Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

01/2012
EN

Methodological aspects of labour market data

New methods to estimate models with large sets of fixed effects with an application to matched employer-employee data from Germany

Nikolas Mittag

Bundesagentur für Arbeit

New methods to estimate models with large sets of fixed effects with an application to matched employer-employee data from Germany

Nikolas Mittag (University of Chicago)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract

This paper will introduce new methods to estimate the two-way fixed effects model and the match effects model in datasets where the number of fixed effects makes standard estimation techniques infeasible. The methods work for balanced and unbalanced panels and increase the speed of estimation without imposing excessive computational demands. I will apply the methods to a new and unusually detailed matched employer-employee dataset from Germany. The analysis shows that the omission of match effects leads to biased inference particularly concerning the effects of individual characteristics and underlines the importance of accurate biographic data.

Zusammenfassung

In Datensätzen mit einer hohen Anzahl von fixen Effekten sind Standardschätzmethoden nicht durchführbar. Das Papier stellt neue Methoden zum Schätzen von two-way fixed effects und match effects Modellen vor. Die Methoden sind für balanced und unbalanced Panel durchführbar und erhöhen zudem die Schätzgeschwindigkeit, ohne hohe Rechenleistung zu benötigen. Im vorliegenden Papier werden die Methoden auf den deutschen Linked Employer-Employee Datensatz (LIAB) angewandt. Die Analysen zeigen vor allem für die Effekte zu individuellen Charakteristika, dass das Auslassen von match effects zu verzerrten Inferenzen führt. Dies unterstreicht die Wichtigkeit präziser biografischer Daten.

Keywords: multi-way fixed effects, matching, linked employer-employee data, wage dispersion

I would like to thank Dan Black, Jeffrey Grogger and Bruce Meyer for helpful comments and suggestions as well as Stefan Bender and the staff at the IAB for the excellent cooperation with the data. All remaining errors are mine.

1 Introduction

This paper introduces a new method to estimate the two-way fixed effects model (TWFE) and the match effects model in datasets where the number of fixed effects makes standard estimation techniques infeasible. Following the seminal article by Abowd et al. (1999), these models have frequently been applied to linked employer-employee data (see Abowd et al. 2008, Woodcock 2008 and the references therein). With large datasets becoming increasingly available, more recent applications have also included student-teacher data (Kramarz et al. 2008) and doctor-patient data (Bennett 2010). The number of fixed effects in these examples ranges from about 700.000 to 1.8 million. The application in this paper includes up to 7.5 million fixed effects, which makes it impossible to estimate these models with the standard technique of including dummies for all fixed effects as this would require inverting an enormous matrix. As a consequence of these computational difficulties, many applications have not even considered the match effects model and/or estimated random or mixed effects models. The former is problematic since one would expect match effects to matter in many applications on theoretical grounds (see e.g. Jovanovic 1979, Mortensen 1978) and several recent papers have shown them to matter empirically in common applications (e.g. Jackson 2011, Woodcock 2008). As the application in this paper shows, the omission of match effects can lead to severely biased coefficient estimates, so tests for their presence should be conducted. A simple test is proposed in this paper. Random and mixed model specifications rely on orthogonality restrictions (e.g. that the two sets of fixed effects are uncorrelated), which greatly reduce the computational complexity and increase efficiency, but also lead to biased estimates if they are invalid. Because these restrictions are not required in the fixed effects models, they should be tested by the Hausman test (Hausman 1978) that one would expect researchers to do in the one-way fixed effects model. Consequently, as the restrictions imposed to circumvent the computational problems associated with fixed effects models are rarely justified by theory and often rejected by the data, it is important to be able to estimate

fixed effects specifications for both the TWFE and the match effects model in order to at least test these restrictions empirically.

The methods I propose in this paper greatly reduce the dimensionality of the matrix that needs to be inverted, which makes estimation a lot faster without excessive memory requirements. They work for both balanced and un-balanced panels. The methods can be used to estimate the slope coefficients only or the full model including the fixed effects. They provide computational advantages in both cases for the TWFE and the match effects model, but the advantages are more pronounced when estimating the full model including the fixed effects, particularly for the match effects model. The next part of this paper will introduce the two models, discuss estimation problems and previous solutions. Part 3 will introduce the new methods of estimation and discuss specification issues and tests. Finally, I will apply the methods to a very rich matched employer-employee dataset from Germany. The application confirms that match effects matter for the conclusions one can draw from such data and underlines the importance of accurate biographical information in wage regressions.

2 The Two-Way Fixed Effects (TWFE) and the Match Effects Model

This part provides background information on the two models at stake. The first section will introduce the TWFE and the match effects model and review identification. Then I will discuss estimation problems, how they have been solved in previous applications and how these solutions are related to the methods presented in this paper.

2.1 The Models

A common specification in panel data models is the unbalanced two-way fixed effects model which includes a set of fixed effects for primary units indexed by i=1,…,N and secondary units indexed by j=1,…,J. Thus, there are N primary units and J secondary units. Applications of this model include, among others, matched employer-employee data (in which the units are individuals and firms, i.e. a fixed effect for each individual and each firm is included; see

e.g. Abowd and Kramarz 1999) and student-school data (including fixed effects for pupils and schools; see e.g. Kramarz et al. 2008). The model is defined by:

$$y_{ijt} = x_{ijt}\beta + \theta_i + \psi_j + \varepsilon_{ijt} \tag{1}$$

Where y_{ijt} is unit *i*'s (scalar) outcome at time *t*, x_{ijt} is a 1×K vector of time-varying observed covariates, β is a vector of coefficients and θ_i and ψ_j are time-invariant scalar fixed effects. ε_{ijt} is the error term that satisfies the usual conditional mean independence assumption:

$$E(\varepsilon_{ijt} \mid x_{ijt}, \theta_i, \psi_j) = 0\,. \tag{2}$$

T_i indicates the number of observations on primary unit *i*, i.e. the subscript *t* runs from 1 to T_i for unit *i*. Similarly, let F_j stand for the number of observations on secondary unit *j*, so that the total number of observations N* is given by

$$N^* = \sum_{i=1}^{N} T_i = \sum_{j=1}^{J} F_j \tag{3}$$

The model allows both T_i and F_j to vary between units, i.e. the panel does not have to be balanced. In matrix notation, the model can be expressed as

$$\underset{N^*\times 1}{y} = \underset{N^*\times K}{X}\ \underset{K\times 1}{\beta} + \underset{N^*\times N}{D_\theta}\ \underset{N\times 1}{\theta} + \underset{N^*\times J}{D_\Psi}\ \underset{J\times 1}{\Psi} + \underset{N^*\times 1}{\varepsilon} \tag{4}$$

Where *y* is an N*×1 vector of outcomes, *X* is an N*×K matrix of observable time-varying covariates, D_θ is the N*×N matrix of indicators for the primary unit and D_Ψ is the N*×J matrix of indicators for the secondary unit and ε is the N*×1 vector of error terms. The parameters of the model are β, the K×1 vector of slopes, θ, the N×1 vector of fixed effects for the primary units and Ψ, the J×1 vector of fixed effects for the secondary units. Conditions for identification of these parameters will be discussed below. To ease notation, define

$$
\begin{aligned}
\underset{N\times N}{T} &= \underset{N\times N^{*}}{D_{\theta}'}\ \underset{N^{*}\times N}{D_{\theta}} = diag(T_1,...,T_N) \\
\underset{J\times J}{F} &= \underset{J\times N^{*}}{D_{\Psi}'}\ \underset{N^{*}\times J}{D_{\Psi}} = diag(J_1,...,J_J) \\
\underset{N\times J}{K} &= \underset{N\times N^{*}}{D_{\theta}'}\ \underset{N^{*}\times J}{D_{\Psi}}
\end{aligned}
\tag{5}
$$

So that *T* is an N×N diagonal matrix with the number of observation on primary unit *i* as the i[th] diagonal element and *F* is a J×J diagonal matrix with the number of observations on secondary unit *j* as the j[th] diagonal element. Element (*i*,*j*) of the N×J matrix K indicates how many observation on primary unit *i* belong to secondary unit *j*, e.g. how many periods individual *i* worked for firm *j*.

The match effects model (see e.g. Woodcock 2007, 2008) is an extension of this model that includes an interaction between the two fixed effects:

$$
y_{ijt} = x_{ijt}\beta + \theta_i + \psi_j + \lambda_s + \varepsilon_{ijt} \tag{6}
$$

Where the index *s*=1,…,*S* is redundant and for notational convenience only as it is determined by *i* and *j*: *s*=f(*i*,*j*). Compared to the TWFE model, the match effects model additionally includes an effect that within which both the firm and individual fixed effect are nested. As the discussion of identification conditions below shows, the mean of the match effects within each *i* and *j* is not identified and has to be normalized. In matrix notation, the model can be expressed as

$$
\underset{N^{*}\times 1}{y} = \underset{N^{*}\times K}{X}\ \underset{K\times 1}{\beta} + \underset{N^{*}\times N}{D_{\theta}}\ \underset{N\times 1}{\theta} + \underset{N^{*}\times J}{D_{\Psi}}\ \underset{J\times 1}{\Psi} + \underset{N^{*}\times S}{D_{\lambda}}\ \underset{S\times 1}{\lambda} + \underset{N^{*}\times 1}{\varepsilon} \tag{7}
$$

In addition to the matrices and vectors defined as above, D_{λ} is the N*×S matrix of indicators for matches between the two units and λ is the S×1 vector of match fixed effects. The mean independence assumption of the error term in this case becomes

$$
E(\varepsilon_{ijt} \mid x_{ijt}, \theta_i, \psi_j, \lambda_s) = 0 \tag{8}
$$

Both models allow the errors to be correlated arbitrarily within firms and individuals as will be discussed below and the panel does not have to be balanced. This paper deals with the case in which both sets of fixed effects include a large number of units, making estimation by standard techniques infeasible. The models can easily be amended to include an additional set of fixed effects such as time dummies in x_{ijt} as long as it is small enough to keep inversion of the *X'X*-matrix feasible. In the remainder of this paper, I will discuss these models in terms of matched employer-employee data, i.e. I will refer to *y* as (log) wages, θ and ψ as individual- and firm-fixed effects and λ as the match-fixed effects. Given that one can arbitrarily choose which of the two units is considered the primary unit, I will define the secondary unit to be the smaller unit in the sense that J<N without loss of generality. In this discussion, I will assume that there are fewer firms than individuals for ease of exposure, but the estimation method extends to other applications and cases with the variables defined analogously.

Conditions for identification of the TWFE model are discussed in Abowd et al. (2002). Essentially, they show how to sort individuals into connected groups and demonstrate that within each group N_g+J_g-1 effects are identified where N_g and J_g are the number of individuals and firms in group g. A connected group contains all the workers that have ever worked for any of the firms in the group and all firms that have ever employed one of the workers in the group. On the contrary, disconnected groups are marked by no realized mobility between the firms in the two groups. Because only N_g+J_g-1 effects are identified per group, a normalization is necessary. The estimation strategy below achieves identification by excluding an overall intercept and constraining the individual fixed effects within each group to sum to zero. Other normalizations can easily be implemented. The match effects model additionally includes an effect within which both the firm and individual fixed effect are nested. Consequently, its mean for each individual and firm is not identified and normalized to zero, i.e. match effects are constrained to sum to zero for each individual and each firm (see Woodcock 2008 for a discussion). Intuitively, the average match quality is an invariant characteristic of a firm and individual by construction. Consequently, it cannot be separately identified from the person

and firm effect. Not being identified implies that the data does not contain any information to tell the parameters apart, so if one allows non-zero individual and firm effects, it does not make much sense to talk about average match effects, since the distinction between the two effects cannot be made based on empirical facts. Normalizing the effect to be orthogonal to the other fixed effects has convenient computational properties and other normalizations can easily be implemented after estimation.

2.2 Estimation Issues

This section will first discuss the computational problems involved in estimating the models presented above when the number of fixed effects is large. I will then briefly present estimation strategies that have been used in previous applications of each model first if only the slopes are of interest and second if estimates of the fixed effects are required and compare them to the strategy I propose. More detailed discussions of these strategies and other methods (such as random effects or mixed models, which I will only briefly touch here) can be found in Abowd et al. (2002, 2008) and Andrews et al. (2005).

Standard OLS estimation of the models above would require calculation of the inverse of a matrix of size K+N+J for the two-way fixed effects model or K+N+J+S for the match effects model. In typical matched datasets, this number easily exceeds a million. Standard software programs cannot handle matrices of this size (e.g. the limit in Stata is 11.000). Even if the software does not impose such limits, calculations with matrices of this size require tremendous amounts of memory and computational power. For example, a full, square matrix with one million rows stored in double precision requires approximately 8000 GB of memory[1]. Conventional matrix inversion algorithms require the matrix and its inverse to be loaded into main memory, which is clearly infeasible even on large servers. The computation time required by these algorithms aggravates the problem, because one would need a large number of servers, potentially for days. The problems of computational time and memory require-

[1] These matrices are usually sparse, so they can be stored more efficiently, but they often have more than one million rows.

ments can be traded off against each other by using various versions of indirect solvers that do not require the whole matrix to be loaded into memory, but tend to be much slower.

In many applications (e.g. Bennett 2010) only the slope coefficients are of interest while the fixed effects are regarded as nuisance parameters. One has to control for the fixed effects, because they are correlated with the observables, but one is not interested in the estimates of the fixed effects themselves. In the case of the match effects model, the slopes are easily estimated, as subtraction of match specific means sweeps out all fixed effects and the slopes can be estimated by OLS on the transformed data. I will make use of this method in the proposed estimation strategy below and explain it in more detail there. In this case, estimation can be done using any standard statistical software, the only advantages of the programs that implement the algorithm described below is that they easily extend to cases where the errors are correlated within firm and/or individual (multi-way clustering) and that they adjust for degrees of freedom automatically. It is more common, however, that the researcher is interested in the slopes of the TWFE model only. The TWFE model is a restricted version of the match effects model (it restricts all match effects to equal zero), so the strategy above will still yield consistent, but inefficient estimates of the TWFE slopes. However, regressors that do not vary within match cannot be included, because they are purged when subtracting match specific means. Consequently, it is only a viable strategy to estimate the slopes if efficiency is not a concern and there are no match specific regressors. Wansbeek and Kapteyn (1989) show a transformation of the data for the two-way fixed effects model that yields the fully efficient estimates of the TWFE model when applying OLS to the transformed data. This transformation is extended to multi-way fixed effects models by Davis (2002). However, is computationally demanding and to my knowledge has not been implemented yet. The algorithm I propose makes use of this transformation and can thus be used to obtain efficient estimates of the TWFE slopes even if there are match specific regressors.

In other cases, at least one set of the fixed effects is of interest. The prevalent case here is the effects of time-invariant person and firm characteristics in the TWFE model. See e.g. Ab-

owd et al. 2008 for several parameters of interest that can be calculated from the estimates of the individual and firm fixed effects. In these applications, the researcher needs to estimate the full model given by equation 4. Early applications relied on approximate solutions to avoid excessive use of computational time. More recent solutions solve the problem by avoiding inversion of the cross-product matrix in the normal equations. There are two main approaches to this. The first approach relies on the fact that one does not need the inverse of the design matrix to obtain the coefficient estimates, which are defined by the normal equations. These equations can be solved much more easily by techniques that do not require matrix inversion, such as the conjugate gradient algorithm (CGA) employed by Kramarz et al. (2008) to estimate a model of academic achievement that includes fixed effects for pupils and schools. This algorithm is implemented in the widely used a2reg program for Stata. It relies on an iterative technique that converges to the exact OLS solution (up to rounding error) if and only if the matrix is invertible. The downside of this approach is that it is slow and does not yield standard errors of the coefficients, because these require the actual inverse to be computed. The second approach attempts to reduce the dimensionality of the matrix to be inverted by performing a first-difference or within transformation on the larger of the sets of fixed effects. This reduces the size of the matrix to K+min(N,J) for the two-way fixed effects model. Abowd et al. (1999) use the first-difference transformation on French employer-employee data. Andrews et al. (2005) advocate the within transformation and discuss several refinements of this algorithm. Both still have to invert a potentially very large matrix. An advantage is that one obtains standard errors for the slopes and the smaller set of fixed effects.

The estimation strategy I suggest in this paper can be seen as a combination of the three strategies above in that it uses a transformation to obtain the slopes, reduces the dimensionality of the matrix and then solves a system of linear equations rather than inverting the matrix. It will be computationally more efficient at estimating the full TWFE model than the two approaches above in most cases, particularly when compared to the last approach. Contrary to the first approach, it also produces standard errors of the slopes, even in cases where

errors are not iid. It does not produce standard errors of the fixed effects, but contrary to the second approach, it yields estimates of both sets of fixed effects.

All of the applications mentioned above use the TWFE model and to my knowledge there is no estimation strategy that works well for the full match effects model, because the inclusion of the match effect increases the size of the resulting problem to be solved tremendously. Consequently, there are very few applications that actually compute estimates of the fixed effects for the match effects model. However, estimating the match effects is of interest in models such as matching on the labor market (see e.g. Woodcock (2008) or the application in this paper) or international migration (e.g. Grogger and Hanson 2010). While the sample in Grogger and Hanson is small enough to estimate the model by standard methods, Woodcock has to restrict estimation to a subsample, because the inclusion of match effects makes the approaches that work for the TWFE model infeasible. Other applications have made additional assumptions to make estimation feasible. Woodcock (2007), for example, estimates mixed model specifications that rely on firm, person and match effect being orthogonal. The advantage of this assumption is that it greatly reduces the computational burden. However, the estimates will be biased if the orthogonality conditions do not hold, which is particularly questionable in the case of the firm and person effect. The algorithm I propose below makes estimation of these models feasible even in very large samples.

3 Estimation

This part will deal with applied issues in the estimation of the TWFE and the match effect models. Most importantly, I will propose an algorithm that allows quick estimation of the two models and is relatively undemanding of computational resources. I will first discuss the algorithm in some detail for the match effects model and will then point out how it can be adapted to estimate the TWFE model. Then I spell out how to address two specification issues: how to test for match effects and how to deal with two-way clustering of the standard errors. Programs that implement different versions of the algorithm in Matlab and Stata will be made available. Section 3.4 will briefly discuss the advantages of the two versions and present some Monte Carlo evidence on their performance.

