

Vouchers and caseworkers in training programs for the unemployed

Ulf Rinne · Arne Uhlenдорff · Zhong Zhao

Received: 27 February 2012 / Accepted: 3 September 2012 / Published online: 20 November 2012
© Springer-Verlag Berlin Heidelberg 2012

Abstract This paper studies the role of vouchers and caseworkers in training programs for the unemployed. We explore the unique features of the Hartz reform in Germany which simultaneously introduced training vouchers and imposed more selective criteria on participants. This allows us to go beyond the standard approach when we estimate the treatment effects for the most important type of training. Next to assessing the overall impact of the reform on the training's effectiveness, we isolate the impact induced by changes in the composition of program participants (selection effect) from the impact based on the introduction of vouchers and related institutional changes (institutional effect). Our results show a small positive overall impact of the reform. The decomposition suggests that the selection effect is, if at all, slightly negative, and that the introduction of the voucher and related institutional changes increased both

The data used in this paper originate from the evaluation of public training programs as part of the evaluation of the proposals of the Hartz Commission. [Schneider et al. \(2007\)](#) contains details.

U. Rinne (✉) · A. Uhlenдорff · Z. Zhao
Institute for the Study of Labor (IZA), Schaumburg-Lippe-Straße 5-9, 53113 Bonn, Germany
e-mail: rinne@iza.org

A. Uhlenдорff
University of Mannheim, Mannheim, Germany
e-mail: uhlenдорff@uni-mannheim.de

A. Uhlenдорff
IAB, Nuremberg, Germany

A. Uhlenдорff
DIW, Berlin, Germany

Z. Zhao
Renmin University of China, Beijing, China
e-mail: mr.zhong.zhao@gmail.com

employment and earnings of participants. It furthermore appears that our findings are driven by skilled participants.

Keywords Active labor market policy · Program evaluation · Matching · Voucher · Caseworker · Training

JEL Classification J64 · J68 · H43

1 Introduction

Active labor market policy (ALMP) aims at increasing employment and wages of unemployed individuals. The effectiveness of measures such as training programs and subsidized employment is analyzed in a large number of evaluation studies, which typically present results for *given* participants under a *given* allocation procedure. However, a number of related and important questions are rarely addressed. Who should participate in training programs, i.e., which groups of the unemployed benefit the most from participation? Who should allocate the training and choose the appropriate training provider? Should caseworkers decide on behalf of the unemployed or should the unemployed themselves make these choices?

These questions can only be addressed in an appropriate setting—which a recent labor market reform in Germany provides us with. When Germany reorganized its ALMP in a series of reforms, known as the Hartz reforms, the provision of training programs for the unemployed was substantially changed. The most important change was the introduction of a voucher scheme. The former contracting-out system was abandoned and replaced by a system in which job seekers are free to select their training provider in the market. This choice was previously made by the caseworker. Participants are, however, not completely free in their choice because the content of the training is still assigned by the caseworker. In addition to the voucher scheme, stricter criteria for the caseworkers to select training participants were introduced. The new rules imply that the caseworkers should select individuals with higher reemployment probabilities for participation—independently of the individual gain resulting from participation. The issue of creaming, or cream skimming, is, therefore, important because caseworkers' incentives are set to maximize participants' outcomes after treatment rather than treatment effects. However, those individuals who would likely succeed also without treatment are not necessarily the ones who benefit the most from training.

This paper takes advantage of the unique setting in Germany. Next to estimating the overall effect of a training program, which is the standard practice in the current literature, we try to disentangle the effect induced by the introduction of vouchers and related institutional changes (institutional effect) from the effect which is due to a more positive selection, or creaming, of participants by the caseworkers (selection effect). Although we cannot directly identify the voucher effect because we do not observe voucher receipt in our data, we carefully control for changes in the composition of participants after the reform. We can thus identify the institutional effect that comprises the voucher effect. Compared to a large body of empirical studies on the overall effect

of training programs, the roles of caseworkers and vouchers are still under-researched. Only few other studies aim at disentangling the impacts of different program components. For example, [Card and Hyslop \(2005\)](#) also identify a selection effect. However, our decomposition is based on a two-step propensity score matching procedure using a rich administrative dataset. This approach allows a comparison between the pre- and post-reform participants who have similar observable characteristics. We furthermore apply regression analysis to the matched data to adjust for possible remaining unbalanced covariates, to address the issue of potential effect heterogeneity, and to account for additional changes in economic and labor market characteristics.

