HEDG Working Paper 09/35

The effects of expanding the generosity of the statutory sickness insurance system

Nicolas R. Ziebarth
Martin Karlsson

December 2009

york.ac.uk/res/herc/hedgwp

The effects of expanding the generosity of the statutory sickness insurance system[‡]

Nicolas R. Ziebarth
SOEP at DIW Berlin and TU Berlin[*]

Martin Karlsson
TU Darmstadt[**]

October 23, 2009

[‡]We would like to thank Daniela Andrén, Mattias Bokenblom, Jörg Breitung, Daniel S. Hamermesh, Lars Hultkrantz, Per Johansson, Michael Lechner, Martin Olsson, Mårten Palme, Per Petterson-Lidbom and participants at various seminars for their helpful comments and discussions. We are responsible for all remaining errors and shortcomings of the article. Nicolas R. Ziebarth gratefully acknowledges support from the *Stiftung der Deutschen Wirtschaft* (sdw, Foundation of German Business) in the form of a scholarship.

[*]Corresponding author: Nicolas R. Ziebarth, German Institute for Economic Research (DIW) Berlin, Socio-Economic Panel Study (SOEP), Mohrenstraße 58, 10117 Berlin, Germany, and Berlin University of Technology (TU Berlin), phone: +49-(0)30-89789-587, fax: +49-(0)30-89789-109, `e-mail: nziebarth@diw.de`

[**]Darmstadt University of Technology (TU Darmstadt), Marktplatz 15 - Residenzschloss, `e-mail: karlsson@vwl.tu-darmstadt.de`

Abstract

In 1999, in Germany, the statutory sick pay level was increased from 80 to 100 percent of foregone earnings for sickness episodes of up to six weeks. We show that this reform has led to an increase in average absence days of about 10 percent or one additional day per employee, per year. The estimates are based on SOEP survey data and parametric, nonparametric, and combined matching-regression difference-in-differences methods. Extended calculations suggest that the reform might have increased labor costs by about €1.8 billion per year and might have led to the loss of around 50,000 jobs.

Keywords: sickness absence, statutory sick pay, natural experiment, Socio-Economic Panel Study (SOEP)

JEL classification: H51; I18; J22

1 Introduction

Workplace absences are hugely important for labor costs, labor productivity and population health, and place a burden on most social security systems. At the same time, we observe huge differences in the average absence days per year and per employee across OECD countries. The variation ranges from 4 to 29 days (OECD, 2006), suggesting that institutional settings and cultural peculiarities matter. In light of this, it seems odd that labor and health economists have devoted little attention to the relationship between sick leave benefits and sickness absence. This fact is clearly demonstrated when we consider the vast number of studies on the relationship between unemployment benefits and unemployment duration.

One reason for this paucity of research might be confusion about cross-country differences between institutions. Typically, in Europe, federal laws insure employees, to a varying extent, against income losses due to workplace absences. Sickness absence insurance and legally-fixed sick pay levels are part of the social legislation and insure employees against work-related *and* non-work-related sickness absence. Depending on the country, either the employer or the tax payer funds sick leave payments. Usually, employers provide short-term sick pay and, through social insurance, taxpayers finance long-term absences resulting from sickness or disability.

The US only know one main sickness absence insurance program at the federal level. Disability Insurance (DI) compensates employees for income losses as a result of long-term sickness or permanent disability. In contrast, the "workers' compensation insurance" solely covers *work-related* absences and is administered on a state-by-state basis. What Europeans call "sickness absence insurance" is, in the US, referred to as "temporary disability insurance" or "cash sickness benefits" and covers absence from work due to temporary *non*-work-related sickness or injury. Five US states, among them the most populous state of California, have such insurance programs. Their relevance is illustrated by the fact that, in California in 2005, the total sum of net benefits for temporary disability insurance amounted to \$ 4.2 billion while the total sum for unemployment insurance amounted to \$ 4.6 billion (Social Security Administration, 2006, 2008).

As yet, only a handful of studies have empirically analyzed the relationship between the level of sick pay and non-work-related sickness absence. All of them come from Sweden and find that employees adapt their sickness absence behavior to legislative changes in the level of statutory sick pay (Johansson and Palme, 1996, 2002, 2005; Henrekson and Persson, 2004; Pettersson-Lidbom and Skogman Thoursie, 2008). Apart from the Swedish studies, there are also two papers providing correlation-based evidence with English data from the 1970s (Doherty, 1979; Fenn, 1981). Two further studies are based on Canadian data and examine the effects of an increase in the Disability Insurance benefit levels in

1987 on the labor market participation of older male workers. The outcomes of these two studies are contradictory. Finally, two papers from the US analyze the impact of changes in the benefit levels of workers' compensation insurance. While Meyer et al. (1995) find that an increase in benefits in 1987 has led to increased injury duration, the results from the Curington (1994) study using data from the 1960s and 1970s are mixed. It also has been found that the generosity of the disability insurance affects labor supply decisions on the extensive margin (Bound, 1989; Gruber, 2000; Chen and van der Klaauw, 2008), although there is also evidence that this might not alway be the case (Campolieti, 2004).

With this study, we expand the scarce literature on the effects of sick leave on sickness-related workplace absences. The estimates draw on a well-defined control group and the underlying reform is the most recent in the relevant literature. We make use of SOEP data that are representative for the most populous European country and provide a rich set of background information, including health status which is the most important determinant of sickness absences. Exploiting panel data enables us to tackle issues like selection into the treatment and panel attrition. We also attempt to relate reform-induced changes in workplace absences to changes in labor costs and job creation. Finally, unlike all other studies, we do not only rely on parametric methods but also make use of matching methods, as well as the most sophisticated evaluation methods combining matching and regression (Imbens and Wooldridge, 2009).

In Germany, on January 1, 1999, the statutory sick pay level for sickness episodes of up to six weeks was raised from 80 to 100 percent of foregone gross wages. This law applies to all private sector employees. Comparing private sector employees to those who were clearly not affected by the reform, namely self-employed and public sector employees, we show that this measure has increased the average annual number of absence days by about 10 percent or one day. Since employers are obliged to provide sick pay for the first six weeks and since we know the gross wage of the respondents, we are able to roughly calculate the reform-induced increase in labor costs. We estimate that employers have to carry the burden of extra costs amounting to around €1.8 billion per year as a consequence of the reform. Using estimates from other studies for Germany which were conducted, at the end of the 90s, by means of macroeconomic simulation models, a back-of-the-envelope calculation suggests that between 40,000 and 80,000 jobs might have been lost due to the reform.

Section 2 explains Germany's sickness absence insurance system. Section 3 gives more details about the data, this is followed by Section 4 which discusses the empirical estimation strategies. We then estimate the reform's impact on sickness absence and calculate the increase in labor costs which might have led to job losses. Section 6 concludes.

2 The German Sick Pay Scheme and Policy Reform

2.1 The Sick Pay Scheme and Monitoring System

Before the implementation of the new law, every German employer in the private sector was legally obliged to continue to pay 80 percent of usual wages for up to six weeks per sickness episode.[1] Obviously, the self-employed are not eligible for employer-provided sick pay. Public sector employees and apprentices were constantly guaranteed 100 percent sick pay for up to six weeks per sickness spell. Henceforth, we use the term *short-term sick pay* as a synonym for employer-provided sick pay and *short-term absenteeism* as a synonym for absence spells of less than six weeks.

In the case of illness, employees are obliged to inform their employer immediately about both the sickness and the expected duration. From the fourth day of a sickness episode, a doctor's certificate is required and is usually issued for up to one week, depending on the illness. If the sickness lasts more than six continuous weeks, the doctor must issue a different certificate. From the seventh week onwards, sick pay is disbursed by the sickness fund and lowered to 70 percent of foregone gross wages for those who are insured under the Statutory Health Insurance (SHI).

The monitoring system mainly consists of an institution called *Medical Service of the SHI*. One of the original objectives of the Medical Service is to monitor sickness absence. German social legislation codifies that the SHI is obliged to contact the Medical Service and request a medical opinion to clarify any doubts about work absences. Such doubts may arise if the insured person is short-term absent with unusual frequency or is regularly sick on Mondays or Fridays. Similarly, if doctors certify sickness with unusual frequency, the SHI may ask for expert advice. The employer also has the right to request the assistance of the Medical Service and ask for expert advice. Expert advice is based on available medical documents, information about the workplace, and a statement which is requested from the patient. If necessary, the Medical Service has the right to conduct a physical examination of the patient and to cut their benefits.[2] In 2007, about 2,000 full-time equivalent and independent doctors worked for the medical service and examined 1.7 million cases of absenteeism (Medizinischer Dienst der Krankenversicherung (MDK), 2008).

[1] The entitlement is codified in the so called *Gesetz über die Zahlung des Arbeitsentgelts an Feiertagen und im Krankheitsfall (Entgeltfortzahlungsgesetz)*, article 3, 4. Sick pay is only provided for regular earnings and not for overtime payments.

[2] The wordings of the laws can be found in the Social Code Book V, article 275, article 276.