3.1 Estimating the match effects model

In this section I will show how to estimate the match effects model. I will first discuss three key facts that greatly simplify the estimation and then present the algorithm to estimate the match effects model. A brief summary of the steps that are necessary to obtain the estimates can be found in Appendix A.

The approach outlined below yields the exact OLS estimates of all slopes and fixed effects of the match effect model. Standard errors for the slopes can be calculated by standard techniques or as in Cameron et al. (2006), which will be discussed in section 3.3. Standard errors for the fixed effects can be bootstrapped if desired. Techniques for doing so are discussed in Cameron et al. (2008) and depend on the correlation structure of the error.

The estimation strategy for both models rests on three key properties:

1. Partial Regression (e.g. Yule 1907, Lovell 1963)
2. If $\hat{\beta}^{OLS}$ are the OLS coefficients on *X* from a regression of *y* on [*X Z*], then the OLS coefficients on *Z* from regressing $(y - X\hat{\beta}^{OLS})$ on *Z* and regressing *y* on [*X Z*] are numerically identical.
3. The residuals sum to zero for every firm, individual and match.

Property 1 and 3 are well known, proofs can be found in most econometric text books. Property (2) is often shown in the context of partial regression (e.g. Green 2008 p. 27). For the application in this paper, *Z* should be thought of as the design matrix of the fixed effects, e.g. $Z = \begin{bmatrix} D_\theta & D_\Psi \end{bmatrix}$ for the TWFE model. It follows from the OLS normal equations:

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \begin{pmatrix} \hat{\beta}^{OLS} \\ \hat{\delta}^{OLS} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

As $\hat{\beta}^{OLS}$ is assumed to be known, the second line is a system of equations which is uniquely solved by $\hat{\delta}^{OLS}$. Rearranging and solving yields

$$\begin{aligned} Z'X\hat{\beta}^{OLS} + Z'Z\hat{\delta}^{OLS} &= Z'y & \Leftrightarrow \\ Z'Z\hat{\delta}^{OLS} &= Z'(y - X\hat{\beta}^{OLS}) & \Leftrightarrow \\ \hat{\delta}^{OLS} &= (Z'Z)^{-1}Z'(y - X\hat{\beta}^{OLS}) \end{aligned}$$

Regressing $(y - X\hat{\beta}^{OLS})$ on *Z* as proposed by property 2 yields:

$$\hat{\delta} = (Z'Z)^{-1}Z'(y - X\hat{\beta}^{OLS}) = \hat{\delta}^{OLS}$$

which shows that the estimates from the auxiliary regression are equal to $\hat{\delta}^{OLS}$, the estimates obtained from the full regression. Estimation of the match effects model then proceeds as follows:

First, obtain estimates of $\hat{\beta}^{ME}$, the slopes from the match effects model defined by eq. 7. As was pointed out in section 2.2, this can be done by an OLS regression of deviations of y_{ijt} and x_{ijt} from their match means. This is equivalent to running a partial regression, i.e. regressing the residuals from a regression of *y* on the three sets of fixed effects on the residuals from regressions of each column of *X* on the three sets of fixed effects. Because individual and firm fixed effects are constant within match, these residuals are easily computed as the deviations of y_{ijt} and x_{ijt} from their match means (see e.g. Andrews et al. 2005 and

Woodcock 2008 for formal proof). The estimates, residuals and standard errors (corrected for the degrees of freedom) from these partial regressions are numerically identical to those that would be obtained from OLS estimation of the full model (Yule 1907).

Because $\hat{\beta}^{ME}$ from the previous step is the OLS estimate, estimates of the fixed effects can be obtained by a regression of $\tilde{y} = y - X\hat{\beta}^{ME}$ on the two sets of fixed effects by property (2). Note that the normalization of having match effects sum to zero within each firm and individual allows one to omit the match effects in this regression. In a regression that includes the three sets of fixed effects only, the effect of omitting the match effects is given by the usual formula for omitted variable bias: $\left[\begin{bmatrix} D_\theta & D_\Psi \end{bmatrix}' \begin{bmatrix} D_\theta & D_\Psi \end{bmatrix}\right]^{-1} \begin{bmatrix} D_\theta & D_\Psi \end{bmatrix}' D_\lambda \lambda$. This bias will always be exactly 0 by construction, because the normalization of having match effects sum to zero within firm and individual implies that $D_\theta' D_\lambda = 0$ and $D_\Psi' D_\lambda = 0$ hold by construction within sample. Because bias is exactly zero, one can compute the match effects separately. Note that this does not hold in a regression that additionally includes covariates, since the match effects need not be exactly orthogonal to the firm and individual effects after conditioning on the covariates. The estimates of the regression of $\tilde{y}$ on the two sets of fixed effects are then given by the standard formula:

$$\begin{pmatrix} \hat{\theta}^{ME} \\ \hat{\Psi}^{ME} \end{pmatrix} = \begin{pmatrix} T & K \\ K' & F \end{pmatrix}^{-1} \begin{pmatrix} D_\theta' \tilde{y} \\ D_\Psi' \tilde{y} \end{pmatrix} = \begin{pmatrix} T & K \\ K' & F \end{pmatrix}^{-1} \begin{pmatrix} T\bar{\tilde{y}}_i \\ F\bar{\tilde{y}}_j \end{pmatrix} \tag{9}$$

where $\bar{\tilde{y}}_i$ and $\bar{\tilde{y}}_j$ are vectors of individual and firm means of $\tilde{y}_{it}$ and *T*,*F* and *K* are defined in eq. 5. The second equality results from the fact that pre-multiplying a vector by a transposed matrix of dummies gives a vector of group totals. This is equivalent to multiplying the group mean by the number of observations in the group. To calculate the firm fixed effects (or whichever set of fixed effects is smaller), one only needs the lower blocks of the inverse

of the partitioned matrix. These blocks can be obtained by applying the formula for the inverse of a partitioned matrix (see Theil 1971 section 1.2 for the formula and proof), yielding:

$$\begin{aligned}\hat{\Psi}^{ME} &= \left[-(F-K'T^{-1}K)^{-1}K'T^{-1} \quad (F-K'T^{-1}K)^{-1}\right]\begin{pmatrix} T\bar{\tilde{y}}_i \\ F\bar{\tilde{y}}_j \end{pmatrix} \\ &= (F-K'T^{-1}K)^{-1}F\bar{\tilde{y}}_j - (F-K'T^{-1}K)^{-1}K\bar{\tilde{y}}_i \\ &= (F-K'T^{-1}K)^{-1}(F\bar{\tilde{y}}_j - K\bar{\tilde{y}}_i)\end{aligned} \tag{10}$$

As was pointed out above, *T* and *F* are diagonal matrices and *K* has at most *S* non-zero elements, which makes it sparse in most cases. All three contain only integers. Consequently, the matrices and vectors involved in this expression are very simple to obtain and can be stored efficiently. The only computational difficulty involved in estimating the firm fixed effects is caused by the matrix in the first brackets. However, it is of size *J*×*J*, so it is already a lot smaller than the matrix that would need to be inverted for standard OLS. Additionally, it is symmetric positive definite, so its Cholesky factorization exists and can be used to calculate the matrix product, which is computationally much simpler than obtaining its inverse.[2] In case there are multiple connected groups, an additional advantage is that the system of equations given by the normal equations can be separately solved by group. In terms of matrices, this means that $(F-K'T^{-1}K)$ becomes block diagonal with one block for every connected group. Consequently, its inverse/Cholesky factorization can be computed group by group. As the order of computational complexity increases by the square of the size of the matrix, this can make the problem considerably easier. This is not possible without subtracting $X\hat{\beta}$ to get rid of *X*, because including *X* adds *K* full rows and columns to the matrix, thereby destroying the block-diagonal structure. In extremely large applications, it may be infeasible to work even with this reduced matrix. In such cases, one could still use algorithms such as the conjugate gradient algorithm to solve the now much smaller system of equations implied by equation

[2] In some cases, there may be better ways to solve the system than by using the Cholesky factorization or it may be numerically instable. Most programs have internal algorithms that automatically determine the best way to solve this equation.

10. Abowd et al (2008), among others, use the CGA to solve the full normal equations, the adaptation to equation 10 is straightforward and implemented in a Stata program that will be discussed further in section 3.4.

The fact that residuals sum to zero for each individual, firm and match implies that the individual fixed effects can be recovered from the individual means. This mean contains the average of the firm effects of the firms for which individual i worked (weighted by the length of the spell). The vector of these averages can be seen to equal $T^{-1}K\hat{\psi}^{ME}$:

$$\begin{aligned} \overline{y}_i &= \overline{X}_i\hat{\beta}^{ME} + \hat{\theta}^{ME} + T^{-1}K\hat{\psi}^{ME} \quad \Leftrightarrow \\ \hat{\theta}^{ME} &= \overline{y}_i - \overline{X}_i\hat{\beta}^{ME} - T^{-1}K\hat{\psi}^{ME} \quad \Leftrightarrow \\ \hat{\theta}^{ME} &= \tilde{\overline{y}}_i - T^{-1}K\hat{\psi}^{ME} \end{aligned} \tag{11}$$

Finally, the match effects can be computed from the match means:

$$\begin{aligned} \overline{y}_s &= \overline{X}_s\hat{\beta}^{ME} + \hat{\theta}_i^{ME} + \hat{\psi}_j{}^{ME} + \hat{\lambda}_s{}^{ME} \quad \Leftrightarrow \\ \hat{\lambda}_s{}^{ME} &= \overline{y}_s - \overline{X}_s\hat{\beta}^{ME} - \hat{\theta}_i^{ME} - \hat{\psi}_j{}^{ME} \quad \Leftrightarrow \\ \hat{\lambda}_s{}^{ME} &= \tilde{\overline{y}}_s - \hat{\theta}_i^{ME} - \hat{\psi}_j{}^{ME} \end{aligned} \tag{12}$$

Where $\overline{y}_s$ and $\tilde{\overline{y}}_s$ are S×1 vectors containing the match means of y_{ijt} and $y_{it} - X_{it}\hat{\beta}^{OLS}$ and $\overline{X}_s$ is the S×K matrix containing the spell means of x_{ijt}. Both calculations are computationally trivial.

3.2 Estimating the TWFE model

In this section, I will discuss how to estimate the TWFE model using a similar procedure as the one outlined above. The main difference is that the partial regression step to obtain estimates of β is more complicated in this case.

The algorithm described in the previous section can be applied to the TWFE model with one caveat: In the TWFE model, the OLS predictions based on the fixed effects are different from the match means, so the estimates $\hat{\beta}$ obtained in the first step are not the same as the OLS estimates. If there are no match effects, the estimates will still be unbiased, but inefficient. In

order to obtain the exact OLS estimates, one needs to run a partial regression as one usually would, i.e. by regressing y and each column of X on the two sets of fixed effects and using the residuals from these regressions to obtain $\hat{\beta}^{TW}$. Implementing this is greatly simplified by the fact that these regressions have the same covariates (the two sets of fixed effects only) as the regression of $\tilde{y}$ on the two sets of fixed effects that is solved by equation 9. Consequently, one can obtain the estimates of these regressions the same way as above by solving equation 10 and 11 and repeating this for each column of *X* in place of *y*. Subtracting these estimates from *y* and *X* gives the "Yulized residuals" (Yule 1907) needed for the partial regression. Wansbeek and Kapteyn (1989) use quite different notation, but it can be shown that this procedure is equivalent to the transformation they propose. Rather than inverting the whole (N+J)×(N+J) matrix, this only requires the Cholesky factorization of $(F - K'T^{-1}K)$ to be computed. As this is the same matrix that is later used to compute the firm fixed effects, it has the same desirable properties. More importantly, its Cholesky decomposition only needs to be computed once and can be stored and reused for all partial regressions and to obtain the estimates of the firm fixed effects from $\tilde{y}$ after $\hat{\beta}^{TW}$ has been calculated. This advantage does not apply when using the conjugate gradient algorithm to solve equation 10, so it has to be repeated for each covariate. In most cases, however, solving equation 10 with the CGA is a matter of seconds.

With the "Yulized residuals" computed, estimation proceeds exactly the same way as in the match effects model: One obtains estimates of $\hat{\beta}^{TW}$ from a partial regression using the "Yulized residuals", subtracts the fitted values from y to obtain $\tilde{y}$ and uses the Cholesky factorization of $(F - K'T^{-1}K)$ from the partial regressions to obtain the firm fixed effects according to equation 10. The individual effects are then given by equation 11.

3.3 Specification Issues

This section will deal with two common specification issues. First, I will discuss conditions under which each of the two models should be used and point out how to test which model is applicable. Second, I will show that the assumption of iid errors that is commonly made with both models can easily be relaxed. Most of the methods described here are commonly known and at most slightly adapted to the setting at stake, so I will only briefly review them.

In theory, the decision between the two models is straightforward: if employers and employees are heterogeneous, but there are no idiosyncratic effects arising from special combinations of workers and firms, the TWFE is sufficient. If, on the other hand, there are combinations of workers and firms that produce outcomes that are systematically different from what one would expect based on the firm and worker fixed effect alone (i.e. there are “better” and “worse” combinations), the match effect model should be used. However, even if theory may provide some guidance in certain cases, it seems more desirable to have a formal framework of model selection.

The first thing to note here is that the TWFE is a special case of the match effects model that restricts all match effects to equal zero. Consequently, the TWFE is a restricted match effects model and will be biased if the imposed restrictions are violated (see Woodcock 2008 for derivations of the biases). Thus, choosing between the TWFE and the match effects model is a choice between a more general and a restricted model just as whether a variable should be included in a regression or not. A desirable strategy for model selection would thus start with the more general model and test whether the restrictions are valid using one of the standard model selection tests. If the restrictions are rejected, one should stick with the more general match effects model, while one may want to take advantage of the more efficient TWFE model if the restrictions are not rejected by the data. This can be tested using the standard F-Test for a restricted vs. an unrestricted model. As the available algorithms usually do not estimate the full covariance matrix, this test should be based on the difference in the sum of squared residuals. Note that the calculation of the degrees of freedom should be

done based on the difference between parameters and restrictions, so the test has S-N-J and N*-N-J degrees of freedom. It is easy to obtain the sum of squared residuals for the match effects model, because the residuals are just the residuals from OLS on deviations from match specific means. Consequently, an F-Test should always be performed when estimating the TWFE model to avoid biased coefficients, particularly because most applications that test for match effects reject the TWFE model (see e.g. Woodcock 2008). On the other hand, if one has estimated the match effects model and the question is whether the TWFE would be more efficient one will unfortunately have to estimate both models to perform a formal test. Similar tests can be used to test either firm or individual fixed effects or both, by simply estimating the corresponding one-way fixed effects model. Note that if one wants to omit one or both of the fixed effects, but keep the match effects in the model, the match effects should be computed with the restriction that they sum to zero for each firm and individual (e.g. by using equation 12 and subtracting the means). Otherwise they will pick up the omitted effects and the residual sum of squares will not change.

It is also often assumed that ε_{ijt} is iid across observations, which is questionable in panel data and can lead to seriously distorted rejection rates even when including fixed effects (Kezdi 2003). While the specific correlation structure of the error obviously depends on the application at hand, the sampling unit in such applications practically never is a single observation, so clustered errors should at least be tested. In the case of worker and firm data, it is likely that errors are correlated both within firms and within individuals, leading to two-way clustered errors. Two-way clustered standard errors for $\hat{\beta}$ can be calculated using the method proposed in Cameron et al. (2006). They generalize the formula for the variance matrix proposed by White (1984) to the case of arbitrary clustering patterns:

$$V(\hat{\beta}) = (X'X)^{-1}\hat{B}(X'X)^{-1} \tag{13}$$

where $\hat{B} = X'(\hat{\varepsilon}\hat{\varepsilon}'.*C)X$. C is a selection matrix with element (k,l) equal to 1 if observation k and l share any cluster and zero otherwise, and .* denotes element-wise multiplication. In the models described in this article, it seems natural to cluster errors by individual and by firm. As the match effect is nested within the two other effects, it does not require additional clustering. Consequently, I will focus on how to allow two-way clustered errors. Extensions to additional levels of clustering (e.g. on time period) are straightforward with the formulas and methods Cameron et al. (2006) provide. Rather, the problem in these models is that the size of the datasets usually prohibits creating the matrix C, which has as many rows and columns as there are observations in the dataset, due to memory constraints. This problem can be avoided by using the formula to calculate $V(\hat{\beta})$ from one-way clustering matrices. In the two-way clustering case, the formula Cameron et al. provide reduces to

$$V(\hat{\beta}) = (X'X)^{-1}\left[\hat{B}^J + \hat{B}^I - \hat{B}^S\right](X'X)^{-1} \tag{14}$$

where $\hat{B}^J$, $\hat{B}^I$ and $\hat{B}^S$ are the $\hat{B}$-matrices obtained from clustering at the firm, individual and match level. Note that these matrices can be computed as the sum of the *B*-matrices for every individual, firm and match, which requires very little memory and can thus easily be parallelized.

3.4 Implementation

One of the main obstacles to implementation is that the size of the matrix that needs to be inverted still is potentially quite large. As was pointed out above, Stata does not work with matrices that have more than 11.000 columns. Matlab does not impose this limit, but one will need to have sufficient main memory to store both the dataset and the Cholesky factorization. Both problems can be circumvented by using the CGA to solve equation 10, but this tends to be slower, particularly for the TWFE. The Matlab program that implements the algorithm above uses the Cholesky factorization while the Stata program uses the CGA.