Our results indicate a small positive overall impact of the reform. The decomposition of this increase in the overall effectiveness reveals three important results. First, we find that the selection effect is, if at all, slightly negative. This implies that using post-training outcomes as a performance standard does not improve the effectiveness of training programs for the unemployed. This finding is consistent with most of the available empirical studies reporting a modest impact of creaming on the effectiveness of training programs, if at all (see, e.g., [Heckman et al. 1997, 2002](#)). Second, we find that the introduction of the voucher and related institutional changes increased both employment and earnings of participants. The institutional effect becomes substantially positive around 6 months after entering training, and decreases slightly at the end of our observation period (1.5 years after program entry). Third, we find that the positive institutional effect is mainly driven by skilled participants. We do not find any significant institutional effect (or reform effect) for unskilled individuals.

The remainder of this paper is organized as follows. Section 2 summarizes the related literature and Sect. 3 outlines the institutional background of public training programs in Germany. After describing our analytical framework in Sect. 4 and our data in Sect. 5, we discuss the matching quality in Sect. 6 and present our results in Sect. 7. Finally, Sect. 8 concludes.

2 Related literature

A number of studies analyze the effectiveness of training programs in Germany before the Hartz reforms.¹ Their results are quite heterogeneous—depending on the investigation period and the underlying dataset. Recent studies which are based on rich administrative datasets often find at least positive treatment effects for some subgroups ([Lechner et al. 2011, 2007](#); [Fitzenberger et al. 2008](#); [Biewen et al. 2007](#); [Rinne et al. 2011](#)). However, there are also recent studies finding insignificant or negative effects ([Hujer et al. 2006](#); [Lechner and Wunsch 2008](#)). Besides differences in the investigation period and the underlying dataset, the mixed results may also be due to different methodological approaches. For instance, [Stephan \(2008\)](#) finds that estimated treatment effects differ considerably across different definitions of non-participants. Overall, the major lesson from the evaluation studies analyzing the pre-reform period (i.e., before 2003) seems to be that positive effects mainly occur in the longer run,

¹ The international literature on the evaluation of ALMP is summarized by [LaLonde \(2003\)](#) and [Kluve \(2010\)](#), among others.

and that studies which find positive medium- or long-term effects also report negative short-term effects.

We try to isolate the impact of the introduction of vouchers on the effectiveness of training programs. In general, the arguments supporting vouchers when allocating training programs are similar to the ones put forward for vouchers in education (Steuerle 2000; Barnow 2000, 2009). It is argued that allowing participants to choose the training provider in accordance to their preferences leads to better matches between the unemployed and training providers, which in turn increases the effectiveness of participation. Vouchers may moreover reduce organizational costs in allocating training programs. In addition, greater freedom of choice may encourage more competition among providers. Training providers may have to compete more intensively if they must regularly face the demand of participants instead of having a longer-term contract with the employment agency. This could lead to a further increase in the match quality and, hence, in the training's effectiveness.

On the other hand, there may be obstacles which could counteract the positive impacts of training vouchers (Hipp and Warner 2008; Barnow 2009). It is argued that the consumer—in our case the unemployed—may lack competence or resources to optimally choose, and that information asymmetries may lead to choices which do not truly reflect preferences. For example, caseworkers may know more about the availability of training courses and the quality of training providers than the unemployed because of their experience with previous participants. They may also have more information about occupations in demand and wages in the labor market, and participants may not correctly perceive their capabilities for specific occupations and training programs. Government and participant goals may also not necessarily coincide, which is due to the public good character of training. Finally, vouchers could lead to problems with market formation. It could, for example, be the case that the best-known providers rather than the most effective providers benefit.

Vouchers have been widely used in other fields of public services—in particular in the field of education—and are quite extensively studied in the literature.² Ladd (2002) presents a review of major studies on school vouchers. She concludes that the overall picture that can be drawn from these studies is rather inconclusive, and that results are not very robust. What can be learned seems to be that large-scale universal school voucher programs do not generate substantial gains and could even be detrimental to sub-populations. More narrowly targeted programs appear more promising. Neal (2002) argues along similar lines. He draws no general conclusions about the outcomes of school voucher programs, but he highlights the importance of taking into account the specific details about funding, targeting, and discretion.

