.08	0.58^{**}	0.43^{***}	0.23	-0.08	-0.01	-0.01	0
		(0.29)	(0.15)	(0.16)	(0.19)	(0.57)	(0.25)	(0.24)	(0.12)	(0.14)	(0.19)	(0.16)	(0.13)	

Notes: The table shows coefficient estimates $\hat{\tau}_{jk}$ from model (iv), estimated using the parametric selection control described in Section 5.3. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4) and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. 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. $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Figure F.3: Full Matrix of Returns and Sample Fraction

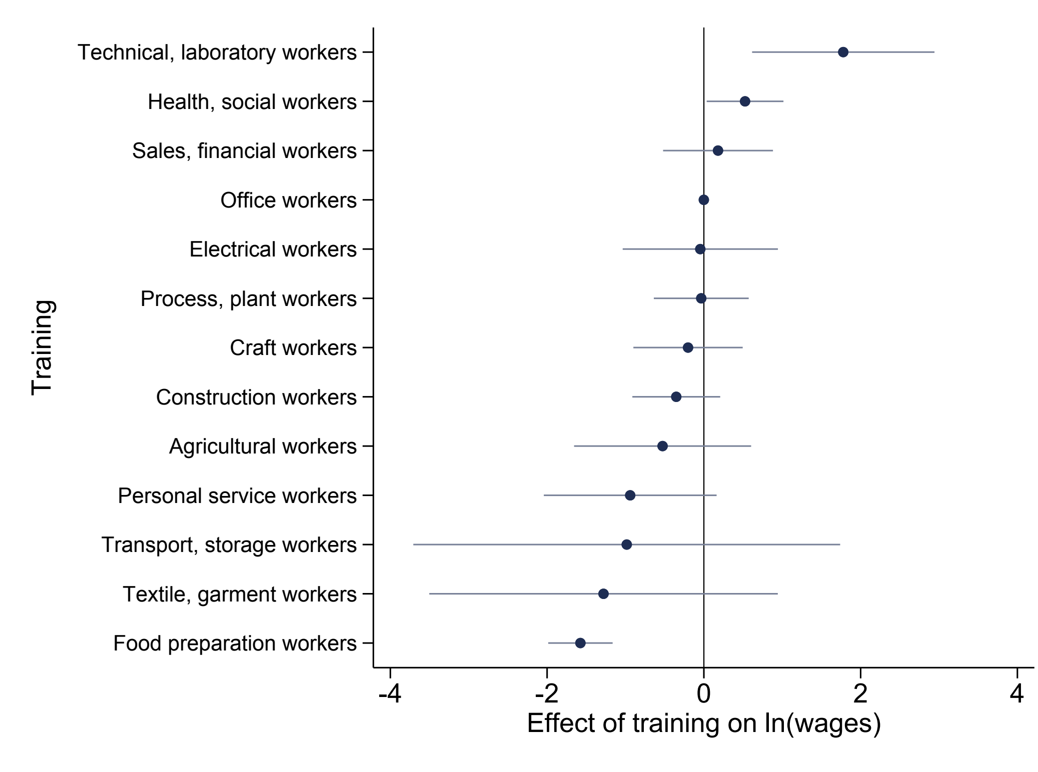


Notes: The figure plots training-occupation cell returns against the within-training sample fraction of workers in the relevant occupation for each off-diagonal training-occupation pair. The fitted line corresponds to a weighted OLS regression where each training-occupation pair is weighted by the fraction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

F.1.4 On-Diagonal Returns

Figure F.4 presents the results from equation (C.2.3) in Appendix C.2, where the omitted share is the one for the largest training, *office workers*. It can be seen that while many trainings appear to have on-diagonal returns which are similar to the one for *office workers*, there are significant differences on either side. The highest level effect is estimated for *technical and laboratory workers*, the lowest for *food preparation workers*.

Figure F.4: On-Diagonal Returns Across Trainings



Notes: The figure plots results from a district-level regression (equation (C.2.3)) which controls for population size and density. All coefficients are relative to training as *office workers*. Standard errors are clustered at the region level. 95% confidence intervals shown.