Table 1 provides some simulations that illustrate the difference between the programs. There are 10 observations per individual and each individual is allowed to move to another firm once. Changing these parameters does not affect the simulation times much. All simulations were carried out on a multicore computer with 24 GB of main memory. Computation time depends a lot on the structure of the data and the computational setup. However, the simulations suggest that the for the match effects model, the CGA is faster if the sample gets large. The number of covariates does not affect estimation speed and memory requirements much. For the TWFE on the other hand, the Cholesky factorization is always faster than the CGA, particularly when there are a lot of covariates. However, memory requirements are much lower with the CGA, which can be a huge advantage in large applications. Because sufficient memory was available, I used the Cholesky factorization for the application below.

4 Application to linked employer-employee data from Germany

In this part, I will apply the models and estimation techniques discussed above to administrative data from Germany. I find that the availability of detailed biographic data has important effects on the inference about how individual characteristics matter and that match effects play an important role. I also examine characteristics that lead to good matches and find that they are more important for the type of firm an individual is matched with than the quality of the match. The next section will discuss how the data is created and address some potential problems, section 4.2 will discuss the models to be estimated, the covariates included and the results.

4.1 The Data

I will estimate the models discussed above using linked employer-employee data from Germany. In particular, I will use the LIAB mover model of the IAB (German Institute for Employment Research), which is created by linking social security records to panel data on firms. The firm data stems from the IAB Betriebspanel (see Fischer et al. 2008 for details), a panel that is based on yearly interviews with the managers of the firms. It dates back to 1993

(1996 for the former GDR) and is a stratified random sample of establishments in Germany. There are 43.617 firms in total and between 4265 and about 16.000 firms per year, with a large part of the variation explained by successive expansion of the panel. It includes weights based on sample and population distributions that adjust for non-response and the non-random sampling. The IAB created a linked employer-employee dataset by matching this data to social security records of the individuals working at the firms in the panel (see Jacobebbinghaus 2008 for a description of the matching process and the individual data).

As the firm data is based on a survey rather than administrative records, the amount of firm specific information is unusually rich. Among others, I have very detailed information on workforce composition, investments, revenue, hiring and firing, training programs, collective bargaining and worker representation as well as information on location, industry and legal form. See the summary statistics in appendix B for details. The individual data, on the other hand, stem from administrative records on social security payments and benefit receipt from the Federal Employment Agency (Bundesagentur für Arbeit), with the main share of benefits being unemployment insurance and employment subsidies. Besides daily histories of earnings and benefits, the data includes information on the person's year of birth, gender, nationality and education. The accuracy of the data is extremely high, partly because social security contributions and benefit rates are based on this information, but also because information such as a person's year of birth and gender can be inferred from the social security number in Germany. The fact that such a long history is available makes it possible to create a number of biographic covariates such as the exact work experience, tenure, year and age at which a person entered the labor market as well as information on job transitions and unemployment.

A downside of the data is that records for the former GDR do not exist prior to 1990, which causes many employment history variables to be inaccurate for people from the former GDR. Consequently, I exclude people from the former GDR that entered the labor market before 1990. Another problem is that income is topcoded at the social security limit for some indi-

viduals, because employers can report the social security limit instead of the actual income if the latter exceeds the former. This limit is different for the states that belonged to the former GDR and varies by year. The exact values can be obtained from the website of the IAB, in 2007, for example, it was € 63.000/year (East: € 54.600). This affects 6.9% of the sample (4.7% after weighting). In order to account for this problem, the analysis below contains a dummy if an observation is topcoded. Additionally, I repeat the analysis excluding individuals with any form of university education. In this restricted sample, only 3.9% of all observations are affected by topcoding (3% after weighting), but the results do not change much. Because the data is derived from social security records, it does not include work that is not subject to social security such as self employment. However, approximately 75-80% (Koch & Meinken 2004) of employment in Germany is subject to Social Security.

The dataset I will use is the LIAB Mover Model 9308, which covers the time period from 1993 to 2008 and was particularly designed to estimate models with individual and firm fixed effects, as it only includes firms that employed at least one worker that also worked for another firm in the panel. That is, firms for which the firm effect is not identified are excluded. After additionally excluding firms with low data quality, there are 25.236 firms in the data that employ movers. The individual data contains all people that move between the firms in the sample (713.559) and up to 500 randomly selected employees of each firm that did not move or moved to firms that are not in the sample. If the firm had less than 500 employees subject to social security, all employees are selected. The data contains a total of 4.666.926 individuals. I adjust the weights from the firm survey to make the resulting sample representative of the population of employees subject to social security. I exclude observations with missing values on continuous covariates, but include a “missing” dummy to retain observations for which categorical variables have missing values, which leaves a sample of 9,891,519 observations for the full sample and 8,822,456 for the low education sample. As was pointed out above, the means of the fixed effects within each connected group are only identified up to a normalization. To ease the interpretation of the analysis of the fixed effects, I follow the

common practice to restrict the sample to observations in the largest connected group, which contains 99% of all observations.

4.2 Analysis

This section will present the models I will estimate, the covariates used and results. Put briefly, I will first estimate models including varying covariates and fixed effects and then analyze these fixed effects using covariates that are constant within a firm, an individual or a match.

One of the main questions of interest concerns the importance of match effects. In order to test whether these effects are relevant, I estimate both the match effects and the corresponding TWFE model that constrains all match effects to be zero. Not only does the data reject this restriction, but the omission of match effect also biases the estimates of the slopes of individual specific characteristics. Additionally, the effect of firm and individual characteristics on wages is of interest. Besides having relatively many firm characteristics, a particular advantage of the data is the availability of accurate information on employment histories. Tenure measured accurately to the day and I have information on past employment, allowing me to construct actual work experience. This is often proxied for by potential experience defined as age-education-6. I will show that using this proxy instead of actual work experience significantly changes the results by estimating the model first with a quartic of actual experience and then re-estimate it using age-education-6 as the measure of experience. Note that the linear term of potential experience is perfectly collinear with the individual fixed effects and the time dummies and thus has to be omitted. The same applies to the linear term of age, so only the second, third and forth order terms are included in both models. Consequently, I estimate a total of 8 models: The match effects model and the TWFE model with actual experience as well as both models with potential experience first using the whole sample and then repeating the analysis on the low-education sample. In a second step, I will analyze the estimates of the individual and firm fixed effects from these models. This analysis shows that the omission of match effects and the lack of biographic variables such as actual experience lead to biased inference in these models. The bias is large for some indi-

vidual specific characteristics, most notably the gender wage gap, but less important for firm characteristics. Finally, I show that pre-match characteristics are important for the kind of firm an individual gets matched with subsequently, but have little impact on the quality of the match.

The first step makes use of the algorithm described above in order to regress the log of person *i*'s daily wage at firm *j* in time period *t* (w_{ijt}) on characteristics of person *i* (x_{ijt}^{I}), and firm j (x_{ijt}^{F}) as well as year, firm, individual and match fixed effects:

$$\log(w_{ijt}) = \theta_i + \Psi_j + \lambda_s + \phi_t + x_{ijt}^{F}\beta^{F} + x_{it}^{I}\beta^{I} + \varepsilon_{ijt} \tag{15}$$

The TWFE model omits λ_s from equation 15. The error term in all regressions is allowed to be arbitrarily correlated within firms and individuals, i.e. I allow for two-way clustering. However, I do not use weights in these regressions, as they increase the computational complexity considerably. Because I condition on all stratification variables, the estimates will be consistent nonetheless. Summary statistics of the firm and individual covariates used in this regression are given in Table A1 in appendix B for the full sample and the sample that restricts the education range to alleviate the topcoding problem. Results for the match effects model with actual and potential experience, the TWFE model with actual experience using the full sample as well as the match effects model with actual experience using the reduced sample are presented in Table 2. Results from the other models are available upon request, but they are similar to the ones presented here and the differences between the reported models illustrate the overall differences well.

Regardless of the model, an F-Test rejects the null hypothesis that match effects do not matter at any conventional level (p-value of 0), so the TWFE is rejected by the data because match effects do explain a relevant part of the overall variation in daily wages. A more important question, however, is whether the exclusion of match effects would lead to substantively wrong conclusions regarding the relation between wages and observed characteristics, i.e. whether the omission of match effects biases coefficients in a relevant way. This can be as-

quite close to the higher order terms of actual experience. However, it seems quite likely that these biases would be worse if no accurate measure of tenure was available.

The algorithm I proposed also yields estimates of the individual, firm and match fixed effects. The correlations of the firm fixed effects across the 8 models exceed 0.9 with the exception of the match effects model with potential experience that produces correlations around 0.6. Overall, the difference tends to be greater between actual and potential experience than between match effects and TWFE model, cautiously suggesting that using potential experience may introduce some bias in the analysis of firm fixed effects while using the TWFE or the match effects model may make little difference here. The differences between models are more pronounced when looking at the correlations between individual fixed effects (ranging from 0.04 to 0.98) with the differences again being larger when using actual instead of potential experience than between the match effects and the TWFE model. Finally, the estimates of the match effects are not affected by the sample restriction, but the correlations are only around 0.7 for models with different measures of experience. Overall, this suggests that estimates of the effect of individual specific characteristics are less robust to model specification than firm characteristics. The TWFE model also suggests that the correlation between individual and firm fixed effects is much less negative than in the match effects model. In the match effects model with actual experience, it is -.37, which suggests substantial negative sorting of high wage individuals into low wage firms. This correlation is known to be biased towards zero, but as Andrews et al. (2008) show, the bias is declining in the number of people that move between firms, so it should be small in this case.

Obtaining estimates of the fixed effects has the advantage that it allows the researcher to assess whether the fixed effects are systematically related to characteristics that do not change over time. In order to examine which permanent characteristics make some firms pay high wages and some individuals receive high wages, I will run the following regressions:

$$\begin{aligned} \hat{\Psi}_j &= z_j^F \delta + \eta_j^F \\ \hat{\theta}_i &= z_i^I \gamma + \eta_i^I \end{aligned} \tag{16}$$

where z_j^F are time invariant characteristics of firm *j* and z_i^I are time invariant characteristics of individual *i*. Interpretation of the coefficients in these regressions as structural parameters requires the error terms in (16) to be uncorrelated with the time invariant characteristics (see e.g. Greene 2011 for a discussion). Summary statistics of the variables included are given in Table A2 and Table A3 in appendix B. Table 3 shows results from the firm fixed effects, Table 4 reports the results for the individual fixed effects. The tables include results using the fixed effects from the four models discussed above, using the fixed effects from the other four models leads to similar conclusions. I adjust the weights from the establishment panel to correct for the sampling design of the individuals, but I do not use GLS to correct for the fact that the estimation error in the fixed effects may be correlated. This will not bias the coefficients, but may affect the standard errors and lead to inefficient estimates (see Hausman and Taylor 1981). However, as long as the unexplained variance in the (true) fixed effects is large relative to the estimation error, this effect will be minimal. It could be solved by bootstrapping the covariance matrix of the fixed effects, but none of the conclusions below would be affected even if the true standard errors were much larger than estimated.

The results in Table 3 confirm the conclusions that using the restricted sample or the TWFE model instead of the match Effects Model leads to very similar conclusions regarding the effects of firm specific characteristics. The differences are more pronounced, but still small when using potential instead of actual experience. The directions of most coefficients are not surprising: collective wage agreements and the existence of a worker's council lead to higher wages, whereas non-monetary benefits such as paid on the job training reduce wages. The regression also includes dummies for the year a company was founded. Because this information is only available since 1990, companies founded before 1990 are pooled in the excluded category. The correlation between these coefficients and GDP growth in the corresponding year is 0.44, suggesting that firms founded in good years pay higher wages. The correlation becomes stronger when looking only at more recent years, which indicates that this effect slowly fades out over time. This finding is robust to de-trending the data.

where z_j^F are time invariant characteristics of firm *j* and z_i^I are time invariant characteristics of individual *i*. Interpretation of the coefficients in these regressions as structural parameters requires the error terms in (17) to be uncorrelated with the time invariant characteristics (see e.g. Greene 2011 for a discussion). Summary statistics of the variables included are given in Table A2 and Table A3 in appendix B. Table 3 shows results from the firm fixed effects, Table 4 reports the results for the individual fixed effects. The tables include results using the fixed effects from the four models discussed above, using the fixed effects from the other four models leads to similar conclusions. I adjust the weights from the establishment panel to correct for the sampling design of the individuals, but I do not use GLS to correct for the fact that the estimation error in the fixed effects may be correlated. This will not bias the coefficients, but may affect the standard errors and lead to inefficient estimates (see Hausman and Taylor 1981). However, as long as the unexplained variance in the (true) fixed effects is large relative to the estimation error, this effect will be minimal. It could be solved by bootstrapping the covariance matrix of the fixed effects, but none of the conclusions below would be affected even if the true standard errors were much larger than estimated.

The results in Table 3 confirm the conclusions that using the restricted sample or the TWFE model instead of the match Effects Model leads to very similar conclusions regarding the effects of firm specific characteristics. The differences are more pronounced, but still small when using potential instead of actual experience. The directions of most coefficients are not surprising: collective wage agreements and the existence of a worker's council lead to higher wages, whereas non-monetary benefits such as paid on the job training reduce wages. The regression also includes dummies for the year a company was founded. Because this information is only available since 1990, companies founded before 1990 are pooled in the excluded category. The correlation between these coefficients and GDP growth in the corresponding year is 0.44, suggesting that firms founded in good years pay higher wages. The correlation becomes stronger when looking only at more recent years, which indicates that this effect slowly fades out over time. This finding is robust to de-trending the data.

Table 4 reports the coefficients from a regression of the individual fixed effects on observed characteristics of the individual. Contrary to the firm characteristics, model specification has a sizeable impact on these coefficients. This is particularly pronounced for the gender wage gap, which drops from more than 23% to less than 6% when replacing potential by actual experience. The likely reason for this is that the amount by which potential experience overstates actual experience is greater for women as they tend to spend more time not being employed. The TWFE also overestimates the gender wage gap by about 5%, while reducing the sample to the less educated has little effect. Such observed gender gaps should be interpreted cautiously, however, as both positive and negative gaps are consistent with discrimination, while on the other hand, gaps of arbitrary size can arise if there are gender-based preferences for factors that are not included in the model, such as occupation or non-monetary benefits. Using potential experience or the TWFE model also distorts the returns to education. Regardless of the model, however, the returns to vocational training are quite high both for people who get the equivalent of a high school degree (upper secondary school) and those with 9-10 years of schooling. Obtaining a high school degree without any further training or education actually seems to impose a wage penalty of around 10%. Note, however, that there are few people in these two categories and they are likely to have attended different types of schools. It seems more likely that the wage penalty stems from a bad choice of school type than from 3 more years of schooling. Most of the nationality dummies are positive, with the exception of China, Morocco and several countries from southern Europe. Because the omitted category is “German” and it seems unlikely that there is positive discrimination, this indicates that immigrants from these countries have unobserved characteristics that increase their productivity. Such a situation may arise from differences in training and schooling, but could also be due to selective migration from these countries. The regression also includes dummies for the year a person first entered the labor market, which raises the question whether there are any long run impacts of entering the labor market when the economy is doing well. However, there is only a very weak relationship with GDP growth.

When de-trending the data, there is a sizeable positive correlation with unemployment. This indicates that the people who enter the labor market in a bad year tend to be high wage earners, but the overall evidence is quite weak.

Finally, I will examine what drives match quality. In order to do so, I will regress the match effects and the corresponding firm effect on several match specific characteristics:

$$\begin{aligned} \hat{\lambda}_s &= z_s^M \pi_1 + \eta_s^M \\ \hat{\Psi}_s &= z_s^M \pi_2 + \upsilon_s^M \end{aligned} \tag{17}$$

where $\hat{\Psi}_s$ is the estimated firm fixed effect of the firm that match *s* corresponds to and z_s^M is a vector of characteristics that are invariant within match *s*. So the first regression provides information on the characteristics that lead to a good fit between the individual and the firm, whereas the second regression examines whether certain characteristics of the matching process can account for employees being matched with high-wage or low-wage firms. Summary statistics of the covariates used are provided in Table A4 in appendix B, results from the match effects model with actual experience are presented in Table 5.

Both the low R-squared and the small coefficients in the first column indicate that the relationship between match quality and observable pre-match characteristics is rather weak. Employees who start part time tend to be matched worse and match quality tends to decline in later matches (although it increases after the 4th match, but there are very few observations in these categories). This is consistent with the biases found in the main regression when omitting the match effects, but it is noteworthy that the negative partial correlation between match effects and tenure and age arises from the match count and not from age at the beginning of the match. While is seems reasonable that part time workers tend to be matched worse (e.g. because there are search costs), declining match quality is somewhat surprising. A potential explanation could be that switching jobs is regarded as a bad signal in a labor market with low turnover such as the German labor market. There is no evidence that labor force status before the current match or the amount of time one spent in it affects the

size of the subsequent match effect. While a joint F-Test on the interactions with female suggests that the matching process is different for females, the differences seem to be small in practical terms.

Pre-Match characteristics are much more predictive of the size of the subsequent firm effect. Job-to-job transitions lead to employment at firms that pay higher wages than employment after unemployment or training, with the difference increasing in the duration of the previous spell. The sorting process is quite different by gender, with females sorting into lower wage firms. As was pointed out above, this is consistent with discrimination, but could also arise from gender based preferences without discrimination. Overall, people tend to move to higher wage firms over the life-cycle, which is indicated by the positive coefficients on age and the match count dummies. This effect is more pronounced for females. Several papers (e.g. Topel and Ward 1992 for the US) have argued that the main increases in wages happen by switching jobs rather than by raises on the job. My results suggest that this effect is mainly due to switching from low-wage to higher wage firms, rather than by obtaining a better match between employer and employee.

Summarizing the main findings, the TWFE is rejected by the data in favor of the match effects model and leads to qualitatively important biases to the effects of individual characteristics. Availability of accurate biographic information such as actual experience and tenure also has an important impact on the estimates of these effects. The effects of firm specific characteristics seem to be more robust. While match effects matter in the main regression, very little of their variance is explained by pre-match characteristics. On the other hand, such characteristics are systematically related to the kind of firm an individual is matched with.