2.2 The Policy Reform

In their election campaign in 1998, the Social Democrats and the Greens promised to increase statutory sick pay from 80 to 100 percent of foregone gross wages should they form the new government. The announcement was a mechanical reaction to a former cut in sick pay, implemented by the previous Conservative government under Chancellor Kohl in October 1996. At that time, together with a reduction of long-term sick pay, short-term sick pay was decreased from 100 to 80 percent of forgone gross wages. Ziebarth and Karlsson (2009) analyze the effect of the cut in short-term sick pay and find that it reduced the average number of absence days by about 5 percent.[3] The cut in sick pay was perceived as unfair and unsocially unjust by the majority of the Germans and was accompanied by many strikes. After their successful election in September 1998, the new left-wing government immediately passed a law which became effective from January 1, 1999, increasing statutory short-term sick pay from 80 to 100 percent of foregone gross wages.[4]

Although statutory sick pay was increased, not every employee in the private sector was affected by this increase since employers may voluntarily provide sick pay over and above the minimum requirements. In Germany, employers from various branches agreed upon a 100 percent sick leave compensation in collective wage agreements. There are no official figures as to how many employees took advantage of these bonus payments when they were sick; in 1998, union leaders proudly declared that 13 out of 27 million employees would receive 100 percent sick pay (Jahn, 1998).[5] In 1997, a poll among handcraft establishments revealed that 51 percent of the companies provided 100 percent sick pay voluntarily, probably due to strong mutual trust between employers and employees in these small companies (Ridinger, 1997). As in all other studies that have evaluated the impact of changes in statutory benefit levels, we assess the overall impact of the law among private sector employees comparing them with totally unaffected groups such as the self-employed and public sector employees.

Already before the reform, Germany's sickness absence insurance was one of the most generous worldwide. In 1998, the total sum of employer-provided sick pay amounted to DM 44.8 billion (€ 22.3 billion) (German Federal Statistical Office, 2001). At that time, there was a consensus among German economists that high (non-wage) labor costs – in

[3] For various reasons, it makes sense to analyze the effects of a generosity contraction separetly from the effects of a generosity expansion: First, one may expect that the effects differ. Second and more importantly, in this case, the cut in short-term sick pay was accompanied by a cut in long-term sick pay that serves as a confounding factor in the estimation. Consequently, the treatment and control groups differ in all reforms which translates into different identification strategies.

[4] Passed on December 19, 1998 this law is the *Gesetz zu Korrekturen in der Sozialversicherung und zur Sicherung der Arbeitnehmerrechte, BGBl.I 1998 Nr. 85 S.3843-3852.*

[5] Both figures include around 3.3 million employees in the public sector (German Federal Statistical Office, 1999).

Germany total labor costs per hour were ranked among the top among OECD countries – were one of the main reasons for the persistently high unemployment rate in Germany. Since sick leave payments are non-wage labor costs and function like a tax on labor, the German Council of Economic Advisors disagreed with the increase of the minimum sick pay level, predicted job losses, and warned about a new obstacle to job creation (Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung, 1998).

3 Data and Variable Definitions

We rely on the German Socio-Economic Panel Study (SOEP) for the empirical analyses. Apart from the SOEP, there is no other data set that includes representative information on sickness absence in Germany. The SOEP is a 25-year-old household panel survey with a focus on labor market activities and earnings. Further details can be found in Wagner et al. (2007).

For the main specifications, two pre-reform and two post-reform years are exploited, i.e., we rely on sickness absence information for the years 1997 to 2000.[6] We restrict our working sample to respondents who are full-time employed and who are between 25 and 55 years of age. We do not use respondents with item non response on relevant variables.

3.1 Sick Leave Measure and Covariates

The SOEP offers detailed information about employment histories, job characteristics, the type of job, and the various income sources. Information on self-assessed health, medical care usage, and the number of sick leave days is likewise sampled.

We call the dependent variable *Daysabs* and generate this count measure one-to-one from the answers of the following question: *"How many days off work did you have in 19XX [200X] due to illness? Please enter all days, not just those for which you had a doctor's certificate."* Relying on self-reported information rather than administrative data has both caveats and benefits.

Clearly, the issue of measurement errors is an important one. The more sickness periods a respondent had in the previous year, the larger the probable recall bias. However, although measurement errors inflate standard errors and lead to less precise estimates, they would only seriously hamper our analysis if the reform had any impact on them.

[6] Since current as well as retrospective information is sampled in every wave, we match the retrospective information which we are interested in with the current information of the corresponding year as long as the respondent was interviewed in both years. If this was not the case, we use both types of information from the same interview and assume that the current statements have not changed since the last year.

This is very unlikely since the reform of statutory sick pay was the subject of political and media debate during the whole period under consideration.

On the other hand, the big advantage of self-reported over administrative data is that it provides a measure of the *total* number of absent days. With register data, we face the issue that solely doctor-certified absence periods are covered. Moreover, employer-provided sick leave periods are often not recorded. This almost always leaves the researcher with censored data and makes certain types of analyses intangible.[7] However, having an uncensored measure of the total number of absence days based on survey data comes at the cost of not having detailed spell data.

The whole set of explanatory variables can be found in Appendix A and is categorized as follows: The first group incorporates variables on personal characteristics, like the dummies *Female, Immigrant, East German, Partner, Married, Children, Disabled, Good health, Bad health, No sports*, and *Age (Age*^{2}*)*. The second group consists of educational controls such as the degree obtained, the number of years with the company, and whether the person was trained for the job. The last group contains explanatory variables on job characteristics. Among them are *Blue-collar worker, White-collar worker*, the size of the company, or *Gross wage per month*. Apart from including various interaction terms between these covariates and *years with company* as well as *Gross wage per month*, we also control for the annual state unemployment rate. In the parametric approaches, state dummies net out permanent differences across states and year dummies take account of common time shocks.

3.2 Treatment and Control Group

The treatment sample consists of all private sector employees except apprentices. The control sample incorporates public sector employees, apprentices, and the self-employed – all those who did not experience a change in their sick pay levels during the period under consideration. The dummy *Treatment Group* has a one for the treated and a zero for the controls. In total, we have 15,402 observations in the treatment group and 8,024 observations in the control group.

[7] For instance, take the case of Sweden and the impact of changes in the waiting period: before 1987, there was a waiting period with zero compensation during the first day of a sickness spell. During the 1990s, the waiting period and the employer-provided sick pay period were changed several times, generating a register base which is censored while the censoring varies with the reforms (see Henrekson and Persson (2004) for more details). In addition to the absence of a natural control group, this makes it difficult to identify causal effects in the case of Sweden. Interestingly, as has been mentioned in Section 1, hitherto, all studies on the relationship between sick pay levels and short-term overall sickness absence come from Sweden.

4 Estimation Strategy and Identification

4.1 Regression

OLS

We start by estimating conventional parametric difference-in-differences (DiD) models. Consider the following equation:

$$y_{it} = \lambda p99_t + \pi D_{it} + \theta DiD_{it} + s'_{it}\psi + \rho_t + \phi_s + \epsilon_{it} \tag{1}$$

where y_{it} stands for the annual number of absence days of individual i in year t, $p99_t$ is a post-reform dummy, D_{it} is a treatment group dummy, and DiD_{it} is the regressor of interest. It has a one for treated respondents in post-reform years and gives us the causal reform effect should certain assumptions hold. It can also be interpreted as the interaction term between the treatment group dummy and the post-reform dummy. By including additional time dummies ρ_t we control for common time shocks that might affect sick leave. State dummies ϕ_s account for permanent differences across the 16 German states along with the annual state unemployment rate that controls for changes in the tightness of the regional labor market and that is included in the $K \times 1$ column vector s'_{it}. The other $K-1$ regressors are made up of personal controls including health status, educational controls, and job-related controls as explained in Section 3.1. In addition to the covariates displayed in Appendix A, we also include various interaction terms. As usual, ϵ_{it} stands for unobserved heterogeneity. To begin with, equation (1) is estimated by OLS.

Zero-Inflated NegBin-2 (ZINB-2)

The number of absent days is a highly skewed count variable with excess zero observations (about 50 percent of the sample) and overdispersion, i.e., the conditional variance exceeding the conditional mean. Hence, it seems to be appropiate to fit count data models which might capture the skewed distribution better than simple OLS. Based on the Akaike (AIC) and Bayesian (BIC) information criteria and various Vuong tests, we found the so called *Zero-Inflated Negative Binominal Model (NegBin)* to be appropriate for our purposes.