There exist only few studies of vouchers for job training programs. Barnow (2009) gives an overview of studies on vouchers for vocational training programs in the United States. The empirical evidence on the effectiveness of training vouchers for unemployed workers is mixed. However, these studies are rather descriptive and to the best of our knowledge there exists no econometric study evaluating the effectiveness

² The discussion about vouchers in the field of education started with Friedman (1962).

of vouchers in training programs for the unemployed.³ A recent example for vouchers in the context of ALMP—although not in the field of education—is the job placement voucher. It was introduced in Germany in 2002 to end the public placement monopoly and to subsidize private competitors. [Winterhager et al. \(2006\)](#) find a positive impact on the employment probability of voucher recipients in West Germany.

Next to the introduction of vouchers, the stricter selection criteria that were imposed after the reform imply creaming, or cream skimming, of participants. This issue has been recognized as a possible concern in the literature on performance standards. It depends on the correlation between the outcomes in case of treatment (or in the absence of treatment) and treatment effects. Performance standards typically provide incentives to focus on individuals with relatively good labor markets outcomes—without considering the gains from participation. Treatment effects are then only maximized if outcomes and treatment effects are positively correlated (or, at least, not negatively correlated).

So far, most of the available empirical studies find that the impact of creaming on the effectiveness of training programs is modest, if at all (see, e.g., [Heckman et al. 1997, 2002](#)). This is due to the fact that short-term outcomes have essentially zero correlation with long-term treatment effects, and that treatment effects are relatively homogenous over a broad range of observable characteristics. The latter appears to hold also in the context of training programs for the unemployed in Germany, at least in the pre-reform period ([Rinne et al. 2011](#)).

From a slightly different point of view, however, focusing on untreated outcomes may actually be a good strategy. [Bell and Orr \(2002\)](#) find that caseworkers can predict untreated outcomes relatively well, but they do a poor job in predicting treatment effects. Therefore, the result that statistical treatment assignment rules on the basis of predicted impacts substantially increase post-program employment rates ([Lechner and Smith 2007](#)) may not be easily transferable into practice. Nevertheless, there appears to be considerable scope to improve the efficacy of caseworkers in assigning individuals to programs. Existing allocation procedures typically lie in the middle of post-program employment rates and are not superior to random assignment (see [Lechner and Smith 2007](#) and references mentioned therein).

3 Institutional background

ALMP aims to improve the labor market prospects of unemployed individuals. For this purpose, the Federal Employment Agency (FEA) in Germany spends a substantial amount of money on programs such as job creation schemes, training programs, or employment subsidies. The most important part of ALMP in Germany are training programs. With almost 7 billion Euros per year, these programs accounted for almost one third of the total expenditures before their major reform in 2003. In the early 2000s, the annual number of entrants into these programs was about half a million (see [Fig. 1](#)). However, this number declined in subsequent years. For example, in 2005 only around

³ [Görlitz \(2010\)](#) evaluates the impact of training vouchers for employed workers at the establishment level. This voucher program was implemented in one German federal state in 2006.

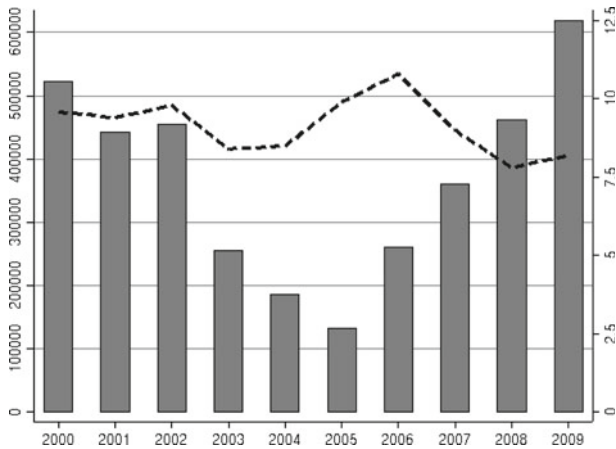


Fig. 1 Entrants into public training programs, unemployment rate (2000–2009). *Source* FEA. *Notes* Bars show annual number of entrants into public training programs (left axis). The dashed line represents the average annual unemployment rate (right axis, in percent).

130,000 individuals entered training programs. After it had reached a low in that year, the number of entrants increased again and reached a peak of more than 600,000 individuals in 2009.