5 Conclusion

In this paper, I have proposed a new algorithm to estimate the TWFE and the match effects model in large applications. These algorithms not only offer advantages in terms of speed and computational resources needed, but also allow estimation of the fixed effects and multi-way clustered standard errors of the slopes. An application to matched employer-employee data from Germany shows that it is important to at least consider the inclusion of match effects and that estimates of the fixed effects can yield insights into the determination of wages and the matching process at the labor market. The unusually rich dataset also shows that, even when including multiple fixed effects, inference can be quite misleading if accurate information on important covariates such as labor market experience is not available.

References

Abowd, J.M. and F. Kramarz 1999, The Analysis of Labor Markets Using Matched Employer-Employee Data, in: O. Ashenfelter and D. Card (Eds.), Handbook of Labor Economics, Volume 3(B) Chapter 26. North Holland, Amsterdam, pp. 2629-2710.

Abowd, J. M., F. Kramarz, and D. N. Margolis, 1999, High wage workers and high wage firms. Econometrica 67 (2), 251-334.

Abowd, J. M., R. H. Creecy, and F. Kramarz 2002, Computing person and firm effects using linked longitudinal employer-employee data. Census Bureau Technical Paper TP-2002-06.

Abowd, J.M., F. Kramarz and S.D. Woodcock 2008, Econometric Analyses of Linked Employer-Employee Data, in: L. Mátyás, P. Sevestre (Eds.), The Econometrics of Panel Data, Chapter 22. Springer, Heidelberg, pp. 727-759.

Andrews, M.J., T. Shank and R. Upward 2005, Practical Fixed Effects Methods for the Three-Way Error Components Model. Mimeo.

Andrews, M.J., L. Gill, T. Shank and R. Upward 2008, High wage workers and low wage firms: negative assortative matching or limited mobility bias? Journal of the Royal Statistical Society A 171 (3), 673-697.

Bennett, D. 2010, Health Care Competition and Antibiotic Use in Taiwan. Mimeo.

Cameron, C. A., J. B. Gelbach and D.L. Miller 2006, Robust Inference With Multi-Way Clustering. Mimeo.

Cameron, C. A., J. B. Gelbach and D.L. Miller 2008, Bootstrap-based Improvements for inference with Clustered Errors. The Review of Economics and Statistics 90 (3), 414-427.

Davis, P. 2002, Estimating multi-way error components models with unbalanced data structures. Journal of Econometrics 106, 67-95.

Fischer, G., F. Janik, D. Müller and A. Schmucker 2008, The IAB establishment panel - from sample to survey to projection. FDZ Methodenreport 01/2008 (en).

Greene, W.H. 2008, Econometric Analysis 6th Edition. Prentice Hall, New York.

Greene, W.H. 2011, Fixed Effect Vector Decomposition: A Magical Solution to the Problem of Time Invariant Variables in Fixed Effects Models. *Political Analysis* 19(2), 135-146.

Grogger, J. and G.H. Hanson 2011, Income maximization and the selection and sorting of international migrants. Journal of Development Economics 95, 42-57.

Hausman, J.A. 1978, Specification Tests in Econometrics. Econometrica 46(6), 1251-1271.

Hausman, J.A. and W.E. Taylor 1981, Panel Data and Unobservable Individual Effects. Econometrica 49(6), 1377-1398.

Jackson, K. 2011, Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers. NBER Working Paper No. 15990.

Jacobebbinghaus, P. 2008, LIAB-Datenhandbuch, Version 3.0. FDZ Datenreport, 03/2008 (de).

Jovanovic, B. 1979, Firm-specific Capital and Turnover. Journal of Political Economy 87(6), 1246-1260.

Kezdi, G. 2003, Robust Standard Error Estimation in Fixed-Effect Panel Models. Mimeo.

Koch, I. and H. Meinken 2004, The Employment Panel of the German Federal Employment Agency. Journal of Applied Social Science Studies 124(2), 315-325.

Kramarz, F., S. Machin, A. Ouazad 2008, What Makes a Test Score? The Respective Contributions of Pupils, Schools, and Peers in Achievement in English Primary Education. IZA Discussion Paper.

Lovell, M.C. 1963, Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association 58, 993-1010.

Mortensen, D.T. 1978, Specific Capital and Turnover. The Bell Journal of Economics 9(2), 572-586.

Theil, H. 1971, Principles of Econometrics. John Wiley and Sons, New York.

Topel, R.H. and M.P. Ward 1992, Job Mobility and the Careers of Young Men. The Quarterly Journal of Economics 107(2), 439-479.

Wansbeek, T. and A. Kapteyn 1989, Estimation of the Error-Components Model with Incomplete Panels. Journal of Econometrics 41, 341-361.

White, H. 1984, Asymptotic Theory for Econometricians. Academic Press, San Diego.

Woodcock, S.D, 2007, Wage Differentials in the presence of unobserved worker, firm and match heterogeneity. Labor Economics 15, 772-794.

Woodcock, S.D. 2008, Match Effects. Mimeo.

Yule, G.U. 1907, On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character 79 (529), 182-193.

Appendix A: Summary of the computational steps in the algorithm for the case of multiple groups

1. Identify connected groups. Abowd et al. (2002) describe this algorithm. This is not necessary when using the CGA.
2. Calculate individual, firm and spell means.
3. Estimate the slope coefficients:
   - For the TWFE-model, do the WK transformation for y and X using the Cholesky factorization of $\left[F - K'T^{-1}K\right]$, which can be calculated separately by group and stored for later use or by applying the CGA to every variable. Regress the transformed y on the transformed X to obtain the slopes.
   - For the match effects model, run OLS on the deviations of y and X from spell means
4. Calculate SEs of the slopes and required test statistics based on the residuals from step 3.
5. Obtain $\bar{\tilde{y}}_{i.}, \bar{\tilde{y}}_{.j}$ and $\bar{\tilde{y}}_{s.}$ by subtracting $\bar{X}\hat{\beta}^{OLS}$ (where the mean of X is taken over the appropriate index) from the individual, firm and spell means from step 2.
6. Calculate the firm fixed effects for each connected group separately using formula 10. This can be done by using the Cholesky factorization of $\left[F - K'T^{-1}K\right]$ (in case of the TWFE-model, it has already been calculated for step 3) or the CGA.
7. Use the firm effects to calculate the individual and match effects.

Appendix B: Summary Statistics

Table A1: Summary Statistics for Main Regression

	Full Sample		Low Education Sample	
	Mean	**SD**	**Mean**	**SD**
Daily Wage	4.12	0.77	4.08	0.77
Total number of employees	326.48	1309.04	308.89	1237.32
Business volume, categorical				
0 to 72,000	0.22%	4.69%	0.21%	4.58%
72,000 to 120,000	0.24%	4.89%	0.24%	4.89%
120,000 to 166,200	0.34%	5.82%	0.35%	5.91%
166,200 to 245,400	0.72%	8.45%	0.75%	8.63%
245,400 to 332,300	1.30%	11.33%	1.34%	11.50%
332,300 to 490,000	1.93%	13.76%	1.97%	13.90%
490,000 to 715,800	2.78%	16.44%	2.84%	16.61%
715,800 to 1,227,100	5.18%	22.16%	5.29%	22.38%
1,227,100 to 3,163,900	11.28%	31.63%	11.48%	31.88%
more than 3,163,900	57.85%	49.38%	57.67%	49.41%
Missing	18.16%	38.55%	17.87%	38.31%
Business volume per employee, categorical				
0 to 21,300	3.31%	17.89%	3.40%	18.12%
21,300 to 30,700	3.67%	18.80%	3.77%	19.05%
30,700 to 39,800	4.36%	20.42%	4.39%	20.49%
39,800 to 50,000	6.13%	23.99%	6.09%	23.91%
50,000 to 59,700	4.96%	21.71%	4.93%	21.65%
59,700 to 71,600	5.47%	22.74%	5.46%	22.72%
71,600 to 92,900	8.10%	27.28%	8.16%	27.38%
92,900 to 128,600	11.67%	32.11%	11.73%	32.18%
128,600 to 230,100	16.79%	37.38%	16.86%	37.44%
more than 230,100	17.36%	37.88%	17.34%	37.86%
Missing	18.16%	38.55%	17.87%	38.31%
Fraction of employees working part time	18.10%	21.24%	18.12%	21.38%
Investment per employee, categorical				
0 to 500	19.09%	39.30%	19.37%	39.52%
2,100 to 3,000	6.88%	25.31%	6.94%	25.41%
3,000 to 4,200	12.00%	32.50%	11.99%	32.48%
4,200 to 6,100	12.54%	33.12%	12.47%	33.04%
6,100 to 10,000	13.35%	34.01%	13.33%	33.99%
10,000 to 18,600	13.78%	34.47%	13.74%	34.43%
more than 18,600	15.53%	36.22%	15.45%	36.14%
Missing	6.82%	25.21%	6.71%	25.02%
Fraction of female employees	40.26%	28.05%	40.19%	28.33%
DHS employment growth index	0.02	0.17	0.02	0.17
More employees than previous year	60.89%	48.80%	60.79%	48.82%
Wanted to hire people, but did not				
successfully hired or did not want to	86.77%	33.88%	86.62%	34.04%
Wanted to hire people, but did not	8.64%	28.10%	8.73%	28.23%
Missing	4.59%	20.93%	4.65%	21.06%
Expected business volume relative to last year				
Same	41.54%	49.28%	41.52%	49.28%
Increasing	31.70%	46.53%	31.50%	46.45%
Decreasing	19.86%	39.89%	20.13%	40.10%
Missing	6.90%	25.35%	6.85%	25.26%
Total number of new employees	12.32	47.25	11.52	43.25
Firm was hiring in current year	69.99%	45.83%	69.46%	46.06%
Total number of employees that left	14.07	61.99	13.26	58.22
Employees have left in current year	75.14%	43.22%	74.83%	43.40%
Number of days in current establishment	3034.97	2688.82	3075.94	2708.80
Age at end of year	40.40	11.65	40.35	11.78
Part time job	20.82%	40.60%	21.30%	40.94%
Daily wage topcoded	4.69%	21.14%	3.01%	17.09%

Year				
1993	6.25%	24.21%	6.35%	24.39%
1994	6.25%	24.21%	6.36%	24.40%
1995	6.25%	24.21%	6.35%	24.39%
1996	6.25%	24.21%	6.36%	24.40%
1997	6.25%	24.21%	6.34%	24.37%
1998	6.25%	24.21%	6.27%	24.24%
1999	6.25%	24.21%	6.28%	24.26%
2000	6.25%	24.21%	6.26%	24.22%
2001	6.25%	24.21%	6.22%	24.15%
2002	6.25%	24.21%	6.23%	24.17%
2003	6.25%	24.21%	6.23%	24.17%
2004	6.25%	24.21%	6.16%	24.04%
2005	6.25%	24.21%	6.16%	24.04%
2006	6.25%	24.21%	6.14%	24.01%
2007	6.25%	24.21%	6.13%	23.99%
2008	6.25%	24.21%	6.14%	24.01%
Potential Experience in years	23.75	11.83	24.20	11.86
Actual Experience in years	14.11	8.41	14.24	8.44

Note: Weighted statistics calculated from the IAB LIAB MM 9308.

Table A2: Summary Statistics for Time Invariant Firm Characteristics

	Mean	SD		Mean	SD
Industry			*Legal Form*		
Agriculture and Forestry	1.57%	12.43%	Individually-owned firm	34.30%	47.47%
Mining, quarrying and electricity	0.32%	5.65%	Partnership	7.63%	26.55%
Food products	2.29%	14.96%	Limited liability company	42.01%	49.36%
Clothing and Textile	0.72%	8.45%	Company limited by shares	2.72%	16.27%
Paper and Printing	1.08%	10.34%	Public corporation	6.28%	24.26%
Wood Products, Furniture, Jewelry, Toys	1.53%	12.27%	other legal form	5.28%	22.36%
Chemical Industry	0.36%	5.99%	missing/don't know	1.78%	13.22%
Rubber/Plastic	0.57%	7.53%	*Main/Exclusive Ownership*		
Non-metallic Mineral Products	0.63%	7.91%	Eastern German property	10.99%	31.28%
Basic Metals, Steel, light Metal	3.88%	19.31%	Western German property	51.87%	49.97%
Recycling	/	/	Foreign property	2.33%	15.09%
Machinery	2.23%	14.77%	Public property	2.65%	16.06%
Motor vehicles: production/sales/repair/fuel	5.01%	21.82%	no principal shareholder	2.86%	16.67%
Other Transport Equipment	/	/	unknown	2.25%	14.83%
Electrical Equipment	1.61%	12.59%	Missing	27.05%	44.42%
Precision and Optical Equipment	1.02%	10.05%	*Year founded (only after 1990)*		
Main Construction Trade	5.52%	22.84%	Founded before 1990	40.65%	49.12%
Building Installation/Completion	5.43%	22.66%	1990	2.63%	16.00%
Sales: Retail and Wholesale	16.60%	37.21%	1991	3.38%	18.07%
Transportation	5.79%	23.36%	1992	2.34%	15.12%
Communication	0.27%	5.19%	1993	2.50%	15.61%
Credit and Financial Intermediation	1.20%	10.89%	1994	2.84%	16.61%
Insurance	0.98%	9.85%	1995	2.89%	16.75%
Computer and Related Activities	1.42%	11.83%	1996	2.02%	14.07%
Research and Development	0.41%	6.39%	1997	2.17%	14.57%
Legal Consulting, Advertising	4.60%	20.95%	1998	2.12%	14.41%
Real Estate	1.78%	13.22%	1999	2.10%	14.34%
Renting, Business Activities	5.82%	23.41%	2000	1.84%	13.44%
Hotel and Restaurant	6.27%	24.24%	2001	1.39%	11.71%
Education/Teaching	2.26%	14.86%	2002	1.27%	11.20%
Human Health, Veterinary and Social Work	9.47%	29.28%	2003	1.40%	11.75%
Sanitation	0.53%	7.26%	2004	1.24%	11.07%
Recreation, Culture, Sports	1.51%	12.20%	2005	1.19%	10.84%
Other Services	2.64%	16.03%	2006	1.01%	10.00%
Organizations, Lobbying	2.31%	15.02%	2007	/	/
Public Administration and Social Security	2.14%	14.47%	2008	/	/
State			Missing	24.43%	42.97%
Schleswig-Holstein	4.85%	21.48%	*Establishment/Department is...*		
Hamburg	3.22%	17.65%	independent company w/o other places of business	75.65%	42.92%
Lower Saxony	10.51%	30.67%	business/office/branch	15.56%	36.25%
Bremen	1.44%	11.91%	head office	5.35%	22.50%
North Rhine-Westphalia	17.77%	38.23%	middle-level authority	1.59%	12.51%
Hesse	8.04%	27.19%	Missing	1.85%	13.48%
Rhineland-Palatinate	5.63%	23.05%	*Company pays for job training/courses*		
Baden-Wuerttemberg	13.24%	33.89%	No	39.80%	48.95%
Bavaria	15.16%	35.86%	Yes	51.61%	49.97%
Saarland	2.09%	14.30%	Missing	8.59%	28.02%
Berlin	3.91%	19.38%	*Has Worker's Council*		
Brandenburg	2.69%	16.18%	No	79.98%	40.01%
Mecklenburg-Western Pomerania	2.19%	14.64%	Yes	18.37%	38.72%

Saxony	4.29%	20.26%	Missing	1.65%	12.74%
Saxony-Anhalt	2.50%	15.61%	*Collective Wage Agreement*		
Thuringia	2.46%	15.49%	Industry-wide wage agreement	43.05%	49.51%
Owner working in Company			company agreement	4.56%	20.86%
No	/	/	no collective agreement	44.65%	49.71%
Yes	73.81%	43.97%	Missing	7.74%	26.72%
Missing	/	/			

Note: Weighted statistics calculated from the IAB LIAB MM 9308. If one or more cells contained too few observations, at least two cell frequencies could not be disclosed (to prevent calculation from totals). This is indicated by /.