F.1.5 Within-Occupation Comparisons

Table F.3 adds the estimated level effects $\hat{\delta}_j$ to the coefficient estimates from Table 5 in Section 6.4, and displays the resulting sums $(\hat{\tau}_{jk} + \hat{\delta}_j)$. Table F.4 further substracts the on-diagonal effect in each *occupation* $(\hat{\tau}_{jk} + \hat{\delta}_j) - (\hat{\tau}_{j'k} + \hat{\delta}_{j'})$, $j' = k$, so that the coefficients can be interpreted relative to the diagonal within each *occupation*. It can be seen that the majority of coefficients is negative, suggesting that workers who were trained in other occupations incur wage penalties relative to their co-workers in that same occupation who also went through the relevant training. Comparing the coefficients in all five occupations suggests that *health and social workers* are the occupation in which workers with training in other occupations incur the largest wage penalties. Perhaps surprisingly, there are two significant *positive* coefficients, for *construction workers* trained as *craft workers* or *sales and financial workers*. However, note that *construction workers* is by far the smallest occupation amongst the ones displayed and only about 2%/0.3% of its workers were trained as *craft workers*/*sales and financial workers*, respectively.

Table F.3: Full Matrix of Returns - Added Effects

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	0	-0.37^{*}	0.01	-0.30^{**}	0.13
			(0.0840)	(0.9201)	(0.0461)	(0.6169)
	Craft workers	-0.00	-0.20	0.19	-0.15	0.23
		(0.9936)	(0.5146)	(0.5659)	(0.6621)	(0.5131)
	Sales, financ. w.	0.14	0.08	0.18	0.05	0.54
		(0.6798)	(0.8013)	(0.5688)	(0.8793)	(0.1495)
	Health, social w.	-0.24	-0.41	0.04	0.53^{**}	0.09
		(0.3483)	(0.1263)	(0.8837)	(0.0365)	(0.7424)
	Construction w.	-0.51^{*}	-0.59^{**}	-0.28	-0.62^{**}	-0.35
		(0.0735)	(0.0223)	(0.3292)	(0.0303)	(0.1815)

Notes: The table shows the sum of coefficients $\hat{\tau}_{jk}$ from model (iv), estimated using the parametric selection control described in Section 5.3 (see Table 5), and $\hat{\delta}_j$ from equation (C.2.3) (see Figure F.4). Results are shown for the five largest occupations. Standard errors are clustered at the region and time level. *p-values* for the test $(\tau_{jk} + \delta_j) = 0$ are shown in parentheses. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table F.4: Full Matrix of Returns - Within-Occupation Comparisons

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	0	-0.17	-0.17	-0.83^{**}	0.48
			(0.6963)	(0.6263)	(0.0309)	(0.2177)
	Craft workers	-0.00	0	0.01	-0.68^{*}	0.58^{**}
		(0.9936)		(0.9696)	(0.0647)	(0.0254)
	Sales, financ. w.	0.14	0.29	0	-0.47	0.89^{***}
		(0.6798)	(0.2393)		(0.1194)	(0.0089)
	Health, social w.	-0.24	-0.20	-0.14	0	0.44
		(0.3483)	(0.4690)	(0.5509)		(0.1062)
	Construction w.	-0.51^{*}	-0.39^{*}	-0.46^{*}	-1.15^{***}	0
		(0.0735)	(0.0588)	(0.0623)	(0.0002)	

Notes: The table shows the sum of coefficients $\hat{\tau}_{jk}$ from model (iv), estimated using the parametric selection control described in Section 5.3 (see Table 5), and $\hat{\delta}_j$ from equation (C.2.3) (see Figure F.4), further substracting the on-diagonal coefficient in the same *column*. Results are shown for the five largest occupations. Standard errors are clustered at the region and time level. *p-values* for the test $(\tau_{jk} + \delta_j) - (\delta_{j'} + \tau_{j'k}) = 0$, $j' = k$, are shown in parentheses. $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

F.2 Further Robustness

Table F.5 shows further robustness results relating to the estimation method. Column (1) approximates the skill price in occupation k using a tenth order polynomial instead of the fourth order polynomial used throughout Chapter 6. This functional form change leaves the baseline estimate unchanged at 12.3%. 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 10%. Columns (3) and (4) address two of the identification concerns discussed in Section 4.4.1, showing that the inclusion of occupation *times* time or industry *times* time fixed effects does not appreciably change the baseline result of 12.3%.