The underlying statistical process differentiates between absent employees and non-absent employees and assigns different probabilities, which are parameterized as functions of the covariates, to each group. The binary process is specified in form of a logit model and the count process is modeled as an untruncated NegBin-2 model for the binary process to take on value one. Hence, zero counts may be generated in two ways: as realizations of the binary process and as realizations of the count process when the binary process is one (Winkelmann, 2008). In contrast to the more restrictive Poisson distribution, the employed negative binomial distribution does not only take excess zeros into account but also allows for overdispersion and unobserved heterogeneity.[8] The NegBin model can be seen as a special case of a continuous mixture model. In the notation of Cameron and Trivedi (2005), the NegBin distribution can be described as a density mixture of the following form:

$$\begin{aligned}
\varphi(y|\mu,\alpha) &= \int f\,(y|\mu,\nu) \times \gamma(\nu|\alpha)\, d\nu \\
&= \int_0^\infty \left(\frac{e^{-\mu\nu}\{\mu\nu\}^y}{y!}\right)\left(\frac{\nu^{\delta-1}e^{-\nu\delta}\delta^\delta}{\Gamma(\delta)}\right) d\nu \\
&= \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y} \qquad (2)
\end{aligned}$$

where $f\,(y|\mu,\nu)$ is the conditional poisson distribution and $\gamma(\nu|\alpha)$ is assumed to be gamma distributed with ν as an unobserved parameter with variance α. Note that in the special case of $\alpha = 0$ the NegBin collapses to a simple Poisson model. $\Gamma(.)$ denotes the gamma integral and $\mu = exp(x'_{it}\beta)$ with x'_{it} incorporating all the regressors as in equation 1.

The marginal effect of the interaction term DiD_{it} is – given the model assumptions are fulfilled – the causal reform effect and is, henceforth, always displayed when output tables are presented.[9]

4.2 Matching

A fundamental alternative to parametric estimation is matching. In principle, matching intends to make treatment and control observations more comparable by assigning each treated unit one or more control units that are similar in terms of observable characteristics. Under the "conditional independence" or "unconfoundedness" assumption which claims that, after having conditioned on observables, the treatment is independent of the outcome, the assignment to treatment can be interpreted as random as if it were generated by a randomized experiment (LaLonde, 1986). Various matching methods exist.

[8] The unobserved heterogeneity allowed for in the NegBin-2 is based on functional form and does not capture unobserved heterogeneity which is correlated with explanatory variables.

[9] Puhani (2008) has shown that the advice of Ai and Norton (2004) to compute the discrete double difference is not of relevance in nonlinear models when the interest lies in the estimation of a treatment effect. The average treatment effect on the treated at the time of the treatment is given by $\varphi(y|\alpha,\bar{s}_1,p99 = 1, D = 1, DiD = 1) - \varphi(y|\alpha,\bar{s}_1,p99 = 1, D = 1, DiD = 0)$, where $\bar{s}_1$ denotes the average values of the covariates for the treatment group in the post-treatment period. This is exactly what we calculate and present throughout this paper.

Most applied matching analyses build upon propensity score matching which avoids the "curse of dimensionality." The curse refers to the fact that matching requires, on the one hand, a sufficiently large number of control variables to make the two samples comparable. On the other hand, finding a sufficiently large number of pairs that are comparable in their observable characteristics becomes increasingly difficult, the larger the number of covariates. Rosenbaum and Rubin (1983) have shown that conditioning on the propensity score – the probability to get treated – is equivalent to selecting pairs of treated and controls based on every single covariate dimension, given unconfoundedness holds. We estimate the propensity score (PS) by means of a logit model and select the covariates to be included out of the total number of covariates (Appendix A) using likelihood ratio tests on zero coefficients – in a first step for control variables in levels and in a second step for their interactions (Imbens, 2008).

Besides a plausible selection-on-observables story, matching requires that the distributions of the covariates for treated and controls overlap to a large extent. In this setting, the common support assumption is fulfilled as can be seen in Figure 6. The PS distribution for both groups shows a large overlap with the region of common support lying between PS values of 0.05 and 0.92.

[Insert Figure 1 about here]

The first method that we employ is stratification matching or blocking. Based on the probability of getting treated, i.e. of belonging to the treatment group in post-reform years, the sample is cut into blocks in such a way that the propensity score for treated and controls is balanced within each block. Then, block-by-block average treatment effects on the treated are obtained by the taking the difference between the average outcome for treated and controls block-by-block . Afterwards, the overall treatment effect on the treated can be computed as the weighted average of the block-by-block treatment effects (Rosenbaum and Rubin, 1984). Cochran (1968) has shown that five blocks are sufficient to reduce the bias that is associated with the overall simple outcome difference between treated and nontreated by more than 95 percent.

The second method is one-to-one nearest neighbor matching with replacement. Based on the propensity score to get treated, the most similar control observation is assigned to every treated observation.[10] Then, the outcome difference of each pair is taken to compute the average treatment effect on the treated (Lechner, 2002).

The third method that we use is kernel matching. Again, pairs are built and averages are taken of the outcomes of these pairs. However, the controls which are assigned to

[10] If there are two control observations with an identical propensity score, both observations are given equal weight.

every treated observation are obtained by a kernel weighted average of various controls (Heckman et al., 1998).

4.3 Combining Regression and Matching

Both regression and matching methods have drawbacks. If treated and control units differ substantially in their observed characteristics, then parametric approaches use the covariate distribution of the controls to make out-of-sample predictions. More specifically, consider the predicted outcome for treated respondents if they were not treated (Imbens and Wooldridge, 2009); the indices i and t are omitted for notational purposes:

$$\widehat{E}(y^1|DiD = 0) = \bar{y}_0 + \hat{\psi}_0'(\bar{s}_1 - \bar{s}_0) \tag{3}$$

where y^1 denotes the potential outcome if person i was affected by the reform and $\bar{y}_0$ stands for the average outcome among the non-treated. As before, s is a vector of covariates, with $\bar{s}_1$ denoting the average covariate values for the treated. The regression parameter $\hat{\psi}_0$ stems from the non-treated. We can infer from the equation that $\hat{\psi}_0$ is used to predict outcomes for treated respondents. If the covariates' means of treated and non-treated differ significantly, then the model predicts out of the sample upon which the estimation is based and yields sensitive results.

Imbens and Rubin (2009) propose to judge the differences in covariates for treated and controls by the scale-free normalized difference:

$$\Delta s = \frac{\bar{s}_1 - \bar{s}_0}{\sqrt{\sigma_1^2 + \sigma_0^2}} \tag{4}$$

with σ_1^2 being the variance for the treated. As a rule of thumb, a normalized difference exceeding 0.25 is likely to lead to sensitive results.

Applied to our case, firstly, we look at how the covariate distribution for the treatment group differs in comparison to the control group, i.e., we compare private sector employees to those whose sick pay was not affected throughout the whole period under consideration. Table 1 shows in column (1) the means of the covariates for the treatment group and in column (2) the means of the covariates for the control group. At first glance, it appears as though the two groups are very similar with respect to their observable characteristics. This presumption is partly reinforced by column (3) which displays the normalized difference. Indeed, most of the values are substantially smaller than 0.25 and some tend towards zero. Only one regressor, job autonomy, has a value larger than 0.25.

[Insert Table 1 about here]

Columns (4) to (6) again show the covariates' for the treatment and control group as well as the normalized difference. The underlying sample was stratified based on the propensity score of belonging to the treatment or the control group. It is easy to see that blocking substantially improves the balance of the covariates between the treatment and control group. The normalized difference for all covariates decreases substantially as compared to the raw sample and yields values of less than 0.1 for almost all regressors. Finally, we perform one-to-one nearest neighbor matching on the probability of belonging to the treatment group and display the values for the corresponding sample in columns (7) to (9). Now the largest normalized difference is 0.05 and the covariates are very well balanced between the treatment and control group. Since the last sample has the best balancing properties, in the second step of the process, we apply the three matching methods which were described above to this subsample.

However, even in the matched sample, small differences between the two groups still remain. These differences may lead to biased matching estimators. In addition, Abadie and Imbens (2007) have shown that the simple nearest neighbor matching estimator includes a bias term which leads to inconsistencies and should be corrected for.

Imbens and Wooldridge (2009) propose two approaches that both combine the strengths of regression and matching estimators which both work well in practice and that both lead to robust results. Approach number one is a combined blocking and regression approach. In the first step, based on the probability of belongingto the treatment group, the sample is stratified into blocks in such a way that the treatment and control groups have equal propensity scores (cf. the blocked sample in columns (3) to (5) of Table 1). In the second step, parametric regressions – as detailed in Section 4.1 – are run within each block. Then the within-block treatment effects are weighted by the relative size of the blocks and aggregated into an overall average treatment effect on the treated. The crucial point is that the covariate distributions within each stratum are very similar and, thus, out-of-sample predictions are avoided, as equation (3) shows. Imbens and Wooldridge (2009) call this approach "considerably more flexible and robust than either subclassification alone, or regression alone." Moreover, according to them, it is "probably one of the more attractive estimators in practice."