Table A3: Summary Statistics for Time Invariant Individual Characteristics

	Full Sample		Low Educ. Sample	
	Mean	SD	Mean	SD
Female	42.72%	49.47%	43.50%	49.58%
Nationality, grouped				
Germany	92.17%	26.86%	91.95%	27.21%
Turkey	2.25%	14.83%	2.39%	15.27%
Italy	0.89%	9.39%	0.93%	9.60%
Yugoslavia, Serbia and Montenegro	0.83%	9.07%	0.88%	9.34%
Greece	0.37%	6.07%	0.39%	6.23%
France	0.26%	5.09%	0.25%	4.99%
Poland	0.27%	5.19%	0.27%	5.19%
Austria	0.29%	5.38%	0.28%	5.28%
Croatia	0.19%	4.35%	0.20%	4.47%
Portugal	0.21%	4.58%	0.22%	4.69%
Spain	0.16%	4.00%	0.16%	4.00%
Netherlands, Luxembourg	0.14%	3.74%	0.13%	3.60%
Russia, Belarus, Former Soviet Union	0.11%	3.31%	0.10%	3.16%
Bosnia and Herzegovina	0.12%	3.46%	0.12%	3.46%
Great Britain, Ireland and Northern Ireland	0.14%	3.74%	0.13%	3.60%
Romania	0.08%	2.83%	0.08%	2.83%
Czech Republic, Slovakia, Former Czechoslovakia	0.09%	3.00%	0.09%	3.00%
Ukraine, Moldova	0.06%	2.45%	0.06%	2.45%
Hungary	0.06%	2.45%	0.06%	2.45%
Albania	0.04%	2.00%	0.04%	2.00%
Belgium	0.03%	1.73%	0.03%	1.73%
Macedonia	0.03%	1.73%	0.03%	1.73%
Switzerland	0.03%	1.73%	0.03%	1.73%
Bulgaria	0.03%	1.73%	0.03%	1.73%
Slovenia	0.02%	1.41%	0.02%	1.41%
Denmark, Sweden	0.04%	2.00%	0.03%	1.73%
Finland	0.02%	1.41%	0.01%	1.00%
Estonia, Latvia, Lithuania	0.02%	1.41%	0.02%	1.41%
Europe (other)	0.02%	1.41%	0.02%	1.41%
Ethiopia	0.01%	1.00%	0.01%	1.00%
Ghana	0.03%	1.73%	0.03%	1.73%
Morocco	0.08%	2.83%	0.09%	3.00%
Tunisia	0.04%	2.00%	0.04%	2.00%
Africa (other)	0.12%	3.46%	0.12%	3.46%
USA, Canada	0.10%	3.16%	0.09%	3.00%
America (other)	0.06%	2.45%	0.06%	2.45%
Afghanistan	0.03%	1.73%	0.03%	1.73%
Sri Lanka	0.05%	2.24%	0.05%	2.24%
Vietnam	0.04%	2.00%	0.04%	2.00%
India	0.03%	1.73%	0.03%	1.73%
Iraq	0.04%	2.00%	0.04%	2.00%
Iran	0.06%	2.45%	0.06%	2.45%
Lebanon	0.03%	1.73%	0.03%	1.73%
Philippines	0.03%	1.73%	0.03%	1.73%
Thailand	0.03%	1.73%	0.03%	1.73%
China, incl. Tibet	0.03%	1.73%	0.03%	1.73%
Asia (other)	0.18%	4.24%	0.17%	4.12%
Oceania	0.01%	1.00%	0.01%	1.00%
Missing	0.04%	2.00%	0.04%	2.00%
School education and vocational training				
Secondary school w/o completed vocational training	15.52%	36.21%	16.65%	37.25%
Secondary school with completed vocational training	58.22%	49.32%	62.47%	48.42%
Upper secondary school (general/subject-specific aptitude for higher education) w/o completed vocational training	1.66%	12.78%	1.78%	13.22%
Upper secondary school (general/subject-specific aptitude for higher education) with completed vocational training	3.79%	19.10%	4.06%	19.74%

Completion of a university of applied sciences	2.78%	16.44%		
College / university degree	4.02%	19.64%		
Missing	14.01%	34.71%	15.04%	35.75%
Year of first employment				
1975 or earlier	23.78%	42.57%	24.58%	43.06%
1976	3.13%	17.41%	3.15%	17.47%
1977	2.53%	15.70%	2.49%	15.58%
1978	2.41%	15.34%	2.37%	15.21%
1979	2.56%	15.79%	2.50%	15.61%
1980	2.58%	15.85%	2.54%	15.73%
1981	2.45%	15.46%	2.41%	15.34%
1982	2.26%	14.86%	2.22%	14.73%
1983	2.29%	14.96%	2.25%	14.83%
1984	2.52%	15.67%	2.46%	15.49%
1985	2.59%	15.88%	2.52%	15.67%
1986	2.84%	16.61%	2.76%	16.38%
1987	2.84%	16.61%	2.76%	16.38%
1988	3.03%	17.14%	2.93%	16.86%
1989	3.66%	18.78%	3.58%	18.58%
1990	4.88%	21.54%	4.80%	21.38%
1991	2.11%	14.37%	2.18%	14.60%
1992	1.78%	13.22%	1.84%	13.44%
1993	2.26%	14.86%	2.26%	14.86%
1994	2.17%	14.57%	2.17%	14.57%
1995	2.22%	14.73%	2.23%	14.77%
1996	2.07%	14.24%	2.05%	14.17%
1997	2.19%	14.64%	2.12%	14.41%
1998	2.18%	14.60%	2.09%	14.30%
1999	5.19%	22.18%	5.14%	22.08%
2000	2.87%	16.70%	2.85%	16.64%
2001	2.19%	14.64%	2.20%	14.67%
2002	1.61%	12.59%	1.63%	12.66%
2003	1.28%	11.24%	1.31%	11.37%
2004	0.99%	9.90%	1.01%	10.00%
2005	0.84%	9.13%	0.86%	9.23%
2006	0.78%	8.80%	0.80%	8.91%
2007	0.64%	7.97%	0.65%	8.04%
2008	0.30%	5.47%	0.30%	5.47%
Age at first employment	24.09	7.64	23.98	7.72

Note: Weighted statistics calculated from the IAB LIAB MM 9308.

Table A4: Summary Statistics for Time Invariant Match Characteristics

	Full Sample		Low Educ. Sample	
	Mean	SD	Mean	SD
Part time job (at beginning of match)	22.98%	42.07%	23.60%	42.46%
Employment Status 8 days before current match				
No Previous Record	17.75%	38.21%	18.18%	38.57%
Unknown, previous spell not benefits	16.12%	36.77%	15.70%	36.38%
Unknown, previous spell was benefit spell	3.93%	19.43%	3.96%	19.50%
Benefit Receipt	19.77%	39.83%	20.26%	40.19%
Employment at other Firm	40.22%	49.03%	39.60%	48.91%
Apprentice/Trainee at other Firm	2.19%	14.64%	2.28%	14.93%
Missing	0.02%	1.41%	0.02%	1.41%
Employment Status 8 days before current match, condensed				
No Previous Record	17.75%	38.21%	18.18%	38.57%
Benefits/Gap	39.82%	48.95%	39.92%	48.97%
Employment at other Firm	40.22%	49.03%	39.60%	48.91%
Apprentice/Trainee at other Firm	2.19%	14.64%	2.28%	14.93%
Missing	0.02%	1.41%	0.02%	1.41%
Number of Days in Emp. Status 8 Days before current match	1117.21	1635.52	1109.38	1650.03
Year Match Started				
1993	18.35%	38.71%	18.69%	38.98%
1994	5.15%	22.10%	5.21%	22.22%
1995	3.11%	17.36%	3.12%	17.39%
1996	3.59%	18.60%	3.61%	18.65%
1997	2.56%	15.79%	2.57%	15.82%
1998	5.21%	22.22%	5.20%	22.20%
1999	5.95%	23.66%	5.98%	23.71%
2000	9.14%	28.82%	9.23%	28.94%
2001	7.41%	26.19%	7.36%	26.11%
2002	6.06%	23.86%	6.02%	23.79%
2003	5.26%	22.32%	5.23%	22.26%
2004	5.59%	22.97%	5.50%	22.80%
2005	5.47%	22.74%	5.40%	22.60%
2006	5.47%	22.74%	5.34%	22.48%
2007	5.66%	23.11%	5.60%	22.99%
2008	6.02%	23.79%	5.95%	23.66%
Days of Benefit Receipt up to beginning of current match	234.37	488.28	243.15	498.10
Days since first Employment at beginning of current match	3127.01	3112.95	3117.67	3140.79
Age at beginning of current match	37.43	12.11	37.36	12.27
Match Count				
1	93.31%	24.98%	93.93%	23.88%
2	6.36%	24.40%	5.80%	23.37%
3	0.32%	5.65%	0.26%	5.09%
4	0.01%	1.00%	0.01%	1.00%
5	/	/	/	/
6	/	/	/	/
Female	42.24%	49.39%	43.12%	49.52%

Note: Weighted statistics calculated from the IAB LIAB MM 9308. If one or more cells contained too few observations, at least two cell frequencies could not be disclosed (to prevent calculation from totals). This is indicated by /.

Appendix C: Results

Table A5: Regressions of log of Daily Wage on Time Variant Characteristics, All Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Total number of employees	**0.000004***	**0.000004****	**0.000004***	**0.000003***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Business volume, categorical				
72,000 to 120,000	**0.0196**	**0.0204**	**0.0165**	**0.0042**
	(0.0158)	(0.017)	(0.0162)	(0.0115)
120,000 to 166,200	**0.0221**	**0.0288**	**0.0207**	**0.0049**
	(0.0157)	(0.0171)	(0.0155)	(0.0127)
166,200 to 245,400	**-0.0036**	**-0.0002**	**-0.0077**	**-0.0277***
	(0.0147)	(0.0162)	(0.0148)	(0.0108)
245,400 to 332,300	**0.0063**	**0.0094**	**0.0029**	**-0.0175**
	(0.0148)	(0.0163)	(0.0147)	(0.0107)
332,300 to 490,000	**-0.0023**	**0.0009**	**-0.0075**	**-0.0301****
	(0.0149)	(0.0163)	(0.0148)	(0.0107)
490,000 to 715,800	**0.0053**	**0.0091**	**0.0036**	**-0.0158**
	(0.0141)	(0.0156)	(0.0141)	(0.0093)
715,800 to 1,227,100	**0.0116**	**0.0141**	**0.0099**	**-0.0083**
	(0.014)	(0.0155)	(0.0139)	(0.0091)
1,227,100 to 3,163,900	**0.0131**	**0.0149**	**0.0115**	**-0.0067**
	(0.0139)	(0.0154)	(0.0138)	(0.009)
more than 3,163,900	**0.0124**	**0.0116**	**0.0127**	**-0.0053**
	(0.0139)	(0.0154)	(0.0139)	(0.0089)
Missing	**0.019**	**0.0235**	**0.0173**	**-0.0033**
	(0.0138)	(0.0152)	(0.0137)	(0.0087)
Business volume per employee, categorical				
21,300 to 30,700	**0.0054***	**0.0055***	**0.0049***	**0.0067****
	(0.0022)	(0.0026)	(0.0021)	(0.0021)
30,700 to 39,800	**0.0026**	**0.0024**	**0.0022**	**0.0041**
	(0.0023)	(0.0028)	(0.0022)	(0.0023)
39,800 to 50,000	**0.0038**	**0.0051**	**0.0036**	**0.0052***
	(0.0024)	(0.0028)	(0.0023)	(0.0022)
50,000 to 59,700	**0.0016**	**0.0004**	**0.001**	**0.0006**
	(0.0024)	(0.0029)	(0.0023)	(0.0023)
59,700 to 71,600	**0.0019**	**0.0021**	**0.0012**	**0.0009**
	(0.0025)	(0.003)	(0.0024)	(0.0024)
71,600 to 92,900	**-0.0011**	**0.0011**	**-0.0022**	**-0.0041**
	(0.0025)	(0.0029)	(0.0024)	(0.0024)
92,900 to 128,600	**0.0006**	**0.0031**	**-0.0009**	**-0.0029**
	(0.0024)	(0.0029)	(0.0023)	(0.0023)
128,600 to 230,100	**0.0027**	**0.0059***	**0.0011**	**-0.0008**
	(0.0024)	(0.0029)	(0.0023)	(0.0023)
more than 230,100	**0.0105*****	**0.0143*****	**0.0089*****	**0.007****
	(0.0024)	(0.0029)	(0.0023)	(0.0023)
Fraction of employees working part time	**-0.0034**	**-0.0011**	**-0.0041**	**-0.0029**
	(0.004)	(0.0044)	(0.004)	(0.0041)
Investment per employee, categorical				
2,100 to 3,000	**0.002**	**0.0016**	**0.0017**	**0.001**
	(0.0012)	(0.0013)	(0.0012)	(0.0013)
3,000 to 4,200	**0.0017**	**0.0025****	**0.0017**	**0.0007**
	(0.0009)	(0.0009)	(0.0009)	(0.0009)
4,200 to 6,100	**0.0031*****	**0.0038*****	**0.0033*****	**0.0027****
	(0.001)	(0.001)	(0.001)	(0.001)
6,100 to 10,000	**0.0041*****	**0.0048*****	**0.0039*****	**0.0034*****
	(0.0009)	(0.0009)	(0.0009)	(0.0009)

10,000 to 18,600	**0.0057*** **	**0.0066*** **	**0.0057*** **	**0.0048*** **
	(0.001)	(0.001)	(0.001)	(0.001)
more than 18,600	**0.0058*** **	**0.0068*** **	**0.0059*** **	**0.0055*** **
	(0.0009)	(0.001)	(0.0009)	(0.0009)
Missing	**0.0039*** **	**0.0047*** **	**0.0038*** **	**0.0036*** **
	(0.001)	(0.001)	(0.001)	(0.001)
Fraction of female employees	**-0.0056**	**-0.0031**	**-0.0043**	**-0.0103** **
	(0.0036)	(0.0039)	(0.0035)	(0.0035)
DHS employment growth index	**0.0082*** **	**0.0098*** **	**0.0084*** **	**0.0083*** **
	(0.002)	(0.002)	(0.002)	(0.002)
More employees than previous year	**0.0037*** **	**0.0042*** **	**0.0038*** **	**0.0037*** **
	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Wanted to hire people, but did not				
Wanted to hire people, but did not	**-0.0002**	**0.0005**	**-0.0005**	**0.0001**
	(0.0011)	(0.0012)	(0.0012)	(0.0012)
Missing	**-0.0002**	**0.0008**	**-0.0004**	**0.001**
	(0.0022)	(0.0023)	(0.0022)	(0.0022)
Expected business volume relative to last year				
Increasing	**0.0013** **	**0.0015** **	**0.0013** **	**0.0012** *
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Decreasing	**-0.0044*** **	**-0.0048*** **	**-0.0044*** **	**-0.0045*** **
	(0.0005)	(0.0006)	(0.0006)	(0.0006)
Missing	**-0.0009**	**-0.0013**	**-0.0009**	**-0.0007**
	(0.001)	(0.001)	(0.001)	(0.0011)
Total number of new employees	**0.000013** *	**0.000012** *	**0.000015** **	**0.00002*** **
	(0.000006)	(0.000005)	(0.000005)	(0.000005)
Firm was hiring in current year	**0.0048*** **	**0.0057*** **	**0.0049*** **	**0.0046*** **
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Total number of employees that left	**-0.000005**	**-0.000007** *	**-0.000006**	**-0.000005**
	(0.000003)	(0.000003)	(0.000003)	(0.000003)
Employees have left in current year	**0.0021*** **	**0.0028*** **	**0.0019** **	**0.002** **
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Number of days in current establishment	**0.000027*** **	**0.000018*** **	**0.000231*** **	**0.000042*** **
	(0.000005)	(0.000001)	(0.000008)	(0.000006)
Age at end of year, squared	**-0.002209*** **	**-0.003944*** **	**-0.000036**	**-0.00285*** **
	(0.000128)	(0.000139)	(0.000332)	(0.000123)
Age at and of year^3	**0.000039*** **	**0.000063*** **	**-0.000004**	**0.000048*** **
	(0.000002)	(0.000002)	(0.000005)	(0.000002)
Age at end of year^4	**-0.00000027*** **	**-0.00000039*** **	**0.00000003**	**-0.00000032*** **
	(0.00000001)	(0.00000001)	(0.00000002)	(0.00000001)
Part time job	**-0.3395*** **	**-0.4446*** **	**-0.3451*** **	**-0.3227*** **
	(0.0046)	(0.0059)	(0.0046)	(0.0042)
Daily wage topcoded	**0.0185*** **	**0.0231*** **	**0.0232*** **	**0.0247*** **
	(0.0012)	(0.0013)	(0.0012)	(0.0012)
Year				
1994	**-0.0351*** **	**0.0757*** **	**0.0171** *	**-0.0303*** **
	(0.0044)	(0.0044)	(0.0076)	(0.0044)
1995	**-0.0562*** **	**0.1658*** **	**0.048** **	**-0.0429*** **
	(0.0082)	(0.0082)	(0.0147)	(0.0083)
1996	**-0.0969*** **	**0.2365*** **	**0.0592** **	**-0.0777*** **
	(0.0122)	(0.0122)	(0.0221)	(0.0123)
1997	**-0.1403*** **	**0.3022*** **	**0.0674** *	**-0.1148*** **
	(0.0162)	(0.0162)	(0.0293)	(0.0163)
1998	**-0.183*** **	**0.369*** **	**0.0772** *	**-0.1505*** **
	(0.0202)	(0.0202)	(0.0367)	(0.0203)
1999	**-0.2154*** **	**0.4417*** **	**0.0958** *	**-0.176*** **
	(0.0243)	(0.0242)	(0.0441)	(0.0244)
2000	**-0.2521*** **	**0.5178*** **	**0.1118** *	**-0.2063*** **
	(0.0283)	(0.0283)	(0.0514)	(0.0284)
2001	**-0.2868*** **	**0.5946*** **	**0.1292** *	**-0.2347*** **
	(0.0324)	(0.0323)	(0.0587)	(0.0325)
2002	**-0.3231*** **	**0.6707*** **	**0.1456** *	**-0.2643*** **
	(0.0364)	(0.0363)	(0.0659)	(0.0365)

2003	**-0.3506*** **	**0.7543*** **	**0.1713***	**-0.2893*** **
	(0.0404)	(0.0404)	(0.0733)	(0.0405)
2004	**-0.3986*** **	**0.8177*** **	**0.1764***	**-0.3307*** **
	(0.0445)	(0.0444)	(0.0807)	(0.0446)
2005	**-0.4413*** **	**0.8871*** **	**0.186***	**-0.3673*** **
	(0.0485)	(0.0485)	(0.0879)	(0.0487)
2006	**-0.4837*** **	**0.9556*** **	**0.1956***	**-0.4028*** **
	(0.0525)	(0.0525)	(0.0953)	(0.0527)
2007	**-0.5172*** **	**1.034*** **	**0.2145***	**-0.4293*** **
	(0.0566)	(0.0566)	(0.1026)	(0.0567)
2008	**-0.5461*** **	**1.1169*** **	**0.2381***	**-0.4525*** **
	(0.0607)	(0.0606)	(0.11)	(0.0609)
Potential Experience in years^2			**-0.0035*** **	
			(0.0002)	
Potential Experience in years^3			**0.000095*** **	
			(0.000006)	
Potential Experience in years^4			**-0.0000009*** **	
			(0.00000004)	
Actual Experience in years	**0.1579*** **	**0.1171*** **		**0.1525*** **
	(0.0026)	(0.0014)		(0.003)
Actual Experience in years^2	**-0.003233*** **	**-0.004723*** **		**-0.002067*** **
	(0.000109)	(0.000152)		(0.000056)
Actual Experience in years^3	**0.000097*** **	**0.00016*** **		**0.000057*** **
	(0.000004)	(0.000006)		(0.000002)
Actual Experience in years^4	**-0.000001*** **	**-0.000002*** **		**-0.00000055*** **
	(0.00000005)	(0.00000008)		(0.00000003)
Number of observations	9792405	9792405	9792405	8693593
Number of individuals	3068373	3068373	3068373	2726651
Number of firms	24323	24323	24323	23348
Number of matches	3413921		3413921	2996587
Total Sum of Squares	4104476	4104476	4104476	3588938
Residual Sum of Squares	175455	244647	176172	151418
R2	0.9573	0.9404	0.9571	0.9578
F-stat of all coefficients	1042.7	1037.6	1027.3	979.3
p-value	0	0	0	0
F-stat of all fixed effects	19.1	15.2	18.3	19.2
p-value	0	0	0	0
F-stat individual FE		11.7		
p-value		0		
F-stat firm FE		52.2		
p-value		0		