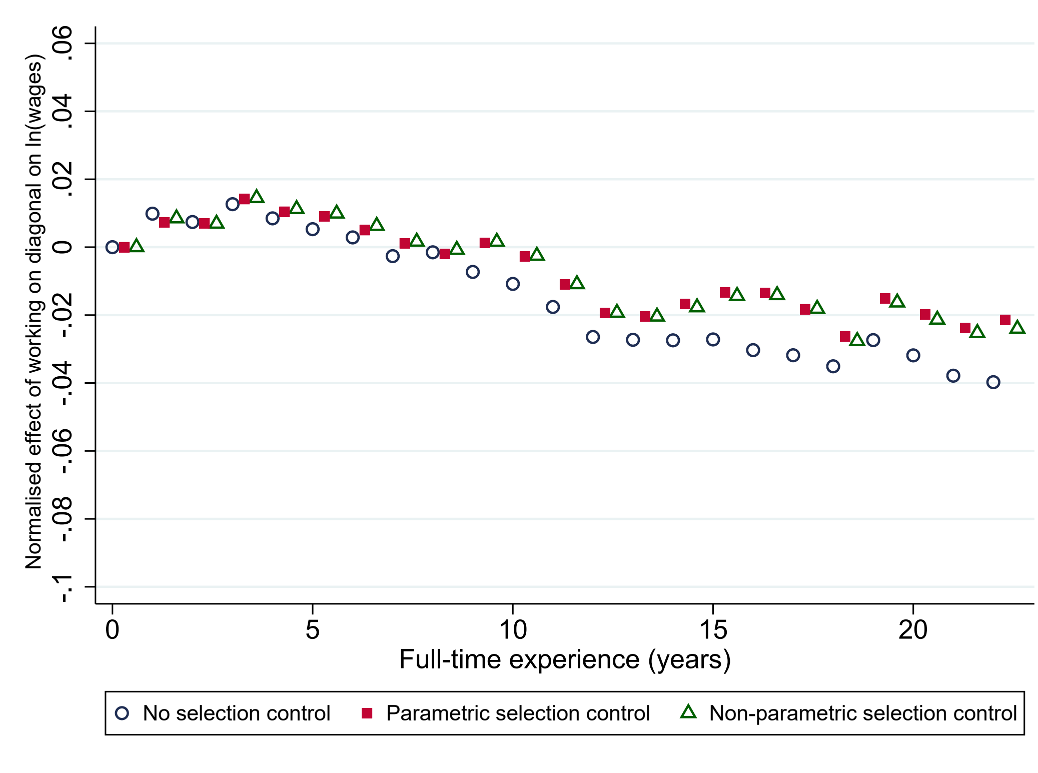
Table F.5: Average On- versus Off-Diagonal Returns - Estimation Robustness

	(1) 10th-order polynomial	(2) full set of cf cells	(3) occ. x time FE	(4) ind. x time FE
$D_{j=k} = 1$	0.1227^{***}	0.0990^{***}	0.1322^{***}	0.1159^{***}
	(0.0351)	(0.0304)	(0.0354)	(0.0325)
exp	0.0599^{***}	0.0598^{***}	0.0587^{***}	0.0577^{***}
	(0.0033)	(0.0025)	(0.0022)	(0.0049)
exp^2	-0.0010^{***}	-0.0004^{***}	-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	1,143,782	1,143,782	1,143,782	1,142,384

Notes: The table reports regression results from model (i). 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. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4) and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010. 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)). $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Figure F.5 provides a comparison between the slope estimates from τ^{exp} from (ii) under no selection control, the parametric selection and the non-parametric selection control where, in contrast to Figure 6 in Section 6.2, the probability of selection into one's training occupation $p_{i(k=j|j)rt}$ has been added as 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 6, closely mapping the slope estimates obtained when using the parametric selection control. This result provides further support to the distributional assumptions made.