The second approach also aims to smooth differences in covariates between treated and controls and additionally corrects for the bias described in Abadie and Imbens (2007). It combines regression and nearest neighbor matching. In the first step, using only the non-treated who were matched to the treated, a linear regression of the outcome on the covariates is carried out.[11] Then, in the second step, the counterfactual potential outcome

[11]Here, a linear model is used. However, various specifications are perceivable.

for the case without treatment, y^0, is calculated as (Abadie et al., 2004):

$$\hat{y}_i^0 = \begin{cases} y_i \text{ if DiD} = 0 \\ \frac{1}{M}\sum_{j\epsilon\Gamma_M}(y_j + \hat{\psi}_0'(s_i - s_j)) \text{ if DiD} = 1 \end{cases}$$

where Γ_M denotes the set of indices for the M closest matches for unit i, and y_j is the outcome which is matched to unit i. According to Abadie and Imbens (2007), the method yields a robust estimator that works well in practice.

$$\hat{y}_i^0 = \begin{cases} y_i \text{ if DiD} = 0 \\ \frac{1}{M}\sum_{j\epsilon\Gamma_M}(y_j + \hat{\psi}_0'(s_i - s_j)) \text{ if DiD} = 1 \end{cases}$$

where Γ_M denotes the set of indices for the M closest matches for unit i, and y_j is the outcome which is matched to unit i. According to Abadie and Imbens (2007), the method yields a robust estimator that works well in practice.

4.4 Identification

In the previous subsections, we discussed in detail how we took advantage of the rich set of socioeconomic background information to make the treatment and control group as comparable as possible. As can be seen in columns (6) and (9) of Table 1, blocking and matching yields two samples that are almost identical in terms of their observable characteristics.

However, the crucial identifying assumption in any DiD analysis claims that all relative post-reform changes in the outcome variable of the treated can be traced back to the reform. In other words, it is assumed that – conditional on all personal, educational, and job characteristics as well as time and year dummies – there are no unobservables that impact the dynamic of the outcome differently for both groups. This common time trend assumption is not directly testable. However, we believe that it is very likely to hold in our context. Firstly, we are analyzing a reform that applied to a large and well-defined group in the labor market – private sector employees. The reform was implemented at the federal level and unambiguously reduced the cost of workplace absences, an outcome which we are able to observe directly. Secondly, the selection on observables argument is very plausible in this setting since treatment is exclusively defined by job characteristics and since we observe all job characteristics that determine treatment. Moreover, as we will explain below, we can exclude the possibility that selection into or out of the treatment distorted our results since we have a panel and can identify job changers. Thirdly, unlike all other studies that we are aware of, we are not only able to control for all treatment relevant job characteristics but also for a rich set of socioeconomic controls and (self-reported) health status which is, by far, the most important determinant of sickness absence.[12]

Fourthly, we also include the lagged number of absence days in one specification. Fifthly, we can test the robustness of our results with respect to sample composition changes over time and labor market attrition. For example, in one robustness check, we weight the parametric regressions with the inverse probability that a respondent, whom we observed as working in the pre-treatment period, will answer the questionnaire without item-non response in the post-treatment period. In another robustness check, we shorten the period under consideration to one pre- and one post-reform year and balance the sample. Sixthly, since it is possible to indirectly test the plausibility of the common time trend assumption, we present the results of placebo regressions. Placebo regressions assume that the reform which is analyzed took place in a year without a reform. Should the coefficient of interest turn out to be significant in a non-reform year, the common time trend assumption would be seriously challenged. Finally, in Figure 2 we display the average number of predicted absence days for pre- and post-reform years and both groups. A clear parallel trend in the pre-reform years is identifiable. From 1998 to 1999, a jump in the average absence days for the treated can be observed, whereas the mean rate for the controls shows a smooth pattern. In the post-reform years, again we observe a parallel evolution of the mean number of absence days for both groups.

[Insert Figure 2 about here]

In recent years, the drawbacks and limitations of DiD estimation has been extensively debated. A particular concern is the underestimation of OLS standard errors due to serial correlation in the case of long time horizons as well as unobserved (treatment and control) group effects. To deal with the serial correlation issue, we focus on short time horizons. As detailed in Bertrand et al. (2004), the main source for understating the standard errors stems from serial correlation of the outcome and the intervention variable. Their findings suggest that the serial correlation issue is, for the most part, eliminated when focusing on less than five periods. While there is consensus about the serial correlation problem, discussions on the issue of unobserved common group effects are more controversial. If one takes the objection of Donald and Lang (2007) seriously, then it would not be possible to draw inferences from DiD analyses in the case of few groups, meaning that no empirical assessment could be performed. We subscribe to the view of Wooldridge (2006) who refers

[12] For the sake of saving space, we do not report detailed results of a regression of *Daysabs* on all covariates in Appendix A. However, the results are in line with the literature on sickness absence. The results are available from the authors upon request.

to it as (p. 18): *"DL [Donald and Lang] criticize Card and Krueger (1994) for comparing mean wage changes of fast-food workers across two states because Card and Krueger fail to account for the state effect (New Jersey or Pennsylvania) [...]. But the DL criticism in the $G = 2$ case is no different from a common question raised for any difference-in-differences analyses: How can we be sure that any observed difference in means is due entirely to the policy change? To characterize the problem as failing to account for an unobserved group effect is not necessarily helpful."*[13] Apart from our focus on short time spans to resolve serial correlation concerns, we use robust standard errors and correct for clustering at the individual level throughout the analysis.

One of the biggest issues in evaluation studies is selection effects. Here the reform was politically determined and the law applied to all private sector firms. It is perceivable that in response to the reform, public sector employees and the self-employed applied for jobs in the private sector since the conditions in the private sector improved. It is also perceivable that the increase in sick pay induced more non-working people to accept job offers in the private market. In any case, since we have a panel dataset and detailed job information, we can tackle such selection concerns.

As already mentioned above, some employers agreed in collective bargaining to pay 100 percent sick pay voluntarily. We cannot identify employees who underlay such collective wage agreements. Since employers are always free to provide additional payments on top of statutory regulations, it is intrinsically difficult to identify all contractually fixed payments. In contrast to other countries, like Sweden, where differences in the labor agreements are more fragmented, polls for Germany at the time of the reform suggest that around half of all employees received the statutory sick pay and half of all employees were provided 100 percent sick pay (Ridinger, 1997; Jahn, 1998). In addition, all studies in this strand of the literature evaluate the *overall* effects of changes in statutory sick pay and do not claim to estimate precise elasticities. We take the same approach.

5 Empirical Estimation Results

A detailed discussion on the identification strategy can be found in the previous section. The same section also includes a description of the estimation methods and the methods

[13] In this very readable extended version of an older published AER paper (Wooldridge, 2003), Wooldridge (2006) discusses several other shortcomings of and assumptions regarding the estimation approach proposed by Donald and Lang (2007). Elsewhere, Wooldridge (2007) asks rhetorically whether introducing more than sampling error into DiD analyses was necessary, or desirable. *"Should we conclude nothing can be learned in such settings?"*, he asks(p. 3). Moreover, he uses the well known Meyer et al. (1995) study, which is similar to ours and obtains marginally significant results, as another example: *"It seems that, in this example, there is plenty of uncertainty in estimation, and one cannot obtain a tight estimate without a fairly large sample size. It is unclear what we gain by concluding that, because we are just identifying the parameters, we cannot perform inference in such cases. In this example, it is hard to argue that the uncertainty associated with choosing low earners within the same state and time period as the control group somehow swamps the sampling error in the sample means."* (p. 3 to 4).

that we have employed to balance the covariate distributions of treatment and control groups as detailed in Table 1. Moreover, in the context of the discussion on identification, we have outlined several arguments as to why the common time trend assumption is likely to hold in this setting. A visual representation of one argument is displayed in Figure 2 where we see parallel pre- and post-reform time trends of the outcome variable but an increase in the absence rate at the time of the reform for the treatment group. The following subsections deal with the difference-in-differences estimation results and the robustness checks.

5.1 Baseline Specifications

We start by estimating parametric OLS and ZINB-2 regressions on the raw sample with all available covariates (Appendix A) included. In the following we always display marginal effects. The DiD estimates are displayed in columns (1) and (2) of Table 2. The OLS model yields an estimate of 1.278 that is significant at the 6.8 percent level. The ZINB-2 model gives an estimate of 0.925 with a standard error of 0.472. The unconditional double difference of the means of the two groups for the two time periods is 1.378 (std. err. 0.722; not shown) and very close to the OLS estimate in column (1).

[Insert Table 2 about here]

Columns (3) to (5) give the results when three different matching variants are conducted using the matched subsample of Table 1. In the first step, the matched subsample of Table 1 was obtained by one-to-one nearest neighbor matching on the probability of belonging to the treatment group vs. the control group. As can be seen in the last column of Table 1, the covariate distribution is almost perfectly balanced between the treatment and control group in this sample. In the second step, we perform the three matching methods as described in the methods' section. In this respect we could also call our nonparametric applications "two-step matching procedures." While we performed matching on the probability of belonging to the treatment group to obtain the matched sample, we now perform matching on the probability of getting treated, i.e. of belonging to the treatment group in the post-reform period, using the matched sample.