Before the Hartz reforms were introduced in 2003, the provision of training programs was organized as follows. After consultation with the job seeker, the caseworker in the local FEA office decided if the unemployed individual should receive training. This was often done in agreement with the job seeker, although the final decision was subject to the discretion of the caseworker (Fitzenberger et al. 2010). Courses were operated by private providers which were approved beforehand. The system is considered as a de facto contracting-out, although there were no legal contracts between providers and local FEA offices. Legally, job seekers paid the courses and were reimbursed, but usually course fees were directly paid to providers to facilitate administration. The degree of competition among providers was limited as approvals were granted only to exactly the number of providers needed to meet regional demand. A public tendering procedure was not in place. This informal procedure entailed a potential for collusive behavior between local FEA offices and private providers that could involve an informal guarantee that capacities would be fully used. It was often reported that approved courses were simply filled up, even though the training provided was inappropriate for some participants (Bruttel 2005). Collusive behavior could moreover arise because training providers are frequently owned by the social partners, and employees' and employers' representatives are in the advisory board of the FEA which decides about strategy and budget policies (Hipp and Warner 2008).

Germany's ALMP has undergone important changes with the Hartz reforms, see Jacobi and Kluge (2007) and Caliendo (2009) for overviews. Figure 2 summarizes important legislative changes in this context. The overall strategy of the reforms involved three aims (Jacobi and Kluge 2007): (a) to improve the effectiveness and

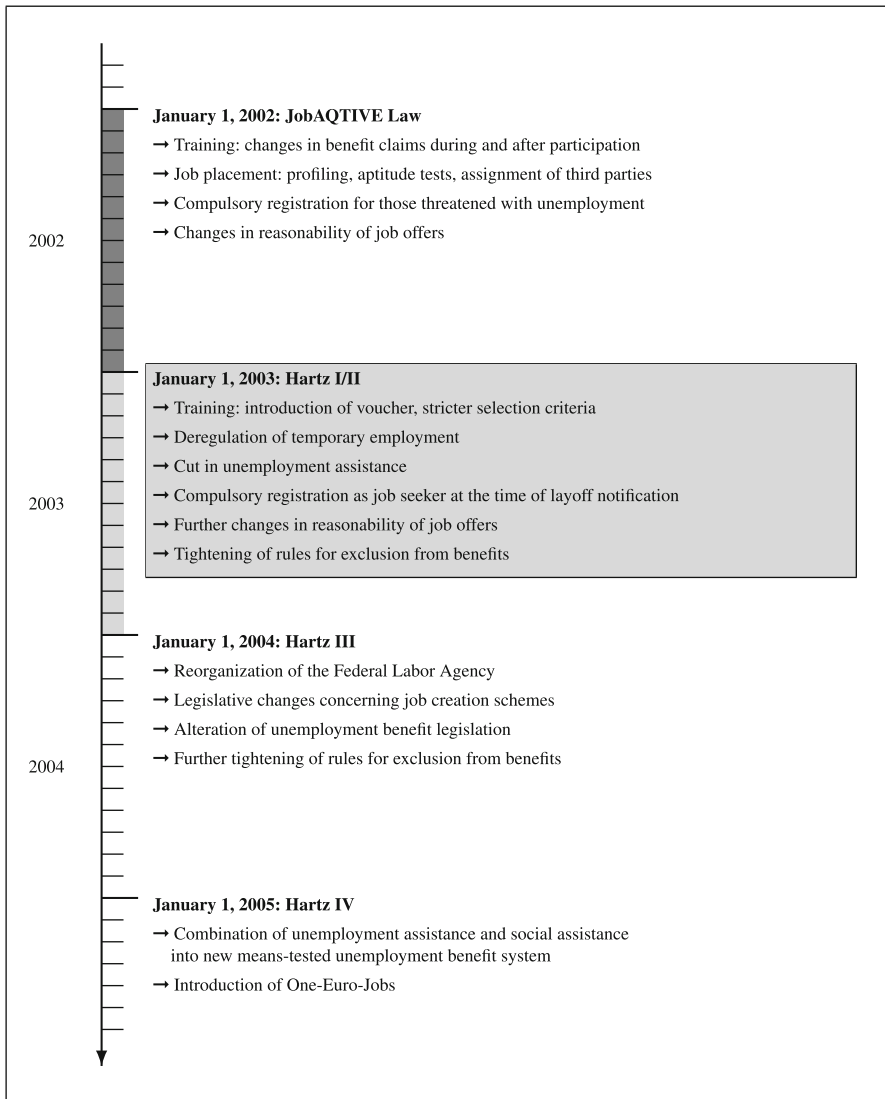


Fig. 2 Chronology of the Hartz reforms. *Source* Authors' illustration.

efficiency of labor market policy measures and services, (b) to activate the unemployed more strongly, and (c) to stimulate labor demand with deregulation.