Figure F.5: Normalised On- versus Off-Diagonal Returns by Work Experience - Robustness

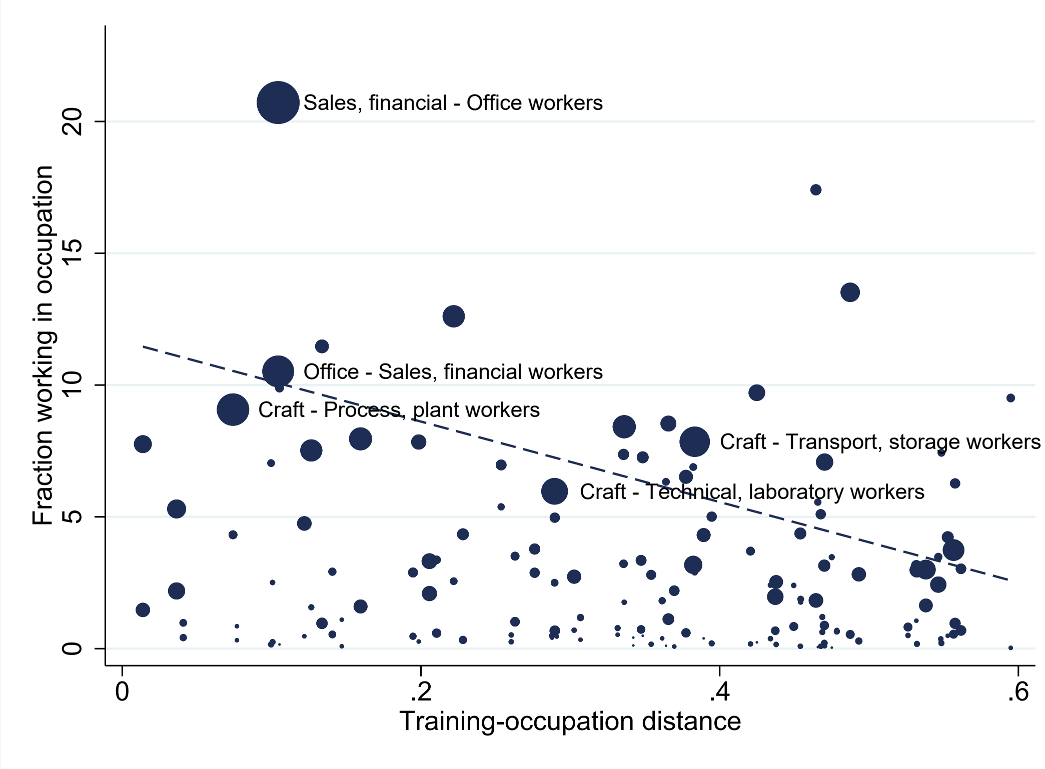


Notes: The figure plots regression coefficient estimates for τ^{exp} from model (ii), where experience levels have been binned into yearly categories. All coefficient estimates are normalised to zero at zero years of work experience. In contrast to Figure 6 in Section 6.2, the non-parametric control function estimator also includes the on-diagonal probability $p_{i(k=j|j)rt}$. Results are based on the baseline sample, further excluding years where the instruments are not available (see Section 5.4), and restricting the sample to spells which started after the end of an apprenticeship. 50% of observations are randomly selected as test sample (see Appendix D.4). Observations are weighted using the empirical training-occupation distribution in 2010.

Appendix G. Task Content

G.1 Descriptives

Figure G.1: 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.