In column (3), the sample is stratified into blocks based on the probability of getting treated which is obtained by another propensity score (PS) logit regression. Then, by taking the average values of treated and non-treated within each block, the block-specific reform effects are calculated, which are finally aggregated to a weighted overall average. This method gives an estimate of 1.148 with a standard error of 0.421. Column (4) yields the estimate for one-to-one nearest neighbor matching based on the PS of belonging to the treatment group in the post-reform period. The estimated reform effect is 1.247 (std. err. 0.561). Column (5) displays the kernel matching estimate which is 0.8729 and has a standard error of 0.421. Kernel matching is similar to one-to-one nearest neighbor matching but assigns more than one control unit to each treated unit and weights these controls by a kernel.

According to Imbens and Wooldridge (2009), the most suited methods in use combine regression and matching and are thus more flexible and robust than other methods. Column (6) shows the result when the sample is first stratified on the probability of belonging to the treatment group (hence it makes use of the blocked sample in Table 1) and then blockwise regressions as in equation 1 are run. The overall treatment effect, which is 0.9829 and significantly different from zero, is obtained as an average of the within block estimates, weighted by the block size. The method used in the last column also combines matching and regression and eliminates a bias that has been proven to exist for nearest neighbor matching. More details can be found in Section 4.3 and Abadie and Imbens (2007). The resulting estimate is very similar to the one in column (6) and yields a reform effect of 1.031 (std. err. 0.4355).

It is remarkable that all estimates differ only slightly in size and that all point estimates are statistically different from zero with the expected sign and size. The size of the coefficients varies between 0.87 and 1.28 and the confidence intervals largely overlap. Remember that each estimate is the treatment effect on the treated. If we take the mean number of absence days in the pre-reform period for the treatment group, which was 9.7, and relate an estimate of 0.97 to it, we would conclude that the increase in statutory sick pay had led to an increase in the average number of absence days of 10 percent among the treated.

Section 2 discusses the political economy of the reform. The generosity expansion was a mechanical reaction to a cut in statutory sick pay in 1996 and was a promise made by the parties that formed the new government during their election campaign. Evaluating the cut in sick pay by using a different identification strategy, Ziebarth and Karlsson (2009) find that it decreased average absence days by about 5 percent. This is in line with the finding here and adds to the credibility of our identification strategy.

5.2 Robustness Checks

Apart from having analyzed the sensitivity of the results with respect to various parametric, nonparametric as well as combined methods, further robustness checks are shown in Table 3. Since the OLS estimates were pretty close to the results of the other models, we now focus on OLS estimation.

Column (1) displays results for specifications where the lagged level of the total number of absence days is included. The specification yield a positive and highly significant estimates of 1.527. Column (2) takes the years 1998 and 1999 and a balanced sample. We, thereby, estimate the short-run effect and control for sample attrition. The OLS point estimate is 1.2349 though it is imprecisely estimated with a p-value of 0.16. Column (3) excludes respondents who answered the following question with "yes:"*"Did you change your job or start a new one after December 31, 19XX?"*. We thereby capture job changers who might have selected themselves into (or out of) the treatment. The estimate is 1.1887 with a p-values of 0.12. Column (4) is the third check on attrition and selection concerns and includes in the four-year sample only those who were observed in the pre- *and* post-reform period at least once. The 1.166 OLS estimate is marginally significant at the 11 percent level. Column (5) displays the OLS estimate when the regression is weighted by the inverse probability of not dropping out of the labor market or sample in the post-reform period. The OLS point estimate has a value of 1.489 but is imprecisely estimated. However, the weighted ZINB-2 estimate (not shown) has a value of 1.0781 and is significant at the 6 percent level.

[Insert Table 3 about here]

As has already been mentioned, an indirect method to test the common time trend assumption is to perform the analyses for the years with no reform. Significant reform estimates for years with no reform would cast doubts on the assumption of no unobserved year-group effects. In this context, however, this is not the case as Table 4 demonstrates.

[Insert Table 4 about here]

5.3 Impact on Labor Costs and Job Creation

We calculate the maximum overall increase in labor costs by comparing the total employer-provided sick pay benefit sum in the pre-reform years 1997/1998 with the total benefit sum in the post-reform years 1999/2000, assuming that every employer only provided the 80 percent statutory sick pay in the pre-reform years. For this overall calculation, we do not need to use any of the empirical results from the last subsection but make use of the full sample, i.e., we consider all employees in the private sector.[14] The first benefit sum – total sick pay in the pre-reform years – is obtained by calculating the product of absence

[14] We consider all employees between 18 and 65 who work in the private sector and who were affected by the law. For employees who claimed that they had had a long-term absence spell of more than six weeks, we set the value for total absence days to 42 as only the first six weeks of sick leave are paid by the employer.

days multiplied by 80 percent of the daily gross wage for each individual in the pre-reform years. This total is then frequency weighted and multiplied with the frequency-weighted number of treated employees.[15] We do the same for the post-reform years but multiply each absence day with 100 percent of the daily gross wage. The difference between the two total sums yields the total labor cost increase if we assume that all employers provided sick pay according to the legal requirements before the reform. We obtain a total maximum increase in labor costs of €5.153 billion for the two post-reform years.

This total increase in labor costs can be decomposed into three components. The first component is rooted in the increase of the statutory sick pay level for the first six weeks from 80 to 100 percent of foregone gross wages. We approximate this amount by comparing the total sick leave payments in the pre-reform period to hypothetical sick leave payments for the same period and the same individuals assuming that the sick pay was already increased to 100 percent at that time. We thus disentangle the direct labor cost effect from the effect that is induced by increasing absence rates as a consequence of the reform. Again, we do not need any regression results for this exercise and use the full sample. Our calculation yields a direct labor cost increase effect of €3.87 billion for both years. If we assume that half of the firms had already provided 100 percent sick pay before the reform, this direct effect reduces to €1.93 billion.[16]

In the next step, we calculate the indirect labor cost effect which was triggered by the reform-induced increase in absenteeism and which represents the second component of the total increase in labor costs. From Table 2, we infer that the overall reform- induced increase in absence days equals approximately one day. Recall that the effect is very robust and homogeneous across various specifications. Hence, we take the average daily gross wage in the pre-reform years and multiply it by the frequency-weighted number of employees in both years, resulting in an indirect labor cost increase of €1.61 billion. If we assume that the increase was 0.9 or 1.1 days, we get indirect effects of 1.45 and 1.77 billion over both years, respectively.[17] The residual is the third component which is caused by time trends, changes in wages, and changes in the employment structure.

The total reform-induced increase in labor costs is thus $(1.93+1.61)/2 =$ €1.77 billion per year.[18]

[15] Frequency weights, which are computed according to data from the Federal Statistical Office, are provided by the SOEP group (SOEPGroup, 2001). Absence days and gross wages are included in the SOEP data. The SOEP group makes great effort to collect income data accurately and impute missing data consistently (Frick and Grabka, 2005).

[16] We, thereby, implicitly assume that employees who worked in companies which voluntarily provided 100 percent sick pay did not differ systematically in terms of absence days and wages from those who worked in companies which only provided statutory sick pay.

[17] Here, we focus on the same dataset that we used to obtain the estimated decrease of one day.

[18] By combining data from the Federal Statistical Office on the total number of employees obliged to pay social insurance contributions in the different years and age groups with SOEP data, we checked the plausibility and sensitivity of this estimate. Using this method, we also control for panel attrition. To

In 1997, the Research Institute of the Federal Employment Agency (IAB) calculated, by means of a general macroeconomic simulation model for Germany, that a reduction of the social security contribution rate by one percentage point would lead to the creation of 120,000 new jobs (Zika, 1997). These statistics were corroborated by other studies (Feil et al., 2008; Meinhardt and Zwiener, 2005). In Germany, social contribution finance five pillars of the German pay-as-you-go Social Security system, are mandatorily deducted from to the salary, equally paid by employer and employee, and amount to around 40 percent of the gross wage. For decades these indirect labor taxes have been of great concern to economists and policy-makers as they make labor more expensive and weaken incentives to take up work. Therefore, a reduction or stabilization of these contributions is one of the most important objectives for every government and was the main objective of various reforms over the last few decades.

For the whole of Germany, in 1997, one percentage point of social security contributions equated to about €5 billion and represents 0.5 percent labor costs for employers and 0.5 percent wage taxes for employees. If we assume that the employment effect in the cited simulation models stems solely from changing labor costs and labor demand, our back-of-the envelope calculation yields that the reform led to a loss of approximately 85,000 jobs (€2.5 bn $\approx$ 120,000 jobs; €1.77 bn $\approx$ 85,000 jobs). If we assume that half of the employment effect in the above models was rooted in changing product demand due to changes in net wages, this number falls to 42,500 when related to our increase in labor costs of €1.77 billion per year.[19]

Based on the combined evidence, one may conclude that between 40,000 and 80,000 jobs might have been lost through the reform due to an increase in labor costs of about €1.8 billion per year. While the latter figure seems to be pretty reliable, it should be borne in mind that the former figures are the result of a rough back-of-the envelope calculation. The claim that between 40,000 and 80,000 jobs might have been lost is based on the results of macroeconomic simulation models from other studies and is derived under various assumptions such as constant labor productivity. Nevertheless, we believe that the numbers are useful to illustrate the dimension of the reform. In Germany in 1999, the working population consisted of 35 million people (German Federal Statistical

calculate the two effects, we multiply the official employment data by SOEP absence rates and income data and get a very similar estimate of $(2.21+1.98)/2 =$ €2.1 billion per year (German Federal Statistical Office, 1996, 1998).