Two important changes affected the provision of training programs after January 1, 2003. The first and the most prominent change was the introduction of the training voucher (*Bildungsgutschein*), see Fig. 3. The voucher prescribes the program's maximum duration, its educational target, its geographical scope, and the maximum course fee that is reimbursed. The voucher is valid for at most 3 months. Within this period, job seekers are free to choose among approved training providers and courses

Training Voucher Number.: _____/____
Customer Number: _____
(In accordance with § 77 para. 3 of the German Social Code III)

Valid until:	
Costs are covered	
<input type="checkbox"/> according to the certification procedure (lump-sum) <input type="checkbox"/> according to actual costs (as evidence is provided)	
Maximum duration: up to ____ months (including an internship if necessary)	
Educational target / qualifications provided:	
Type of course:	<input type="checkbox"/> full-time (35 hours/week) <input type="checkbox"/> part-time (12–24 hours/week) <input type="checkbox"/> on-the-job course <input type="checkbox"/> correspondence course
Provider:	<input type="checkbox"/> in-firm training <input type="checkbox"/> off-the-job training
Location:	<input type="checkbox"/> outside a daily commuting distance

Fig. 3 Training voucher. *Source* Authors' illustration

in the market—subject to the requirements stated in the voucher.⁴ Local FEA staff are not allowed to make recommendations, but they can, for example, provide a list of approved courses. A transitional arrangement was in place when the reform was introduced and the allocation of participants was *exclusively* based on vouchers only from March 2003 onwards (Schneider et al. 2007).⁵

However, a voucher is only granted if the caseworker considers participation in a given type of public training program as a successful strategy to reintegrate the job seeker in the primary labor market—without taking into account the *relative* gain compared to the counterfactual situation without participation. This is the second important change after the reform. The selection criteria for participants, therefore, became stricter, and the matching between program types and participants by the caseworkers based on the expected reemployment probability is novel. A voucher is only granted if two conditions are met (this is the so-called “70 % rule,” see Hipp and Warner 2008). First, the probability that the unemployed will immediately find a new job after finishing the training should be at least 70 %. Second, the training program should have had a placement rate of at least 70 % in the past.

Schneider and Uhlendorff (2006) and Schneider et al. (2007) find that the overall effectiveness of training programs has increased after the reform. Nevertheless, the question remains which features of the reform have caused this increase—and to what particular extent. We, therefore, decompose the overall reform effect into an

⁴ The approval of providers and courses is subject to a new quality management system which adopts a two-level approach. For details see, e.g., Bruttel (2005).

⁵ The official transitional arrangement was as follows: “Individuals who were counseled before January 1, 2003 and participation in a public training program was agreed upon do not receive a training voucher if they enter the program before March 1, 2003.”

institutional effect and a selection effect. This is the most significant difference of our study compared to previous studies analyzing the reform, and also compared to the related literature on the evaluation of training programs.

Although the innovative voucher system should increase consumer sovereignty and competition between training providers, [Bruttel \(2005\)](#) and [Hipp and Warner \(2008\)](#) are rather pessimistic about the actual impacts. Based on the initial evidence, these studies identify a number of practical obstacles to fully achieve the positive effects of training vouchers. For instance, information asymmetries constrain consumer sovereignty. In particular, low-skilled job seekers may lack the abilities to navigate the training market and to take an active role in searching for an appropriate course. Although the overall voucher redemption rate is comparatively high with 86 % in the period from 2003 to 2006, low-skilled individuals are significantly less likely to redeem a granted voucher than high-skilled unemployed ([Kruppe 2009](#)). On the supply side, a potential obstacle for competition between providers is their unequal distribution across German regions. Providers also reacted to the reform and increased co-operation and collusive behavior, for example by stopping to offer courses with the same or similar contents.

4 Analytical framework

We apply a two-step matching approach to disentangle the effect of the reform which is due to the introduction of the voucher and related institutional changes that potentially lead to a better match between participants and providers and to an improved quality of the offered training programs (institutional effect) from the effect which is due to changes in the composition of participants (selection effect). In order to isolate these two effects and to avoid complications and confusions with other elements of the reform process, we focus on individuals who entered training in 2002 (pre-reform period) and to individuals who did so in 2003 (post-reform period).