Omitted Categories: *Business volume, categorical:* 0 to 72,000; *Business volume per employee, categorical:* 0 to 21,300; *Investment per employee, categorical:* 0 to 500; *Wanted to hire people, but did not:* successfully hired or did not want to; *Expected business volume relative to last year:* same; *Year:* 1993

Notes: Standard errors are clustered at the firm and individual level. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table A6: Regressions of Firm Fixed Effects on Time Invariant Firm Characteristics, All Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Industry				
Agriculture and Forestry	**0.0097**	**0.023**	**0.035**	**-0.0178**
	(0.1517)	(0.1482)	(0.1382)	(0.1427)
Mining, quarrying and electricity	**-0.0222**	**-0.0342**	**0.0946**	**-0.0444**
	(0.1826)	(0.1815)	(0.1756)	(0.2213)
Food products	**0.0111**	**0.0005**	**0.084**	**-0.02**
	(0.083)	(0.0827)	(0.0813)	(0.0813)
Clothing and Textile	**0.0923**	**0.0581**	**-0.1374**	**-0.015**
	(0.1003)	(0.0996)	(0.1228)	(0.1051)
Paper and Printing	**0.1463**	**0.1269**	**0.0365**	**0.0751**
	(0.1497)	(0.1483)	(0.1197)	(0.1531)
Wood Products, Furniture, Jewellery and Toys	**0.353***** (0.0914)	**0.3296***** (0.0925)	**0.2399**** (0.0906)	**0.3384***** (0.0951)
Chemical Industry	**0.2447***	**0.2355***	**0.0506**	**0.2226**
	(0.1203)	(0.1073)	(0.2508)	(0.1395)
Rubber/Plastic	**0.2319****	**0.2242****	**0.2315****	**0.1788****
	(0.0757)	(0.072)	(0.0753)	(0.0683)
Non-metallic Mineral Products	**0.2292***	**0.2386***	**0.195**	**0.2078***
	(0.0967)	(0.0955)	(0.1097)	(0.0938)
Basic Metals, Steel, light Metal	**0.2922*****	**0.2799*****	**0.1797***	**0.2444*****
	(0.0633)	(0.0606)	(0.0701)	(0.0638)
Recycling	**-0.6061**	**-0.6308**	**-0.4216**	**-0.6119**
	(0.8601)	(0.8529)	(0.6481)	(0.8785)
Machinery	**0.3259*****	**0.2903****	**0.3044****	**0.3036****
	(0.0892)	(0.0884)	(0.0953)	(0.0924)
Motor vehicles: production/sales/repair/fuel	**-0.0523** (0.0753)	**-0.0459** (0.0744)	**-0.122** (0.0752)	**-0.0614** (0.0704)
Other Transport Equipment	**0.2208**	**0.182**	**0.1617**	**0.1701**
	(0.1267)	(0.1282)	(0.1348)	(0.1212)
Electrical Equipment	**0.1582**	**0.1857**	**0.1151**	**0.0964**
	(0.0913)	(0.0955)	(0.1075)	(0.0932)
Precision and Optical Equipment	**0.1238**	**0.1145**	**0.1744**	**0.112**
	(0.0948)	(0.0944)	(0.0924)	(0.0947)
Main Construction Trade	**0.2746*****	**0.2828*****	**0.1295**	**0.241*****
	(0.0668)	(0.0656)	(0.069)	(0.0683)
Building Installation/Completion	**0.2129*****	**0.2182*****	**0.0857**	**0.1888****
	(0.0643)	(0.0629)	(0.0602)	(0.062)
Transportation	**0.0196**	**-0.0077**	**0.0322**	**-0.0017**
	(0.074)	(0.0724)	(0.0728)	(0.0744)
Communication	**-0.2788**	**-0.2588**	**-0.0935**	**-0.2873**
	(0.2347)	(0.2263)	(0.209)	(0.2201)
Credit and Financal Intermediation	**0.0569**	**0.0585**	**-0.0343**	**0.0972**
	(0.0974)	(0.0928)	(0.1077)	(0.1087)
Insurance	**0.001**	**-0.0476**	**0.0323**	**-0.0059**
	(0.1713)	(0.1684)	(0.1535)	(0.1704)
Computer and Related Activities	**0.3042***	**0.3454****	**0.2247**	**0.2972***
	(0.1259)	(0.1125)	(0.1232)	(0.1422)
Research and Development	**0.219****	**0.2032***	**0.1334**	**0.2881****
	(0.0814)	(0.0808)	(0.0792)	(0.098)
Legal Consulting, Advertising	**-0.0672**	**-0.0626**	**-0.124**	**-0.1207**
	(0.1026)	(0.1011)	(0.0946)	(0.1166)
Real Estate	**-0.211**	**-0.1993**	**-0.0892**	**-0.2738**
	(0.1292)	(0.123)	(0.1184)	(0.1506)
Renting, Business Activities	**0.0377**	**0.0197**	**-0.0048**	**-0.0408**
	(0.0702)	(0.068)	(0.063)	(0.0769)

Hotel and Restaurant	**-0.3776*** **	**-0.3595*** **	**-0.3428*** **	**-0.395*** **
	(0.0794)	(0.0771)	(0.0773)	(0.0794)
Education/Teaching	**0.0995**	**0.0811**	**0.0027**	**0.0988**
	(0.0987)	(0.092)	(0.091)	(0.1012)
Human Health, Veterinary and Social Work	**0.0285**	**-0.0026**	**-0.0059**	**-0.0264**
	(0.0652)	(0.0639)	(0.0674)	(0.0655)
Sanitation	**0.1615**	**0.1295**	**0.2748**	**0.0359**
	(0.149)	(0.1394)	(0.1495)	(0.114)
Recreation, Culture, Sports	**-0.4495** **	**-0.451** **	**-0.3342* **	**-0.5665*** **
	(0.1677)	(0.1604)	(0.1461)	(0.1683)
Other Services	**-0.1998* **	**-0.1856**	**-0.1955* **	**-0.2485* **
	(0.1)	(0.1002)	(0.0993)	(0.1035)
Organizations, Lobbying, etc.	**-0.395** **	**-0.4172** **	**-0.2883* **	**-0.3177* **
	(0.1442)	(0.138)	(0.1421)	(0.1532)
Public Administration and Social Security	**0.0378**	**0.0195**	**-0.0048**	**0.0152**
	(0.0903)	(0.087)	(0.0885)	(0.0932)
State				
Schleswig-Holstein	**0.149* **	**0.1147**	**0.2315*** **	**0.1704** **
	(0.0675)	(0.0664)	(0.0702)	(0.0658)
Hamburg	**0.1895** **	**0.178** **	**0.2291** **	**0.238*** **
	(0.0712)	(0.0688)	(0.0776)	(0.0718)
Lower Saxony	**-0.0497**	**-0.0524**	**-0.0125**	**-0.0351**
	(0.0619)	(0.0597)	(0.0611)	(0.0634)
Bremen	**0.2206*** **	**0.2132*** **	**0.0987**	**0.1776** **
	(0.0632)	(0.062)	(0.0681)	(0.0677)
Hesse	**0.0594**	**0.0603**	**0.0842**	**0.0747**
	(0.0604)	(0.0584)	(0.053)	(0.0592)
Rhineland-Palatinate	**-0.1463* **	**-0.1623* **	**-0.0273**	**-0.0919**
	(0.068)	(0.0685)	(0.0675)	(0.0695)
Baden-Wuerttemberg	**0.0046**	**0.0037**	**0.0268**	**0.0325**
	(0.054)	(0.0524)	(0.0547)	(0.0557)
Bavaria	**0.1849*** **	**0.1706** **	**0.1678** **	**0.1601** **
	(0.0547)	(0.0534)	(0.0524)	(0.0563)
Saarland	**0.0867**	**0.055**	**0.145* **	**0.1081**
	(0.0581)	(0.0549)	(0.0569)	(0.0576)
Berlin	**0.058**	**0.0528**	**0.0146**	**0.1185**
	(0.0644)	(0.063)	(0.0683)	(0.066)
Brandenburg	**-0.0266**	**-0.0053**	**-0.1203**	**-0.0566**
	(0.0809)	(0.0774)	(0.0828)	(0.0841)
Mecklenburg-Western Pomerania	**0.091**	**0.1104**	**0.0101**	**0.0955**
	(0.0699)	(0.0673)	(0.0726)	(0.0722)
Saxony	**0.0005**	**0.0344**	**-0.1158**	**-0.0351**
	(0.0788)	(0.0763)	(0.0828)	(0.0796)
Saxony-Anhalt	**-0.0365**	**-0.0033**	**-0.1171**	**-0.0969**
	(0.0818)	(0.0785)	(0.0822)	(0.087)
Thuringia	**0.0183**	**0.0484**	**-0.1243**	**-0.0191**
	(0.0819)	(0.079)	(0.0853)	(0.0839)
Legal Form				
Individually-owned firm	**-0.3585*** **	**-0.3462*** **	**-0.3193*** **	**-0.3453*** **
	(0.0399)	(0.0389)	(0.0385)	(0.0405)
Partnership	**-0.1603*** **	**-0.1479** **	**-0.224*** **	**-0.1735*** **
	(0.0479)	(0.047)	(0.0467)	(0.0466)
Company limited by shares	**0.0728**	**0.0734**	**-0.0197**	**0.0785**
	(0.0619)	(0.0641)	(0.0662)	(0.0663)
Public corporation	**-0.058**	**-0.0635**	**-0.0343**	**-0.0412**
	(0.0795)	(0.076)	(0.0788)	(0.0797)
other legal form	**-0.1069**	**-0.101**	**-0.0982**	**-0.137**
	(0.0733)	(0.0709)	(0.0674)	(0.0742)
missing/don't know	**0.0281**	**-0.0125**	**0.1367**	**0.0068**
	(0.2088)	(0.2011)	(0.2641)	(0.2251)
Main/Exclusive Ownership				
Eastern German property	**0.1201* **	**0.1185* **	**0.0553**	**0.1106**
	(0.0573)	(0.0561)	(0.0641)	(0.0592)

Foreign property	**0.2526*** **	**0.2415** **	**0.2109** **	**0.244** **
	(0.0756)	(0.0746)	(0.0757)	(0.0822)
Public property	**0.0829**	**0.0951**	**0.0846**	**0.0697**
	(0.0675)	(0.0643)	(0.0647)	(0.0698)
no principal shareholder	**0.1496**	**0.1461**	**0.0463**	**0.1283**
	(0.0881)	(0.083)	(0.0845)	(0.0791)
Unknown	**-0.0961**	**-0.0976**	**-0.0943**	**-0.0652**
	(0.0985)	(0.0983)	(0.0939)	(0.1053)
Missing	**0.3172*** **	**0.3807*** **	**-0.0073**	**0.3054*** **
	(0.088)	(0.0781)	(0.0897)	(0.0921)
Year founded (only after 1990)				
1990	**0.0111**	**0.0061**	**0.0573**	**-0.0022**
	(0.0715)	(0.0686)	(0.0626)	(0.0726)
1991	**0.049**	**0.0421**	**0.1081**	**0.0478**
	(0.0951)	(0.0882)	(0.1108)	(0.0992)
1992	**-0.1807**	**-0.1885**	**-0.0634**	**-0.1615**
	(0.1104)	(0.1096)	(0.1183)	(0.1151)
1993	**-0.0171**	**-0.0352**	**-0.0269**	**-0.033**
	(0.0772)	(0.0742)	(0.0787)	(0.0764)
1994	**-0.0531**	**-0.0516**	**-0.0342**	**-0.0436**
	(0.1086)	(0.1056)	(0.1196)	(0.1102)
1995	**0.018**	**0.0042**	**0.0507**	**-0.0064**
	(0.0865)	(0.0861)	(0.0768)	(0.0858)
1996	**-0.055**	**-0.0564**	**0.0164**	**-0.0257**
	(0.0816)	(0.0762)	(0.0738)	(0.0875)
1997	**-0.145**	**-0.1476**	**-0.0371**	**-0.1792**
	(0.0945)	(0.0869)	(0.0924)	(0.0982)
1998	**-0.0483**	**-0.0556**	**0.0401**	**-0.1158**
	(0.1126)	(0.1087)	(0.1032)	(0.1125)
1999	**-0.0718**	**-0.1115**	**0.0404**	**-0.0365**
	(0.0927)	(0.0898)	(0.0883)	(0.0891)
2000	**0.1873**	**0.1291**	**0.3593*** **	**0.1945**
	(0.1023)	(0.1014)	(0.0989)	(0.1072)
2001	**-0.2087**	**-0.2301**	**0.0009**	**-0.1473**
	(0.1475)	(0.1432)	(0.1499)	(0.1592)
2002	**-0.0678**	**-0.102**	**0.2036**	**0.0169**
	(0.1128)	(0.1117)	(0.1072)	(0.1112)
2003	**-0.0607**	**-0.1207**	**0.189**	**-0.0099**
	(0.1189)	(0.1123)	(0.129)	(0.1207)
2004	**-0.0577**	**-0.1243**	**0.2189**	**0.048**
	(0.1302)	(0.1292)	(0.1369)	(0.1156)
2005	**0.0726**	**-0.0112**	**0.3268*** **	**0.1659**
	(0.1008)	(0.0975)	(0.0913)	(0.1046)
2006	**0.3205** **	**0.2202**	**0.6327*** **	**0.3881** **
	(0.1172)	(0.1155)	(0.1126)	(0.118)
2007	**0.238**	**0.1263**	**0.5837*** **	**0.3013** *
	(0.1364)	(0.1349)	(0.1693)	(0.1381)
2008	**-0.1132**	**-0.2126** *	**0.2387** **	**-0.0995**
	(0.0928)	(0.089)	(0.0911)	(0.0942)
missing	**-0.0309**	**-0.0165**	**-0.1324**	**-0.0823**
	(0.0907)	(0.0816)	(0.0915)	(0.0946)
Establishment/Department is...				
place of business/office/branch	**0.0365**	**0.0436**	**0.0648**	**0.0306**
	(0.0444)	(0.0424)	(0.0409)	(0.0469)
head office	**0.0713**	**0.0655**	**0.0282**	**0.0537**
	(0.0561)	(0.0537)	(0.0496)	(0.0573)
middle-level authority	**0.0597**	**0.0667**	**0.0755**	**0.0159**
	(0.0829)	(0.0775)	(0.0743)	(0.079)
Missing	**0.0412**	**0.1051**	**0.0911**	**0.1107**
	(0.1655)	(0.1531)	(0.1847)	(0.1867)
Company pays for job training/courses				
Yes	**-0.119*** **	**-0.1113** **	**-0.1723*** **	**-0.1187*** **
	(0.0355)	(0.0345)	(0.0339)	(0.0359)

Missing	**-0.1222***	**-0.1007**	**-0.0948**	**-0.1208***
	(0.0574)	(0.0562)	(0.0548)	(0.0583)
Has Worker's Council				
Yes	**0.1215*****	**0.1385*****	**0.0696**	**0.1149****
	(0.0359)	(0.0347)	(0.0356)	(0.0381)
Missing	**-0.1565**	**-0.1277**	**-0.1184**	**-0.1649**
	(0.1848)	(0.174)	(0.1888)	(0.1985)
Collective Wage Agreement				
company agreement	**-0.0448**	**-0.0388**	**-0.1046**	**-0.0594**
	(0.0726)	(0.0701)	(0.0705)	(0.0759)
no collective agreement	**-0.1009****	**-0.1187*****	**-0.0342**	**-0.1074****
	(0.0346)	(0.0338)	(0.0338)	(0.0351)
Missing	**-0.0453**	**-0.0317**	**-0.1638***	**-0.0542**
	(0.074)	(0.0728)	(0.0735)	(0.0777)
Owner working in Company				
No	**0.06**	**0.0538**	**0.0399**	**0.0644**
	(0.0391)	(0.0375)	(0.0369)	(0.0406)
Missing	**-0.3195***	**-0.3009***	**-0.5088****	**-0.0562**
	(0.1469)	(0.1293)	(0.1853)	(0.0985)
Constant	**4.0106*****	**5.344*****	**4.404*****	**4.3716*****
	(0.0581)	(0.0566)	(0.057)	(0.0593)
Number of Observations	24291	24291	24291	23318
R-squared	0.1619	0.1814	0.1287	0.1567