Table G.1: Training-Occupation Distances - Selected Categories

Statistics	Training j	Occupation k	$Dist_{jk}$
Overall mean			0.3418
Standard dev.			0.1549
Weight. mean			0.2774
	Craft workers	Electrical workers	0.0138
	Craft workers	Construction workers	0.0364
	Construction workers	Electrical workers	0.0409
	Craft workers	Process, plant workers	0.0742
	Food prep. workers	Textile, garment workers	0.0768
	.	.	.
	.	.	.
	.	.	.
	Process, plant workers	Personal service workers	0.5527
	Office workers	Craft workers	0.5566
	Office workers	Process, plant workers	0.5576
	Craft workers	Personal service workers	0.5616
	Office workers	Textile, garment workers	0.5949

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. $Dist_{jk}$ is computed using survey waves 1885/86, 1991/92 and 1998/99.

Table G.2: Training-Occupation Distances - Five Largest Occupations

		Occupation				
		Office workers	Craft workers	Sales, financ. workers	Health workers	Constr. workers
Training	Office workers	0				
	Craft workers	0.56	0			
	Sales, financ. w.	0.10	0.54	0		
	Health, social w.	0.16	0.44	0.21	0	
	Construction w.	0.53	0.04	0.49	0.44	0

Notes: The table reports the distance measure $Dist_{jk}$, computed using survey waves 1885/86, 1991/92 and 1998/99, for the five largest occupations.

G.2 Further Results

Table G.3: Match Returns and Task Distance - No Selection Control

	(1)	(2)	(3)
$Dist_{jk}$	-0.0202^{*}	-0.0198^{**}	-0.0031
	(0.0104)	(0.0089)	(0.0083)
$Dist_{jk} \times Down_{jk}$			-0.0352
			(0.0242)
$Down_{jk}$		-0.1011^{***}	-0.0987^{***}
		(0.0175)	(0.0188)
Mean of $\hat{\tau}_{jk}$	-0.0188	-0.0188	-0.0188
Train. FE	yes	yes	yes
Adj. R^2	0.3369	0.4192	0.4252
N	156	156	156

Notes: The table reports regression results from equation (37), where the dependent variable τ_{jk} has been estimated without selection control. $Dist_{jk}$ is constructed using survey waves 1985/86, 1991/92 and 1998/99, and scaled by its standard deviation. Diagonal coefficients (where $\hat{\tau}_{jk} = 0$ and $Dist_{jk} = 0$) have *not* been included in the regression. Standard errors are clustered at the training level. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

G.3 Robustness

Table G.4: Match Returns and Task Distance - Include On-Diagonal Coefficients

τ_{jk} estimated	without selection control			with parametric control fcn.		
	(1)	(2)	(3)	(4)	(5)	(6)
$Dist_{jk}$	-0.0164	-0.0090	0.0069	-0.0394^{*}	-0.0294^{*}	0.0134
	(0.0109)	(0.0086)	(0.0077)	(0.0218)	(0.0162)	(0.0289)
$Dist_{jk} \times Down_{jk}$			-0.0454^{*}			-0.1228^{**}
			(0.0209)			(0.0493)
$Down_{jk}$		-0.0929^{***}	-0.0955^{***}		-0.1254^{**}	-0.1326^{**}
		(0.0144)	(0.0176)		(0.0498)	(0.0501)
Mean of $\hat{\tau}_{jk}$	-0.0174	-0.0174	-0.0174	-0.0726	-0.0726	-0.0726
Train. FE	yes	yes	yes	yes	yes	yes
Adj. R^2	0.3087	0.3867	0.4044	0.5585	0.5694	0.5811
N	169	169	169	169	169	169

Notes: The table reports regression results from equation (37), where the dependent variable τ_{jk} has been estimated without selection control. $Dist_{jk}$ is constructed using survey waves 1985/86, 1991/92 and 1998/99, and scaled by its standard deviation. Diagonal coefficients (where $\hat{\tau}_{jk} = 0$ and $Dist_{jk} = 0$) have been included in the regression. Standard errors are clustered at the training level. $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Table G.5: Match Returns and Task Distance - Five Largest Trainings/Occupations