[19] The macroeconomic simulation models used to derive the increased employment effects assume a constant labor supply (Feil et al., 2008). In our rough calculation we do not account for the fact that the increase in sick pay led to higher net wages and that a potential associated increase in product demand might have offset parts of the job loss effect. However, two thirds of German GDP comes from exports, and domestic demand traditionally plays a minor role in Germany; domestic demand is very insensitive to aggregate wage changes, probably also because of the high savings rate which is more than ten percent in Germany. Lastly, we do not account for the possibility that increasing absence rates may increase labor demand.

Office, 2000).

6 Conclusion

The main aim of the study was to estimate the causal effect of an increase in statutory sick pay on sick leave behavior in Germany. The underlying reform became effective on January 1, 1999 and increased the statutory sick pay level from 80 to 100 percent of foregone gross wages for employees in the private sector. Since public sector employees, apprentices, and the self-employed were not affected by the law, we are able to analyze the reform effects by means of difference-in-differences methods and SOEP survey panel data. The panel data structure allows us to eliminate or avoid the typical pitfalls in evaluation studies like selection effects or sample attrition. Moreover, we do not only make use of conventional parametric regression methods but also employ various nonparametric matching methods and combined matching and regression approaches.

All methods unambiguously show that the reform has prompted private sector employees to take one more (fully paid) day off for reasons of sickness. Related to the average number of annual absence days, an increase in workplace absences of one day translates into an increase of about 10 percent.

Common features of contributions to this branch of the literature are that they all evaluate the overall effect among those who were, in principle, affected. We cannot identify private sector employees who were, de facto, not affected by the law since their employers voluntarily provided 100 percent sick pay already before the reform. If we assume that all employers provided solely the statutory sick pay prior to the reform, the labor costs increased nationwide by € 2 billion per year simply because of the rise of the replacement level. This direct labor cost effect can be quite precisely calculated since we have detailed wage information for each respondent and frequency weights for Germany. The direct labor cost increase falls to € 1 billion under the assumption that only half of all private sector employees were affected directly which is what polls suggested. The indirect labor cost effect is rooted in the reform-induced increase in absence rates and amounts to € 800 million per year for the whole of Germany. While we can calculate these numbers in a reliable manner using SOEP data, we need the results of other studies which are based on macroeconomic simulation models to approximate the reform's impact on job creation. A rough back-of-the envelope calculation suggests that between 40,000 and 80,000 jobs might have been lost as a consequence of the reform. In Germany the working population consists of 35 million people.

The question as to whether behavioral reactions to changes in sick pay levels prove the existence of moral hazard or presenteeism, i.e., employees going to work inspite of being sick, is difficult to answer empirically. It was beyond the scope of this paper but should be the subject of further research.

References

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens (2004). Implementing matching estimators for average treatment effects in Stata. *Stata Journal 4*(3), 290–311.

Abadie, A. and G. W. Imbens (2007). Bias corrected matching estimators for average treatment effects. Working paper, Havard University. http://ksghome.harvard.edu/ aabadie/bcm.pdf, last accessed at June 29, 2009.

Ai, C. and E. C. Norton (2004). Interaction terms in logit and probit models. *Economics Letters 80*, 123–129.

Bertrand, M., E. Duflo, and M. Sendhil (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–275.

Bound, J. (1989). The health and earnings of rejected disability insurance applicants. *American Economic Review 79*(3), 482–503.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications* (1 ed.). Cambridge University Press.

Campolieti, M. (2004). Disability insurance benefits and labor supply: Some additional evidence. *Journal of Labor Economics 22*(4), 863–890.

Card, D. and A. B. Krueger (1994). Wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review 84*(4), 772–793.

Chen, S. and W. van der Klaauw (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics 142*(2), 757–784.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics 24*(2), 295–313.

Curington, W. P. (1994). Compensation for permanent impairment and the duration of work absence: Evidence from four natural experiments. *The Journal of Human Resources 29*(3), 888–910.

Doherty, N. (1979). National insurance and absence from work. *The Economic Journal 89*(353), 50–65.

Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics 82*(2), 221–233.

Feil, M., S. Klinger, and G. Zika (2008). Der Beschäftigungseffekt geringerer Sozialabgaben in Deutschland: Wie beeinflusst die Wahl des Simulationsmodells das Ergebnis? *Journal of Applied Social Science (Schmollers Jahrbuch) 128*(3), 431–460.

Fenn, P. (1981). Sickness duration, residual disability, and income replacement: an empirical analysis. *The Economic Journal 91*(361), 158–173.

Frick, J. R. and M. M. Grabka (2005). Item-non-response on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv 89*(1), 49–60.

German Federal Statistical Office (1996). *Statistical Yearbook 1996 for the Federal Republic of Germany.* Metzler-Poeschel.

German Federal Statistical Office (1998). *Statistical Yearbook 1998 for the Federal Republic of Germany.* Metzler-Poeschel.

German Federal Statistical Office (1999). *Statistical Yearbook 1999 for the Federal Republic of Germany.* Metzler-Poeschel.

German Federal Statistical Office (2000). *Statistical Yearbook 2000 for the Federal Republic of Germany.* Metzler-Poeschel.

German Federal Statistical Office (2001). *Statistical Yearbook 2001 for the Federal Republic of Germany.* Metzler-Poeschel.

Gruber, J. (2000). Disability insurance benefits and labor supply. *Journal of Political Economy 108*(6), 1162–1183.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(2), 261–94.

Henrekson, M. and M. Persson (2004). The effects on sick leave of changes in the sickness insurance system. *Journal of Labor Economics 22*(1), 87–113.

Imbens, G. W. (2008). The evaluation of social programs: Some practical advice. Presentation, 2nd IZA/IFAU Conference on Labour Market Policy Evaluation. October 11, 2008.

Imbens, G. W. and D. B. Rubin (2009). *Causal Inference in Statistics and the Social Sciences* (1 ed.). Cambridge and New York: Cambridge University Press. forthcoming.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Jahn, J. (1998). Lohnfortzahlung: Gerichte stehen vor Herkulesaufgabe. *Handelsblatt 124: 02.07.1998*, 4.

Johansson, P. and M. Palme (1996). Do economic incentives affect work absence? Empirical evidence using swedish micro data. *Journal of Public Economics 59*(1), 195–218.

Johansson, P. and M. Palme (2002). Assessing the effect of public policy on worker absenteeism. *Journal of Human Resources 37*(2), 381–409.

Johansson, P. and M. Palme (2005). Moral hazard and sickness insurance. *Journal of Public Economics 89*, 1879–1890.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*(4), 604–620.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labour market policies. *The Review of Economics and Statistics 84*(2), 205–220.

Medizinischer Dienst der Krankenversicherung (MDK) (2008). www.mdk.de, last accessed at June 25, 2009.

Meinhardt, V. and R. Zwiener (2005). Gesamtwirtschaftliche Wirkungen einer Steuerfinanzierung versicherungsfremder Leistungen in der Sozialversicherung. Politikberatung kompakt 7, German Institute for Economic Research (DIW) Berlin. http://www.diw.de, last accessed at December 19, 2008.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review 85*(3), 322–340.

OECD (2006). *OCED Health Data 2006.*

Pettersson-Lidbom, P. and P. Skogman Thoursie (2008). Temporary disability insurance and labor supply: evidence from a natural experiment. Working paper, Stockholm University, Department of Economics. http://people.su.se/ pepet/tdi.pdf, last accessed at June 25, 2009.

Puhani, P. A. (2008). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. IZA Discussion Paper Series 3478, IZA. http://www.iza.org, last accessed at 22.02.2008.

Ridinger, R. (1997). Einfluss arbeitsrechtlicher Regelungen auf die Beschäftigungsentwicklung im Handwerk–Ergebnisse von Befragungen von Handwerksbetrieben im 3. Quartal 1997. Dokumentation, Zentralverband des Deutschen Handwerks. http://www.zdh.de, last accessed at June 19, 2009.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing the bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*(387), 516–524.

Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (1998). *Vor weitreichenden Entscheidungen.* Metzler-Poeschel.

Social Security Administration (2006). *Annual Statistical Supplement 2006, Table 9.A2.* http://www.ssa.gov/policy/docs/statcomps/supplement/2006/9a.html, last accessed at March 19, 2009.

Social Security Administration (2008). *Annual Statistical Supplement 2006, Table 9.C1.* http://www.ssa.gov/policy/docs/statcomps/supplement/2008/9c.html, last accessed at March 19, 2009.

SOEPGroup (2001). The German Socio-Economic Panel (GSOEP) after more than 15 years: Overview. *Quarterly Journal of Economic Research (Vierteljahrshefte zur Wirtschaftsforschung) 70*(1), 7–14.