Omitted Categories: *Industry:* Sales: Retail and Wholesale; *State:* North Rhine-Westphalia; *Legal Form:* Limited liability company; *Main/Exclusive Ownership:* Western German property; *Year founded:* Founded before 1990; *Establishment/Department is...:* independent company/organisation w/o other places of business; *Company pays for job training/courses:* No; *Has Worker's Council:* No; *Collective Wage Agreement:* Industry-wide wage agreement; *Owner working in Company:*Yes;
Notes:. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table A7: Regressions of Individual FE on Time Invariant Firm Characteristics, All Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Female	**-0.0559*** **	**-0.1035*** **	**-0.2092*** **	**-0.0514*** **
	(0.0032)	(0.0029)	(0.0032)	(0.0034)
Nationality, grouped				
Turkey	**-0.0288*** **	**-0.0073**	**0.0244*** **	**-0.0334*** **
	(0.0072)	(0.0067)	(0.0073)	(0.0073)
Italy	**-0.0128**	**-0.0192**	**0.0054**	**-0.0244**
	(0.0123)	(0.0108)	(0.0121)	(0.0128)
Yugoslavia, Serbia and Montenegro	**-0.0465*** **	**-0.0183**	**0.0826*** **	**-0.0566*** **
	(0.0126)	(0.0118)	(0.0139)	(0.0127)
Greece	**0.0152**	**0.0097**	**0.0329**	**0.0086**
	(0.0187)	(0.0171)	(0.0179)	(0.0199)
France	**0.0855*** **	**0.0798*** **	**0.0874*** **	**0.0967** **
	(0.0256)	(0.0208)	(0.022)	(0.0309)
Poland	**0.1499*** **	**0.1187*** **	**0.1293*** **	**0.1539*** **
	(0.0293)	(0.0254)	(0.0232)	(0.0304)
Austria	**0.0062**	**0.0301**	**0.1276*** **	**0.0061**
	(0.0279)	(0.0261)	(0.0357)	(0.033)
Croatia	**0.0279**	**0.0415***	**0.0364**	**0.0118**
	(0.0208)	(0.0184)	(0.0203)	(0.022)
Portugal	**-0.0852*** **	**-0.0354**	**0.0102**	**-0.1062*** **
	(0.0229)	(0.0217)	(0.0244)	(0.0238)
Spain	**-0.1055*** **	**-0.0778*** **	**-0.0567***	**-0.1202*** **
	(0.025)	(0.0213)	(0.0262)	(0.0278)
Netherlands, Luxembourg	**0.218*** **	**0.2009*** **	**0.2082*** **	**0.2042*** **
	(0.0341)	(0.0298)	(0.0326)	(0.0377)
Russia, Belarus, Former Soviet Union	**0.1182*** **	**0.0955** **	**0.0861** **	**0.1167** **
	(0.0336)	(0.0311)	(0.0279)	(0.036)
Bosnia and Herzegovina	**0.0722**	**0.0501**	**0.075***	**0.0692**
	(0.041)	(0.0328)	(0.032)	(0.0435)
Great Britain, Ireland and Northern Ireland	**0.0759***	**0.0788***	**0.1665*** **	**0.1212*** **
	(0.0328)	(0.0309)	(0.0366)	(0.0284)
Romania	**0.1272** **	**0.1034***	**0.1603*** **	**0.091**
	(0.0449)	(0.0414)	(0.043)	(0.0475)
Czech Republic, Slovakia, Former Czechoslovakia	**0.2572*** **	**0.2676*** **	**0.2873*** **	**0.2795*** **
	(0.0702)	(0.0785)	(0.064)	(0.0799)
Ukraine, Moldova	**0.1916**	**0.1719**	**0.1405**	**0.1703**
	(0.1666)	(0.1622)	(0.1443)	(0.1486)
Hungary	**-0.0843**	**-0.1087**	**-0.0201**	**-0.0532**
	(0.167)	(0.1742)	(0.1266)	(0.1284)
Albania	**0.2473** **	**0.2286***	**0.1773***	**0.1797** **
	(0.0916)	(0.0917)	(0.0842)	(0.0557)
Belgium	**0.1845*** **	**0.1981*** **	**0.2577*** **	**0.1896*** **
	(0.0387)	(0.0337)	(0.048)	(0.0468)
Macedonia	**0.0632**	**0.0456**	**0.0066**	**0.0454**
	(0.0771)	(0.0739)	(0.0792)	(0.0806)
Switzerland	**0.1497*** **	**0.1257*** **	**0.1714*** **	**0.1325** **
	(0.0413)	(0.0349)	(0.0433)	(0.0507)
Bulgaria	**0.0117**	**0.0523**	**0.0702**	**-0.2161** **
	(0.1148)	(0.1197)	(0.1068)	(0.0792)
Slovenia	**-0.0439**	**-0.0243**	**0.0507**	**-0.0263**
	(0.0681)	(0.0603)	(0.0685)	(0.0582)
Denmark, Sweden	**0.232*** **	**0.2214*** **	**0.1923***	**0.2315*** **
	(0.0592)	(0.0454)	(0.0777)	(0.0687)
Finland	**0.0323**	**0.0075**	**-0.0123**	**0.0794**

	(0.0895)	(0.0669)	(0.0766)	(0.0966)
Estonia, Latvia, Lithuania	**0.0977**	**0.1153**	**0.0277**	**0.0797**
	(0.0683)	(0.0626)	(0.0903)	(0.0787)
Europe (other)	**0.5538*** **	**0.3933** **	**0.1574**	**0.5417*** **
	(0.131)	(0.1387)	(0.2469)	(0.1446)
Ethiopia	**0.043**	**0.0284**	**0.0542**	**-0.001**
	(0.0689)	(0.0737)	(0.07)	(0.073)
Ghana	**0.0987**	**0.0615**	**0.013**	**0.0945***
	(0.0538)	(0.0563)	(0.0434)	(0.0447)
Morocco	**-0.0627***	**-0.0562***	**-0.0674***	**-0.0909** **
	(0.0298)	(0.0281)	(0.0338)	(0.029)
Tunisia	**0.0908**	**0.0432**	**0.0466**	**0.0847**
	(0.0503)	(0.0421)	(0.0593)	(0.0551)
Africa (other)	**0.1093** **	**0.0857** **	**0.0437**	**0.0914** **
	(0.0343)	(0.0322)	(0.031)	(0.0349)
USA, Canada	**0.1646*** **	**0.1415*** **	**0.2398*** **	**0.14*** **
	(0.0336)	(0.0321)	(0.0334)	(0.0357)
America (other)	**-0.0066**	**-0.0235**	**0.0077**	**-0.0192**
	(0.0605)	(0.0572)	(0.0613)	(0.076)
Afghanistan	**0.1216**	**0.131***	**0.0646**	**0.0636**
	(0.0726)	(0.0585)	(0.0741)	(0.07)
Sri Lanka	**0.2474*** **	**0.2144*** **	**0.151*** **	**0.1069*** **
	(0.0476)	(0.0457)	(0.0437)	(0.0293)
Vietnam	**-0.0471**	**-0.052**	**-0.0736**	**-0.0366**
	(0.06)	(0.0622)	(0.0673)	(0.0584)
India	**0.0325**	**0.0355**	**0.0307**	**0.0141**
	(0.0837)	(0.0875)	(0.0737)	(0.0786)
Iraq	**0.3026*** **	**0.2805*** **	**0.1492*** **	**0.2552*** **
	(0.0668)	(0.0646)	(0.0408)	(0.0632)
Iran	**0.1635** **	**0.1134***	**0.0692**	**0.1383***
	(0.0569)	(0.0511)	(0.052)	(0.0679)
Lebanon	**0.0698**	**0.0195**	**0.0679**	**0.0811**
	(0.0759)	(0.0611)	(0.0761)	(0.082)
Philippines	**-0.0187**	**-0.0015**	**-0.0421**	**-0.0475**
	(0.0545)	(0.0538)	(0.0675)	(0.0587)
Thailand	**-0.1378**	**-0.0994**	**-0.0547**	**-0.163**
	(0.1041)	(0.1005)	(0.079)	(0.105)
China, incl. Tibet	**-0.1812***	**-0.1856** **	**-0.1116**	**-0.3155** **
	(0.0779)	(0.0684)	(0.0764)	(0.1146)
Asia (other)	**0.1114** **	**0.0963***	**0.0845***	**0.0842***
	(0.0405)	(0.0386)	(0.0355)	(0.0429)
Oceania	**0.2028** **	**0.1453***	**0.1316***	**0.1891***
	(0.0661)	(0.0595)	(0.0658)	(0.0766)
Missing	**-0.0722**	**-0.0951**	**-0.0138**	**-0.0829**
	(0.1032)	(0.0739)	(0.0525)	(0.1092)
School education and vocational training				
Secondary / intermediate school w/o completed vocational training	**-0.2325*** ** (0.0042)	**-0.2605*** ** (0.0038)	**-0.2917*** ** (0.0043)	**-0.2233*** ** (0.0042)
Upper secondary school w/o completed vocational training	**-0.3391*** ** (0.0152)	**-0.3914*** ** (0.0136)	**-0.5038*** ** (0.0139)	**-0.3507*** ** (0.0167)
Upper secondary school with completed vocational training	**0.2338*** ** (0.007)	**0.1849*** ** (0.0065)	**0.0761*** ** (0.0067)	**0.2385*** ** (0.007)
Completion of a university of applied sciences	**0.3861*** ** (0.007)	**0.3147*** ** (0.0064)	**0.0495*** ** (0.0077)	
College / university degree	**0.5251*** ** (0.006)	**0.4205*** ** (0.0053)	**0.0809*** ** (0.0061)	
Missing	**-0.0248*** ** (0.0061)	**-0.1286*** ** (0.0056)	**-0.143*** ** (0.0057)	**-0.0143*** (0.0062)
Year of first employment				
1976	**0.1695*** ** (0.0096)	**-0.0001** (0.0078)	**0.0331*** ** (0.0097)	**0.1761*** ** (0.0107)
1977	**0.2175*** ** (0.0096)	**-0.0577*** ** (0.0081)	**-0.0238*** (0.0093)	**0.2127*** ** (0.0104)

1978	**0.2692*** **	**-0.1032*** **	**-0.0582*** **	**0.2501*** **
	(0.0122)	(0.0105)	(0.0109)	(0.0133)
1979	**0.2881*** **	**-0.1924*** **	**-0.1234*** **	**0.2506*** **
	(0.0111)	(0.0094)	(0.01)	(0.0121)
1980	**0.3088*** **	**-0.2679*** **	**-0.1802*** **	**0.2658*** **
	(0.0111)	(0.0097)	(0.0106)	(0.012)
1981	**0.3446*** **	**-0.3277*** **	**-0.223*** **	**0.288*** **
	(0.0093)	(0.0083)	(0.0095)	(0.0097)
1982	**0.3539*** **	**-0.4104*** **	**-0.2802*** **	**0.2737*** **
	(0.0099)	(0.009)	(0.0099)	(0.0106)
1983	**0.3846*** **	**-0.4804*** **	**-0.3371*** **	**0.2946*** **
	(0.0103)	(0.0093)	(0.01)	(0.0113)
1984	**0.4202*** **	**-0.5414*** **	**-0.3727*** **	**0.3155*** **
	(0.0105)	(0.0092)	(0.0103)	(0.0123)
1985	**0.4444*** **	**-0.615*** **	**-0.4388*** **	**0.3359*** **
	(0.01)	(0.0094)	(0.0098)	(0.0109)
1986	**0.4523*** **	**-0.7006*** **	**-0.4817*** **	**0.3266*** **
	(0.0097)	(0.0091)	(0.0098)	(0.0106)
1987	**0.5063*** **	**-0.7474*** **	**-0.5284*** **	**0.369*** **
	(0.0086)	(0.0077)	(0.0088)	(0.0093)
1988	**0.5408*** **	**-0.8142*** **	**-0.5693*** **	**0.3885*** **
	(0.0087)	(0.0081)	(0.009)	(0.0091)
1989	**0.579*** **	**-0.8784*** **	**-0.6156*** **	**0.4203*** **
	(0.0077)	(0.0071)	(0.0078)	(0.0082)
1990	**0.6343*** **	**-0.9326*** **	**-0.649*** **	**0.4754*** **
	(0.0072)	(0.0066)	(0.0072)	(0.0075)
1991	**0.6729*** **	**-1.0128*** **	**-0.7425*** **	**0.503*** **
	(0.0095)	(0.0088)	(0.0094)	(0.0098)
1992	**0.6705*** **	**-1.0868*** **	**-0.8129*** **	**0.4827*** **
	(0.0112)	(0.0105)	(0.011)	(0.0117)
1993	**0.7112*** **	**-1.134*** **	**-0.8396*** **	**0.5086*** **
	(0.01)	(0.0094)	(0.0101)	(0.0102)
1994	**0.7472*** **	**-1.1981*** **	**-0.8991*** **	**0.5335*** **
	(0.0088)	(0.0081)	(0.01)	(0.0094)
1995	**0.772*** **	**-1.2728*** **	**-0.9604*** **	**0.5552*** **
	(0.0095)	(0.009)	(0.0099)	(0.0097)
1996	**0.8177*** **	**-1.3224*** **	**-1.0141*** **	**0.5755*** **
	(0.0094)	(0.0089)	(0.01)	(0.0099)
1997	**0.8261*** **	**-1.4073*** **	**-1.0695*** **	**0.5841*** **
	(0.0108)	(0.0103)	(0.0103)	(0.011)
1998	**0.8605*** **	**-1.4658*** **	**-1.1269*** **	**0.605*** **
	(0.0099)	(0.0093)	(0.0102)	(0.0108)
1999	**0.6538*** **	**-1.7615*** **	**-1.4275*** **	**0.3744*** **
	(0.0087)	(0.0083)	(0.0084)	(0.0093)
2000	**0.7696*** **	**-1.7423*** **	**-1.4012*** **	**0.4832*** **
	(0.011)	(0.0105)	(0.0106)	(0.0117)
2001	**0.7744*** **	**-1.829*** **	**-1.483*** **	**0.4731*** **
	(0.0106)	(0.01)	(0.0103)	(0.0106)
2002	**0.7907*** **	**-1.9012*** **	**-1.5519*** **	**0.4934*** **
	(0.0127)	(0.0121)	(0.0127)	(0.0128)
2003	**0.7585*** **	**-2.0238*** **	**-1.6506*** **	**0.4462*** **
	(0.0155)	(0.0151)	(0.0149)	(0.0154)
2004	**0.7291*** **	**-2.1341*** **	**-1.7672*** **	**0.3967*** **
	(0.0148)	(0.0141)	(0.0158)	(0.0155)
2005	**0.7176*** **	**-2.2329*** **	**-1.8528*** **	**0.3691*** **
	(0.0166)	(0.0159)	(0.0182)	(0.0177)
2006	**0.7364*** **	**-2.3*** **	**-1.919*** **	**0.3755*** **
	(0.0187)	(0.0171)	(0.0208)	(0.0197)
2007	**0.7236*** **	**-2.4004*** **	**-1.9845*** **	**0.3523*** **
	(0.02)	(0.0185)	(0.0215)	(0.0219)
2008	**0.755*** **	**-2.4621*** **	**-2.0181*** **	**0.3536*** **
	(0.0274)	(0.0261)	(0.0273)	(0.0276)
Age at first employment	**0.0455*** **	**0.1075*** **	**0.0604*** **	**0.0644*** **

	(0.0003)	(0.0002)	(0.0003)	(0.0003)
Constant	**-1.4123*** **	**-1.7287*** **	**-0.6584*** **	**-1.652*** **
	(0.0078)	(0.0071)	(0.0076)	(0.0081)
Number of Observations	3062118	3062118	3062118	2720888
R-squared	0.2509	0.7443	0.4988	0.2798

Omitted Categories: *Nationality, grouped:* Germany; *School education and vocational training:* Secondary/intermediate school with completed vocational training; *Year of first employment:* 1975 or earlier;
Notes: Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table A8: Regression of Match and Firm Fixed Effects on Pre-Match Characteristics, All Coefficients

	Match Effect	Firm Effect
Part time job (at beginning of match)	**-0.1187*** **	**-0.6257*** **
	(0.0047)	(0.0151)
Employment Status 8 days before current match		
No Previous Record	**-0.04*** **	**-0.0796*** **
	(0.0066)	(0.0176)
Previous spell was benefits or gap	**-0.0012**	**-0.057*** **
	(0.0039)	(0.0107)
Apprentice/Trainee at other Firm	**0.0074**	**-0.0164**
	(0.0155)	(0.0488)
Number of Days in Labor Market Status 8 Days before current match (Main Effect)	**-0.000001**	**0.000009*** **
	(0.000001)	(0.000002)
…if previous spell was benefits or gap (Interaction)	**0.000001**	**0.000003**
	(0.000003)	(0.000008)
…if previous spell was training (Interaction)	**0.000016**	**0.000009**
	(0.000019)	(0.000058)
Year Match Started		
1994	**0.0024**	**0.0315**
	(0.0059)	(0.0172)
1995	**0.0115**	**-0.0106**
	(0.0079)	(0.0243)
1996	**0.0145***	**0.0104**
	(0.0071)	(0.0173)
1997	**0.0126**	**0.0105**
	(0.0081)	(0.0236)
1998	**0.0217** **	**-0.015**
	(0.0067)	(0.0168)
1999	**-0.0181**	**-0.1885*** **
	(0.0104)	(0.029)
2000	**-0.0035**	**-0.0955*** **
	(0.0071)	(0.0189)
2001	**0.0138***	**-0.0263**
	(0.0063)	(0.0163)
2002	**0.0189** **	**-0.0035**
	(0.0059)	(0.0166)
2003	**0.0243*** **	**-0.1059*** **
	(0.0069)	(0.0213)
2004	**0.0223*** **	**-0.0391**
	(0.0059)	(0.0267)
2005	**0.026*** **	**-0.0119**
	(0.0062)	(0.0189)
2006	**0.0353*** **	**-0.0048**
	(0.0063)	(0.0194)
2007	**0.0436*** **	**0.0327**
	(0.006)	(0.0185)
	0.0561* **	**0.025**
2008	(0.0065)	(0.0188)
Female	**-0.0529**	**-0.376*** **
	(0.0433)	(0.1055)
Number of Days of Benefit Receipt up to beginning of current match	**0.000007***	**-0.000082*** **
	(0.000003)	(0.000009)
Years since first Employment at beginning of current match	**0.0024** **	**-0.0046***
	(0.0009)	(0.0022)
…interacted with female dummy	**-0.0024**	**-0.0132*** **
	(0.0014)	(0.0036)
Years since first employment squared	**-0.000129*** **	**-0.000058**
	(0.00003)	(0.000077)
…interacted with female dummy	**0.000082**	**0.000457*** **
	(0.000048)	(0.000131)

Age at beginning of current match	**0.0024**	**0.0364*** **
	(0.0015)	(0.0038)
…interacted with female dummy	**0.0043**	**0.0221*** **
	(0.0025)	(0.0059)
Age squared	**-0.00002**	**-0.000387*** **
	(0.000019)	(0.000048)
Match Count		
2	**-0.0063**	**0.0612*** **
	(0.0043)	(0.0117)
…interacted with female dummy	**0.0071**	**0.0027**
	(0.0065)	(0.0207)
3	**-0.0229** **	**0.0958*** **
	(0.0087)	(0.0236)
…interacted with female dummy	**0.0208**	**-0.0253**
	(0.0142)	(0.052)
4	**-0.0576***	**0.1722*** **
	(0.0252)	(0.0401)
…interacted with female dummy	**0.0807**	**0.1501**
	(0.0562)	(0.1395)
5	**0.1193*** **	**-0.05**
	(0.0188)	(0.0573)
…interacted with female dummy	**-0.0815**	**0.1983**
	(0.0757)	(0.1426)
6	**-0.2742*** **	**0.2965*** **
	(0.0055)	(0.0263)
…interacted with female dummy	**0.0918*** **	**-0.3977*** **
	(0.0086)	(0.0315)
Constant	**-0.0565***	**3.7472*** **
	(0.0258)	(0.0691)
Number of Observations	665080	665080
R-squared	0.0343	0.2189

Omitted Categories: *Emp. Status 8 days before current match:* Employment at other Firm; *Year Match Started:* 1993; *Match Count:* 1

Note: Estimates of Match and Firm Fixed Effects are from main regression on full sample including actual experience. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table 1: Estimation Time on Simulated Data

			Matlab		Stata	
Individuals	**Firms**	**Covariates**	**TWFE**	**Match**	**TWFE**	**Match**
1000000	*10000*	*5*	0.94	1.26	2.96	1.66
1000000	*10000*	*10*	1.13	1.42	4.25	1.87
1500000	*15000*	*5*	1.74	2.40	4.62	2.57
1500000	*15000*	*10*	2.03	2.65	6.70	2.90
2000000	*20000*	*5*	2.86	4.25	6.35	3.51
2000000	*20000*	*10*	3.31	4.63	9.12	3.95
2500000	*25000*	*5*	4.44	6.72	8.09	4.48
2500000	*25000*	*10*	4.97	7.02	11.06	5.06

Note: All times are the average of 10 iterations in minutes.