Using the potential outcome framework ([Neyman 1923](#); [Roy 1951](#); [Rubin 1974](#)), we assume that each individual has two potential outcomes for the program: Y_{1i} is the outcome if individual i participates, and Y_{0i} if not. Let D_i be an indicator for participation and R_i be an indicator for the post-reform period, then the average treatment effect on the treated individuals before (in 2002) and after the reform (in 2003) are given by:

$$ATT_{2002} = E[Y_{1i} - Y_{0i} | D_i = 1, R_i = 0] \quad (\text{ATT pre-reform period})$$

$$ATT_{2003} = E[Y_{1i} - Y_{0i} | D_i = 1, R_i = 1] \quad (\text{ATT post-reform period})$$

However, a simple comparison of treated and non-treated individuals may be biased if participants and non-participants differ with respect to characteristics which influence the outcome Y . If treatment assignment is *strongly ignorable*, i.e., if selection is based on observed characteristics X (conditional independence) and if observed characteristics of participants and non-participants overlap, the matching approach is

an appealing choice to estimate treatment effects. This implies that if unobserved characteristics play a role for the selection into training, they have to be uncorrelated with the outcome variables once we condition on X . Rosenbaum and Rubin (1983) show that if the matching assumptions hold, i.e., if treatment assignment is strongly ignorable *given* X , it is also strongly ignorable *given any balancing score that is a function of* X .⁶ One possible balancing score is the propensity score $P(X)$, i.e., the probability of participating in a given program. Mueser et al. (2007) present evidence that if rich administrative data are used to measure the performance of training programs, propensity score matching is generally the most effective.

We thus estimate ATT_{2002} ($ATT_n, 2003$) from the pre-reform data (post-reform data) by propensity score matching methods.⁷ However, the difference between ATT_{2002} and ATT_{2003} does *not* equal the effect of the introduction of vouchers, since the participants before and after reform may have different characteristics. As mentioned above, compared to the pre-reform period, the post-reform programs are subject to stricter selection criteria (possibly leading to a selection effect, SE) and vouchers and other institutional changes were introduced (which may cause an institutional effect, IE). If we assume additive separability of the two components, ATT_{2003} is given by:

$$ATT_{2003} = ATT_{2002} + IE + SE \quad (1)$$

and the overall reform effect (RE) can be written as:

$$\begin{aligned} RE &= ATT_{2003} - ATT_{2002} \\ &= IE + SE \end{aligned} \quad (2)$$

In order to isolate the institutional effect, we apply a two-step propensity score matching procedure. In the first step, the pre-reform participants are matched with the post-reform participants. Note that we have a relatively larger sample of the pre-reform participants. This implies that we will find for nearly all of our post-reform participants a corresponding match, which ensures that our matched sample of participants is representative for the post-reform participants.⁸ As a result, the obtained pairs of participants only differ with respect to the timing of participation. Importantly, observable characteristics do not differ anymore. In the second step, the matched pre-reform participants in 2002 are matched with non-participants of the same year. The corre-

⁶ When there are many covariates, it is impractical to match directly on covariates because of the curse of dimensionality. See, e.g., Zhao (2008) for some comments on this problem.

⁷ We have additionally run simple OLS for the pooled sample. The results (available on request) suggest that participants in 2003 have a significantly higher probability of being employed than participants in 2002. This is in line with our results.

⁸ This is precisely the reason why we perform the decomposition exercise in the presented way, namely that we take the post-reform participants and look for similar individuals among the pre-reform participants. The results of the decomposition might change if we would match participants in 2003 to participants in 2002, i.e., if we would hold the composition of the pre-reform constant. Therefore, the institutional effect has to be interpreted with respect to the group of participants in 2003. This issue is conceptually related to the "omitted group" problem in the Oaxaca-Blinder decomposition framework, see Fortin et al. (2011) for a discussion.

sponding treatment effect $ATT_{2002|2003}$ is the effect *only* for those participants under the pre-reform regime who are comparable to participants after the reform. This step controls for the changes in the composition of participants before and after the reform, i.e., the selection effect. Note that to identify $ATT_{2002|2003}$, we assume that there are no anticipation effects which may result in accelerated or delayed participation.