Wagner, G. G., J. R. Frick, and J. Schupp (2007). The German Socio-Economic Panel study (SOEP) - evolution, scope and enhancements. *Journal of Applied Social Science (Schmollers Jahrbuch) 127*(1), 139–169.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5 ed.). Springer.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review 93*(2), 133–138.

Wooldridge, J. M. (2006). Cluster-sample methods in applied econometrics: an extended analysis. Working paper, Michigan State University, Department of Economics. https://www.msu.edu/ ec/faculty/wooldridge/current research/clus1aea.pdf, last accessed at March 19, 2009.

Wooldridge, J. M. (2007). What's new in econometrics? Imbens/Wooldridge lecture notes; summer institute 2007, lecture 10: Difference-in-differences estimation, NBER. http://www.nber.org/minicourse3.html, last accessed at March 19, 2009.

Ziebarth, N. R. and M. Karlsson (2009). A natural experiment on sick pay cuts, sickness absence, and labor costs. SOEPpapers 244, German Institute for Economic Research (DIW).

Zika, G. (1997). Die Senkung der Sozialversicherungsbeiträge. IAB Werkstattbericht 7, Research Insitute of the Federal Employment Agency (IAB).

Figure 1: Distribution of Propensity Scores Showing Region of Common Support

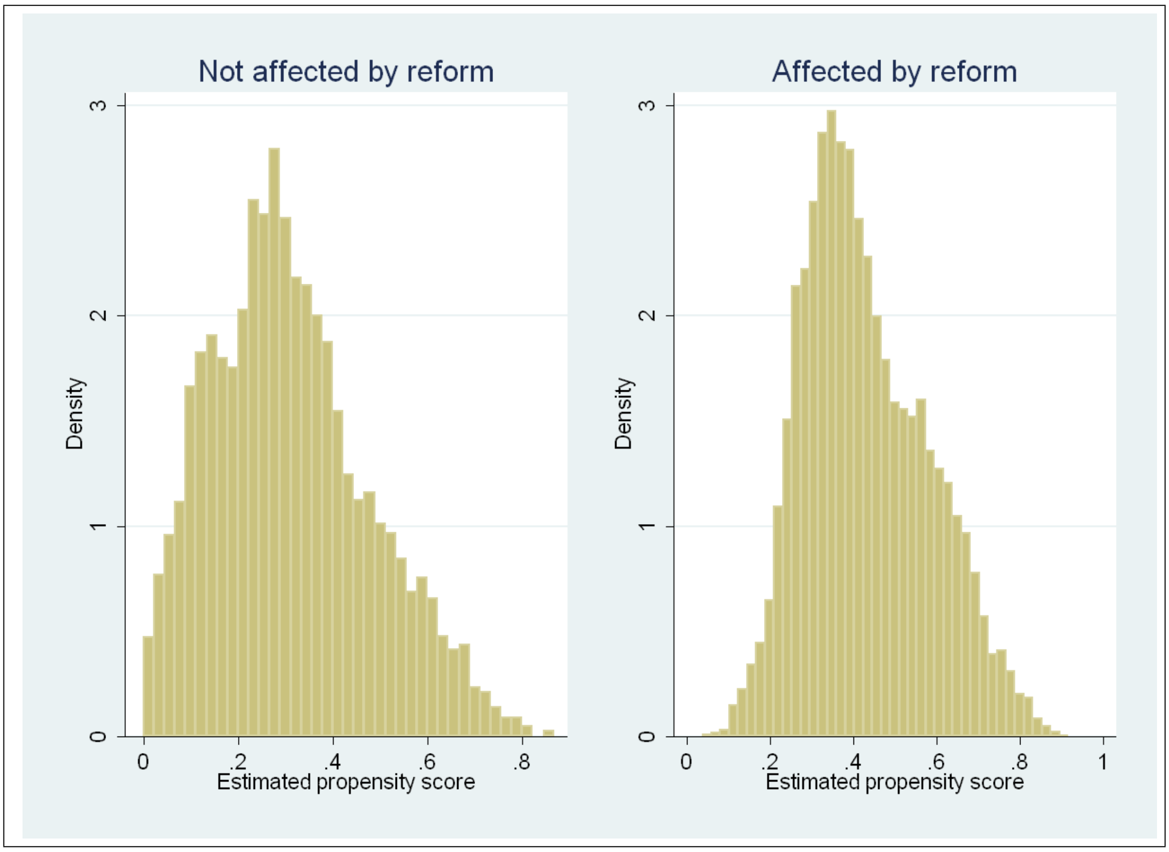

Figure 2: Average Predicted Sickness Absence Days for Treatment and Control Group Over Time

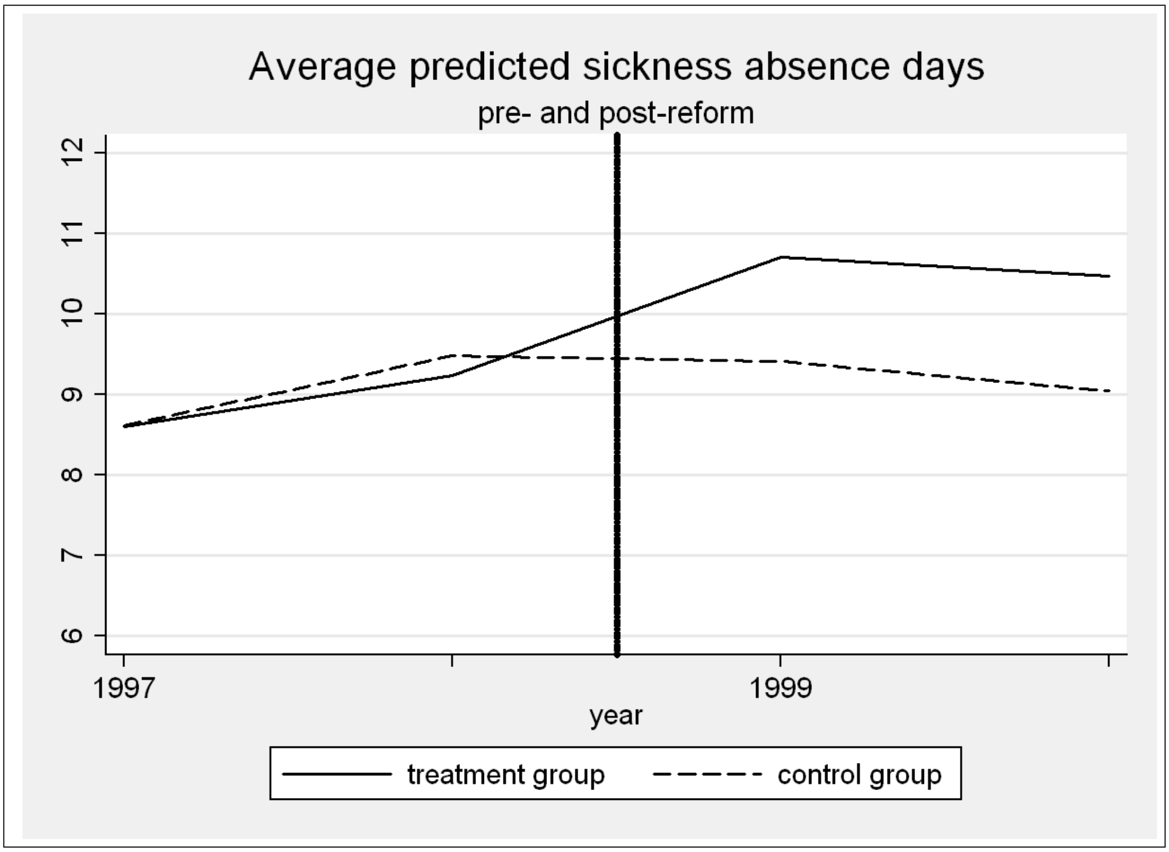

Table 1: Sample Means of Treatment and Control Group: Raw, Blocked, and Matched Sample