Table 2: Regressions of log of Daily Wage on Time Variant Characteristics, Selected Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Total number of employees	**0.000004***	**0.000004****	**0.000004***	**0.000003***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Investment per employee, categorical				
2,100 to 3,000	**0.002**	**0.0016**	**0.0017**	**0.001**
	(0.0012)	(0.0013)	(0.0012)	(0.0013)
3,000 to 4,200	**0.0017**	**0.0025****	**0.0017**	**0.0007**
	(0.0009)	(0.0009)	(0.0009)	(0.0009)
4,200 to 6,100	**0.0031*****	**0.0038*****	**0.0033*****	**0.0027****
	(0.001)	(0.001)	(0.001)	(0.001)
6,100 to 10,000	**0.0041*****	**0.0048*****	**0.0039*****	**0.0034*****
	(0.0009)	(0.0009)	(0.0009)	(0.0009)
10,000 to 18,600	**0.0057*****	**0.0066*****	**0.0057*****	**0.0048*****
	(0.001)	(0.001)	(0.001)	(0.001)
more than 18,600	**0.0058*****	**0.0068*****	**0.0059*****	**0.0055*****
	(0.0009)	(0.001)	(0.0009)	(0.0009)
Missing	**0.0039*****	**0.0047*****	**0.0038*****	**0.0036*****
	(0.001)	(0.001)	(0.001)	(0.001)
Fraction of female employees	**-0.0056**	**-0.0031**	**-0.0043**	**-0.0103****
	(0.0036)	(0.0039)	(0.0035)	(0.0035)
DHS employment growth index	**0.0082*****	**0.0098*****	**0.0084*****	**0.0083*****
	(0.002)	(0.002)	(0.002)	(0.002)
More employees than previous year	**0.0037*****	**0.0042*****	**0.0038*****	**0.0037*****
	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Expected business volume relative to last year				
Increasing	**0.0013****	**0.0015****	**0.0013****	**0.0012***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Decreasing	**-0.0044*****	**-0.0048*****	**-0.0044*****	**-0.0045*****
	(0.0005)	(0.0006)	(0.0006)	(0.0006)
Missing	**-0.0009**	**-0.0013**	**-0.0009**	**-0.0007**
	(0.001)	(0.001)	(0.001)	(0.0011)
Total number of new employees	**0.000013***	**0.000012***	**0.000015****	**0.00002*****
	(0.000006)	(0.000005)	(0.000005)	(0.000005)
Firm was hiring in current year	**0.0048*****	**0.0057*****	**0.0049*****	**0.0046*****
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Total number of employees that left	**-0.000005**	**-0.000007***	**-0.000006**	**-0.000005**
	(0.000003)	(0.000003)	(0.000003)	(0.000003)
Employees have left in current year	**0.0021*****	**0.0028*****	**0.0019****	**0.002****
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Number of days in current establishment	**0.000027*****	**0.000018*****	**0.000231*****	**0.000042*****
	(0.000005)	(0.000001)	(0.000008)	(0.000006)
Age at end of year, squared	**-0.002209*****	**-0.003944*****	**-0.000036**	**-0.00285*****
	(0.000128)	(0.000139)	(0.000332)	(0.000123)
Age at and of year^3	**0.000039*****	**0.000063*****	**-0.000004**	**0.000048*****
	(0.000002)	(0.000002)	(0.000005)	(0.000002)
Age at end of year^4	**-0.00000027*****	**-0.00000039*****	**0.00000003**	**-0.00000032*****
	(0.00000001)	(0.00000001)	(0.00000002)	(0.00000001)
Part time job	**-0.3395*****	**-0.4446*****	**-0.3451*****	**-0.3227*****
	(0.0046)	(0.0059)	(0.0046)	(0.0042)
Potential Experience in years^2			**-0.0035*****	
			(0.0002)	
Potential Experience in years^3			**0.000095*****	
			(0.000006)	
Potential Experience in years^4			**-0.0000009*****	
			(0.00000004)	
Actual Experience in years	**0.1579*****	**0.1171*****		**0.1525*****
	(0.0026)	(0.0014)		(0.003)
Actual Experience in years^2	**-0.003233*****	**-0.004723*****		**-0.002067*****

	(0.000109)	(0.000152)		(0.000056)
Actual Experience in years^3	**0.000097*** **	**0.00016*** **		**0.000057*** **
	(0.000004)	(0.000006)		(0.000002)
Actual Experience in years^4	**-0.000001*** **	**-0.000002*** **		**-0.00000055*** **
	(0.00000005)	(0.00000008)		(0.00000003)
Number of observations	9792405	9792405	9792405	8693593
Number of individuals	3068373	3068373	3068373	2726651
Number of firms	24323	24323	24323	23348
Number of matches	3413921		3413921	2996587
Total Sum of Squares	4104476	4104476	4104476	3588938
Residual Sum of Squares	175455	244647	176172	151418
R2	0.9573	0.9404	0.9571	0.9578
F-stat of all coefficients	1042.7	1037.6	1027.3	979.3
p-value	0	0	0	0
F-stat of all fixed effects	19.1	15.2	18.3	19.2
p-value	0	0	0	0
F-stat individual FE		11.7		
p-value		0		
F-stat firm FE		52.2		
p-value		0		

Omitted Categories: *Business volume, categorical:* 0 to 72,000; *Business volume per employee, categorical:* 0 to 21,300; *Investment per employee, categorical:* 0 to 500; *Wanted to hire people, but did not:* successfully hired or did not want to; *Expected business volume relative to last year:* same;

Notes: Regression also includes dummies for business volume, dummies for business volume per employee, the fraction of employees working part time, a dummy if they wanted to hire, but did not, year dummies and a dummy if an observation was top-coded. Table A5 in Appendix C reports all coefficients. Standard errors are clustered at the firm and individual level. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table 3: Regressions of Firm Fixed Effects on Time Invariant Firm Characteristics, Selected Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Legal Form				
Individually-owned firm	**-0.3585*** **	**-0.3462*** **	**-0.3193*** **	**-0.3453*** **
	(0.0399)	(0.0389)	(0.0385)	(0.0405)
Partnership	**-0.1603*** **	**-0.1479** **	**-0.224*** **	**-0.1735*** **
	(0.0479)	(0.047)	(0.0467)	(0.0466)
Company limited by shares	**0.0728**	**0.0734**	**-0.0197**	**0.0785**
	(0.0619)	(0.0641)	(0.0662)	(0.0663)
Public corporation	**-0.058**	**-0.0635**	**-0.0343**	**-0.0412**
	(0.0795)	(0.076)	(0.0788)	(0.0797)
other legal form	**-0.1069**	**-0.101**	**-0.0982**	**-0.137**
	(0.0733)	(0.0709)	(0.0674)	(0.0742)
missing/don't know	**0.0281**	**-0.0125**	**0.1367**	**0.0068**
	(0.2088)	(0.2011)	(0.2641)	(0.2251)
Main/Exclusive Ownership				
Eastern German property	**0.1201***	**0.1185***	**0.0553**	**0.1106**
	(0.0573)	(0.0561)	(0.0641)	(0.0592)
Foreign property	**0.2526*** **	**0.2415** **	**0.2109** **	**0.244** **
	(0.0756)	(0.0746)	(0.0757)	(0.0822)
Public property	**0.0829**	**0.0951**	**0.0846**	**0.0697**
	(0.0675)	(0.0643)	(0.0647)	(0.0698)
no principal shareholder	**0.1496**	**0.1461**	**0.0463**	**0.1283**
	(0.0881)	(0.083)	(0.0845)	(0.0791)
Unknown	**-0.0961**	**-0.0976**	**-0.0943**	**-0.0652**
	(0.0985)	(0.0983)	(0.0939)	(0.1053)
Missing	**0.3172*** **	**0.3807*** **	**-0.0073**	**0.3054*** **
	(0.088)	(0.0781)	(0.0897)	(0.0921)
Company pays for job training/courses				
Yes	**-0.119*** **	**-0.1113** **	**-0.1723*** **	**-0.1187*** **
	(0.0355)	(0.0345)	(0.0339)	(0.0359)
Missing	**-0.1222***	**-0.1007**	**-0.0948**	**-0.1208***
	(0.0574)	(0.0562)	(0.0548)	(0.0583)
Has Worker's Council				
Yes	**0.1215*** **	**0.1385*** **	**0.0696**	**0.1149** **
	(0.0359)	(0.0347)	(0.0356)	(0.0381)
Missing	**-0.1565**	**-0.1277**	**-0.1184**	**-0.1649**
	(0.1848)	(0.174)	(0.1888)	(0.1985)
Collective Wage Agreement				
company agreement	**-0.0448**	**-0.0388**	**-0.1046**	**-0.0594**
	(0.0726)	(0.0701)	(0.0705)	(0.0759)
no collective agreement	**-0.1009** **	**-0.1187*** **	**-0.0342**	**-0.1074** **
	(0.0346)	(0.0338)	(0.0338)	(0.0351)
Missing	**-0.0453**	**-0.0317**	**-0.1638***	**-0.0542**
	(0.074)	(0.0728)	(0.0735)	(0.0777)
Constant	**4.0106*** **	**5.344*** **	**4.404*** **	**4.3716*** **
	(0.0581)	(0.0566)	(0.057)	(0.0593)
Number of Observations	24291	24291	24291	23318
R-squared	0.1619	0.1814	0.1287	0.1567

Omitted Categories: *Industry:* Sales: Retail and Wholesale; *State:* North Rhine-Westphalia; *Legal Form:* Limited liability company; *Main/Exclusive Ownership:* Western German property; *Year founded:* Founded before 1990; *Establishment/Department is...:* independent company/organisation w/o other places of business; *Company pays for job training/courses:* No; *Has Worker's Council:* No; *Collective Wage Agreement:* Industry-wide wage agreement; *Owner working in Company:*Yes;

Notes: Regression also includes 35 industry dummies, state dummies, dummies for the year the firm was founded (with firms founded before 1990 pooled in one category), dummies for the type of establishment and a dummy whether the owner works in the firm. Table A6 in Appendix C reports all coefficients. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table 4: Regressions of Individual FE on Time Invariant Individual Characteristics, Selected Coefficients

	Actual Experience, Match	Actual Experience, TWFE	Potential Experience, Match	Actual Exp., Match, reduced Sample
Female	**-0.0559*** **	**-0.1035*** **	**-0.2092*** **	**-0.0514*** **
	(0.0032)	(0.0029)	(0.0032)	(0.0034)
School education and vocational training				
Secondary / intermediate school w/o completed vocational training	**-0.2325*** **	**-0.2605*** **	**-0.2917*** **	**-0.2233*** **
	(0.0042)	(0.0038)	(0.0043)	(0.0042)
Upper secondary school w/o completed vocational training	**-0.3391*** **	**-0.3914*** **	**-0.5038*** **	**-0.3507*** **
	(0.0152)	(0.0136)	(0.0139)	(0.0167)
Upper secondary school with completed vocational training	**0.2338*** **	**0.1849*** **	**0.0761*** **	**0.2385*** **
	(0.007)	(0.0065)	(0.0067)	(0.007)
Completion of a university of applied sciences	**0.3861*** **	**0.3147*** **	**0.0495*** **	
	(0.007)	(0.0064)	(0.0077)	
College / university degree	**0.5251*** **	**0.4205*** **	**0.0809*** **	
	(0.006)	(0.0053)	(0.0061)	
Missing	**-0.0248*** **	**-0.1286*** **	**-0.143*** **	**-0.0143***
	(0.0061)	(0.0056)	(0.0057)	(0.0062)
Age at first employment	**0.0455*** **	**0.1075*** **	**0.0604*** **	**0.0644*** **
	(0.0003)	(0.0002)	(0.0003)	(0.0003)
Constant	**-1.4123*** **	**-1.7287*** **	**-0.6584*** **	**-1.652*** **
	(0.0078)	(0.0071)	(0.0076)	(0.0081)
Number of Observations	3062118	3062118	3062118	2720888
R-squared	0.2509	0.7443	0.4988	0.2798

Omitted Categories: *Nationality, grouped:* Germany; *School education and vocational training:* Secondary/intermediate school with completed vocational training; *Year of first employment:* 1975 or earlier;
Notes: Regression also includes dummies for nationality (grouped), the year in which an individual first entered the labor market, but coefficients are not reported. Table A7 in Appendix C reports all coefficients. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Table 5: Regression of Match and Firm FE on Pre-Match Characteristics, Selected Coefficients

	Match Effect	Firm Effect
Part time job (at beginning of match)	**-0.1187*** **	**-0.6257*** **
	(0.0047)	(0.0151)
Employment Status 8 days before current match		
No Previous Record	**-0.04*** **	**-0.0796*** **
	(0.0066)	(0.0176)
Previous spell was benefits or gap	**-0.0012**	**-0.057*** **
	(0.0039)	(0.0107)
Apprentice/Trainee at other Firm	**0.0074**	**-0.0164**
	(0.0155)	(0.0488)
Number of Days in Labor Market Status 8 Days before current match (Main Effect)	**-0.000001**	**0.000009*** **
	(0.000001)	(0.000002)
…if previous spell was benefits or gap (Interaction)	**0.000001**	**0.000003**
	(0.000003)	(0.000008)
…if previous spell was training (Interaction)	**0.000016**	**0.000009**
	(0.000019)	(0.000058)
Female	**-0.0529**	**-0.376*** **
	(0.0433)	(0.1055)
Number of Days of Benefit Receipt up to beginning of current match	**0.000007***	**-0.000082*** **
	(0.000003)	(0.000009)
Years since first Employment at beginning of current match	**0.0024** **	**-0.0046***
	(0.0009)	(0.0022)
…interacted with female dummy	**-0.0024**	**-0.0132*** **
	(0.0014)	(0.0036)
Years since first employment squared	**-0.000129*** **	**-0.000058**
	(0.00003)	(0.000077)
…interacted with female dummy	**0.000082**	**0.000457*** **
	(0.000048)	(0.000131)
Age at beginning of current match	**0.0024**	**0.0364*** **
	(0.0015)	(0.0038)
…interacted with female dummy	**0.0043**	**0.0221*** **
	(0.0025)	(0.0059)
Age squared	**-0.00002**	**-0.000387*** **
	(0.000019)	(0.000048)
Match Count		
2	**-0.0063**	**0.0612*** **
	(0.0043)	(0.0117)
…interacted with female dummy	**0.0071**	**0.0027**
	(0.0065)	(0.0207)
3	**-0.0229** **	**0.0958*** **
	(0.0087)	(0.0236)
…interacted with female dummy	**0.0208**	**-0.0253**
	(0.0142)	(0.052)
4	**-0.0576***	**0.1722*** **
	(0.0252)	(0.0401)
…interacted with female dummy	**0.0807**	**0.1501**
	(0.0562)	(0.1395)
5	**0.1193*** **	**-0.05**
	(0.0188)	(0.0573)
…interacted with female dummy	**-0.0815**	**0.1983**
	(0.0757)	(0.1426)
6	**-0.2742*** **	**0.2965*** **
	(0.0055)	(0.0263)
…interacted with female dummy	**0.0918*** **	**-0.3977*** **
	(0.0086)	(0.0315)
Constant	**-0.0565***	**3.7472*** **
	(0.0258)	(0.0691)
Number of Observations	665080	665080
R-squared	0.0343	0.2189

Omitted Categories: *Emp. Status 8 days before current match:* Employment at other Firm; *Year Match Started:* 1993; *Match Count:* 1
Note: Estimates of Match and Firm Fixed Effects are from main regression on full sample including actual experience. Regression also includes dummies for the year the match started, but coefficients are not reported. Significance levels: *: Significant at 5%; **: Significant at 1%; ***: Significant at 0.1%.

Imprint

FDZ-Methodenreport 01/2012

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Iris Dieterich

Technical production
Iris Dieterich

Download
http://doku.iab.de/fdz/reporte/2012/MR_01-12_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:

Nikolas Mittag,
University of Chicago,
1155 East 60th Street,
Chicago, IL 60637
USA

Email: mittag@uchicago.edu