τ_{jk} estimated	without selection control			with parametric control fcn.		
	(1)	(2)	(3)	(4)	(5)	(6)
$Dist_{jk}$	-0.0145	-0.0124	-0.0028	-0.0396^{*}	-0.0375^{**}	0.0081
	(0.0131)	(0.0073)	(0.0098)	(0.0217)	(0.0150)	(0.0463)
$Dist_{jk} \times Down_{jk}$			-0.0212			-0.1009
			(0.0250)			(0.0878)
$Down_{jk}$		-0.1055^{***}	-0.1082^{***}		-0.1019	-0.1150^{**}
		(0.0160)	(0.0163)		(0.0581)	(0.0482)
Mean of $\hat{\tau}_{jk}$	-0.0175	-0.0175	-0.0175	-0.0898	-0.0898	-0.0898
Train. FE	yes	yes	yes	yes	yes	yes
Adj. R^2	0.3778	0.4733	0.4711	0.6596	0.6639	0.6679
N	100	100	100	100	100	100

Notes: The table reports regression results from equation (37), restricting the sample to the five largest trainings/occupation. $Dist_{jk}$ is constructed using survey waves 1985/86, 1991/92 and 1998/99, and scaled by its standard deviation. Diagonal coefficients (where $\hat{\tau}_{jk} = 0$ and $Dist_{jk} = 0$) have *not* been included in the regression. Standard errors are clustered at the training level. $^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Appendix H. Welfare and Policy

H.1 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.[73] In terms of private cost, the average yearly difference in earnings between an apprentice and a trained worker early in their career was about $17,560$ Euros in 2010.[74].

Net benefits in Euros

My estimates suggest that the annual average gain of retraining of τ corresponds to 10% of wages for the average worker. The cost of a year of foregone work experience is 6%. Assuming that the effective foregone work experience of two years of retraining is one year, the net gain of retraining is therefore equal to 4%. Based on average yearly earnings of $45,000$ Euros in 2010, this amounts to $1,800$ Euros in 2010.

Cost-benefit calculations

Assuming a discount factor of 0.99, retraining costs would be recouped after about 39 years of subsequent work in the new occupation:

$$29,460 + \beta \times 29,460 = \beta^2 \times 1,800 \times \frac{1-\beta^{t+1}}{1-\beta} \tag{H.1.1}$$

$$t \approx 39. \tag{H.1.2}$$

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 ($67-20-2-39=6$ years). Using Figure 2 in Section 2.5, it can be seen that around 35% of workers work off the diagonal by six years of experience. Based on a final share of 45%, this corresponds to a fraction of over three quarters.

In addition, the return to working on versus off the diagonal only drops by around $1pp$ from its peek of 12.5% by six years of experience (see Figure 5 in Section 6.2). Using the calculations in Section 8.1, this suggests that 18% of all workers are still locked in at six

[73] Figures based on cpi adjusted 2012/13 figures. Source: *Finanzierung der beruflichen Ausbildung in Deutschland, BWP 2/2016, BiBB.* Total net expenditures divided by total of 1.4 Mio. apprentices.

[74] Figure computed as average difference in earnings between apprentices and trained workers with less than ten years of full-time work experience in sample

years of expererience. Based on a final share of 20%, this corresponds to a fraction of 90% of locked-in workers.

H.2 Information Provision

Recall the key friction in the proposed framework is the ex-ante imperfect information at the time of training choice. Section 8 considers retraining as a potential policy instrument. This section will briefly discuss 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.[75] 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 programmes. In particular, my findings suggest that government programmes 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 $4,500$ Euros in 2010 for every year participants will subsequently spend working on instead of off the diagonal.[76] Moreover, using the 2010 empirical distribution of workers across training-occupation cells, I find that over 50% of off-diagonal workers could be retrained *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 programmes to be cost-effective.

[75] 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 practice.

[76] 10% of an average of $45,000$ Euros in 2010. Assuming a discount factor of β, future work years on the diagonal lead to a benefit of $(4,500 \times \beta^t)$ Euros.