	Raw Sample			***Blocked Sample***			***Matched Sample***		
Covariates	**Treated mean**	**Controls mean**	**norm. diff.**	**Treated mean**	**Controls mean**	**norm. diff.**	**Treated mean**	**Controls mean**	**norm. diff.**
Age	39.068	40.865	0.156	39.068	39.339	0.083	39.535	40.090	0.048
Female	0.277	0.400	0.185	0.277	0.422	0.130	0.357	0.362	0.006
Partner	0.797	0.788	0.016	0.797	0.792	0.073	0.792	0.788	0.006
Married	0.676	0.687	0.018	0.676	0.684	0.061	0.678	0.671	0.011
Immigrant	0.180	0.078	0.217	0.180	0.080	0.084	0.128	0.109	0.041
Children	0.488	0.462	0.037	0.488	0.432	0.075	0.474	0.461	0.018
Disabled	0.042	0.047	0.018	0.042	0.042	0.049	0.051	0.046	0.017
Health good	0.633	0.640	0.011	0.633	0.597	0.031	0.627	0.637	0.015
Health bad	0.081	0.082	0.004	0.081	0.088	0.044	0.084	0.085	0.004
8 years of schooling	0.327	0.216	0.179	0.327	0.223	0.049	0.283	0.279	0.005
10 years of schooling	0.348	0.396	0.070	0.348	0.409	0.079	0.383	0.415	0.046
13 years of schooling	0.151	0.274	0.215	0.151	0.233	0.060	0.191	0.170	0.038
Trained for job	0.564	0.683	0.175	0.564	0.666	0.062	0.615	0.625	0.016
New job	0.159	0.180	0.123	0.181	0.117	0.041	0.148	0.142	0.011
Years with company	8.736	10.920	0.181	8.736	11.723	0.090	9.786	9.831	0.004
White collar	0.493	0.392	0.145	0.493	0.481	0.188	0.486	0.493	0.009
Job autonomy	0.218	0.407	0.295	0.218	0.409	0.115	0.281	0.248	0.052
Gross wage/1,000	2,389	2,610	0.117	2,389	2,405	0.082	2,474	2,406	0.038
State unemployment rate	11.134	11.755	0.097	11.134	11.870	0.104	11.542	11.718	0.027

"Norm. Diff." stands for "normalized difference" which is calculated according to $\frac{\bar{\mu}_1-\bar{\mu}_0}{\sqrt{\sigma_1^2+\sigma_0^2}}$, where $\bar{\mu}_1$ is the sample mean of the covariate for the treatment group and $\bar{\sigma}_0^2$ stands for the variance of the covariate within the control group. "Blocked sample" means that the sample was blocked to guarantee identicial propensity scores within blocks.

Here, the propensitiy score is the probability to belong to the treatment group and was estimated by a logit model under the inclusion of the displayed covariates and *(years with company)*^{2}, *(years with company)*$*female$, *(years with company)* $*$ *(trained for job)*, *(years with company)* $*$ *(job autonomy)*, *(Annual state unemployment rate)*^{2}, *(gross wage)*^{2}, *(gross wage)*$*female$, *(gross wage)*$*$*(white collar)*, *(gross wage)* $*$ *(13 years of schooling)*, *(gross wage)*$*$*married*.

The covariates in levels as well as their interactions to estimate the propensity score were selected according to likelihood ratio tests on zero coefficients as decribed in Imbens (2008).

After the PS estimation, 224 observations (0.01%) were not considered since they lay outside the common support which is [0.045; 0.919].

The number of blocks is twelve; the smallest block contains 41 treated and 316 untreated observations.

The matched sample has been generated by means of one-to-one nearest neighbor matching.

In total, the raw sample contains 23,426 observations, the blocked sample contains 23,202 observations, and the matched sample contains 14,502 observations.

Table 2: Difference-in-Differences Estimation: Regression, Matching, and Combined Methods

Variable	*Regression*		*Matching*			*Matching + Regression*	
	OLS	**ZINB-2**	**blocking**	**nearest neighbor**	**kernel**	**blocking + regression**	**n.neighbor + regression**
DiD	1.2778*	0.9251**	1.148***	1.247***	0.8729***	0.9829***	1.031**
	(0.7003)	(0.4716)	(0.421)	(0.561)	(0.421)	(0.3929)	(0.4355)
Covariates employed							
Job characteristics	yes	yes	yes	yes	yes	yes	yes
Educational characteristics	yes	yes	yes	yes	yes	yes	yes
Personal characteristics	yes	yes	yes	yes	yes	yes	yes
Regional unemployment rate	yes	yes	yes	yes	yes	yes	yes
Time dummies	yes	yes	no	no	no	yes	no
State dummies	yes	yes	no	no	no	yes	no
N	23,426	23,426	14,502	12,787	14,502	23,202	23,426

* p<0.1, ** p<0.05, *** p<0.01

Standard errors are in parentheses. In the parametric specifications, they are adjusted for intrapersonal correlations.

In all columns, the number of treated observations is 9,139.

The estimate in column (2) is the marginal effect, calculated at the means of the covariates except for *Post reform dummy*(=1), *Treatment Group*(=1), *Year 1999* (=1), *Year 2000* (=1), and *DiDg* (=1). ZINB-2 stands for *Zero-Inflated Negative Binominal Model 2*.

The underlying sample in columns (3) to (5) is the matched sample in Table 1.

In column (3), using the matched sample of Table 1, the propensity score (PS) of belonging to the treatment group in the post-reform period is estimated, based on a logit model and the same covariates as in Table 1.Based on this PS, the sample is stratified into eleven blocks, each with an equal PS for treated and non-treated. Then, the block-specific treatment effects – the difference in average outcomes for treated and non-treated – are weighted by the number of treated to obtain the overall average treatment effect on the treated.

In column (4), the average treatment effect on the treated is obtained by one-to-one nearest neighbor matching and column (5) makes use of kernel matching with a bandwith of 0.6.

Standard errors in column (5) are obtained by bootstrapping with 100 replications.

In column (6), the "blocked sample" of Table 1 is used. Then within each block, a ZINB-2-DiD regression is performed. Finally, the within block estimates are weighted by the number of treated observations to obtain the overall treatment effect on the treated.

In column (7), one-to-one nearest neighbor matching and regression are combined. As explained in Abadie et al. (2004), the estimator is bias corrected and allows for heteroskedastic errors.

Table 3: Robustness Checks

Model	+ lagged daysabs	1998 vs. 1999 + balanced	no job changers	observed pre- and post	weighted
OLS	1.527**	1.2349	1.1887	1.1664	1.489
	(0.7089)	(0.8876)	(0.7670)	(0.7285)	(1.099)
	[0.031]	[0.164]	[0.121]	[0.109]	[0.176]
N	19.458	8,623	19,668	15,347	23,426

* p<0.1, ** p<0.05, *** p<0.01

Standard errors in parentheses are adjusted for clustering on person identifiers. P-values are in square brackets.

All specifications are as in columns (1) and (7) of Table 2, except for the following:

The model in column (1) contains the lagged number of annual absence days as an additional covariate.

The model in column (2) uses a balanced sample contrasting the pre-reform year 1998 to the post-reform year 1999.

The model in column (3) excludes all those who have changed their jobs.

The model in column (4) includes only those who were observed at least once in the pre-reform years *and* the post-reform years.

The model in column (5) weights the regression with the inverse probability to not drop out of the sample in the post-reform period.

Table 4: Difference-in-Differences Estimation on the Number of Absence Days: Placebo Estimates

Model	**2000**	**2001**
OLS	0.2750	-0.6487
	(0.5980)	(0.6073)
ZINB-2 (d)	0.1925	-0.1220
	(0.4485)	(0.4399)
nearest neighbor	0.3468	-0.0617
+ regression	(0.4459)	(0.4129)
N	25,692	27,912

* p<0.1, ** p<0.05, *** p<0.01

Both columns make use of two pseudo pre- and two pseudo post-reform years, i.e., column (1) includes the waves 1999-2002 and column (2) includes the waves 2000-2003. Standard errors in parentheses are adjusted for clustering on person identifiers.

Appendix A

Table 5: Descriptive Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Treatment Group	0.6575	0.4746	0	1	23,426
Daysabs	9.782	24.8137	0	365	23,426
Personal characteristics					
Female	0.3193	0.4662	0	1	23,426
Age	39.7	8.2	25	55	23,426
Age squared	1642.3	661.2	625	3,025	23,426
Immigrant	0.1448	0.3519	0	1	23,426
East German	0.2646	0.4411	0	1	23,426
Partner	0.794	0.4044	0	1	23,426
Married	0.6797	0.4666	0	1	23,426
Children	0.4795	0.4996	0	1	23,426
Disabled	0.0439	0.2049	0	1	23,426
Good health	0.6354	0.4813	0	1	23,426
Bad health	0.0811	0.2731	0	1	23,426
No sports	0.3782	0.485	0	1	23,426
Educational characteristics					
Certificate after 8 years' schooling	0.289	0.4533	0	1	23,426
Certificate after 10 years' schooling	0.3643	0.4813	0	1	23,426
Certificate after 13 years' schooling	0.1929	0.3946	0	1	23,426
Other certificate	0.0754	0.264	0	1	23,426
Years with company	9.5	8.5	0	41	23,426
Trained for job	0.6052	0.4888	0	1	23,426
Job characteristics					
New job	0.1605	0.367	0	1	23,426
Blue-collar worker	0.372	0.4833	0	1	23,426
White-collar worker	0.4586	0.4983	0	1	23,426
High job autonomy	0.2823	0.4501	0	1	23,426
One man company	0.0338	0.1806	0	1	23,426
Small size company	0.2448	0.43	0	1	23,426
Medium size company	0.276	0.447	0	1	23,426
Large company	0.2202	0.4144	0	1	23,426
Very large company	0.2253	0.4178	0	1	23,426
Gross wage per month	2,465	1,287	404	28,632	23,426
Annual state unemployment rate	11.3	4.5	5.4	21.7	23,426