With $ATT_{2002|2003}$ we can calculate the difference in the following treatment effects to obtain an estimate of the institutional effect:

$$IE = ATT_{2003} - ATT_{2002|2003} \quad (3)$$

Finally, the comparison of the institutional effect with the reform effect gives us an estimate of the selection effect:

$$\begin{aligned} SE &= RE - IE \\ &= (ATT_{2003} - ATT_{2002}) - (ATT_{2003} - ATT_{2002|2003}) \\ &= ATT_{2002|2003} - ATT_{2002} \end{aligned} \quad (4)$$

There are several propensity score matching methods suggested in the literature, see, e.g., [Imbens \(2004\)](#), [Caliendo and Kopeinig \(2008\)](#), and [Imbens and Wooldridge \(2009\)](#) for overviews. Based on the characteristics of our data and in particular because of our two-step matching approach, we opt for the nearest-neighbor matching without replacement. This matching method has the advantage of being the most straightforward matching estimator as a given participant is matched with a non-participant (or participant) who is the closest in terms of the estimated propensity score. We avoid an increased variance of the estimator as we match without replacement ([Smith and Todd 2005a](#)), which is justified since the ratio between participants and potential matching partners is comparatively high in our data. Hence, the constructed counterfactual outcome is based only on distinct matches. The nearest-neighbor matching moreover generates a matched sample which can be further analyzed.

The matching method we apply is based on the conditional independence assumption. This is in general a very strong assumption and, hence, its plausibility is crucial. [Caliendo et al. \(2008\)](#) provide a good example of a careful discussion of this issue. The implementation of matching estimators requires a set of variables simultaneously influencing the participation decision or, more generally, the selection process into the program, and the outcome variable. In our specific case, information about socio-demographic characteristics and the educational background is, for example, critical. These are important factors determining individual employment prospects, on which caseworkers also base their assignment decisions. Our data contain a variety of such information. Moreover, our data allow us to construct a detailed employment history for each individual. It seems reasonable to assume that the employment history also contains information about unobserved variables not included in our data, e.g., motivation, attitudes, and aptitudes. As such characteristics are very likely persistent over time, they should be reflected in an individual's labor market history before program entry. For this reason, conditioning on the past labor market outcomes is crucial for the conditional independence assumption to hold ([Heckman et al. 1998](#)). Furthermore,

regional labor market characteristics play an important role for employment prospects as well as in the assignment process. Such information is included in our data (e.g., regional unemployment rate). Altogether, it is quite plausible that once we condition on this rich set of variables, the conditional independence assumption will hold, and that there will not be any additional variables which jointly influence the participation decision and the outcome variable which are not reflected by our rich set of variables.

We first estimate the probability of participation conditional on a number of observable characteristics using binary probit models, where an indicator of program participation is the dependent variable.⁹ These characteristics include socio-demographic characteristics (e.g., age, nationality, marital status, and number of children), regional information (region and unemployment rate), educational and vocational attainment, the employment history (4 years prior to program entry), and information on the last employment spell (duration, income, and business sector).¹⁰ We run these regressions separately for women and men from East and West Germany, respectively.

After estimating the propensity score, we perform the matching by exact covariate matching combined with propensity score matching. The variables used for exact matching are region, previous unemployment duration (in months), and quarter of program entry. Therefore, we stratify the four sub-samples of women and men in East and West Germany by these variables first, and then implement propensity score matching for each cell without replacement. The common support or overlap condition is also implemented at this stage. This procedure ensures that matched participants and non-participants are (a) previously unemployed for the same duration at program entry and (b) entering the program in the same quarter. The latter condition ensures that seasonal influences are constant. Furthermore, we do not condition on future non-participation. This is important in the context of dynamic assignment processes. Following the arguments of [Sianesi \(2004\)](#), in countries like Sweden or Germany in principle any unemployed individual will join a program at some point in time—provided that he or she remains unemployed long enough. Hence, a restriction on future outcomes (i.e., to require non-participation in the follow-up period after the fictitious program entry) is likely to affect estimated treatment effects negatively, since a substantial fraction of the “never treated”-individuals would de facto be observed to leave the unemployment register.

In order to assess reform impacts on the employment probability and earnings of participants, we estimate ordinary least squares models on the matched data. In order to test the robustness of our results with respect to potential differences in observable characteristics X , which may remain after the matching, we run additional regressions controlling for observable characteristics X on the matched sample. Applying regression models to matched data is a standard practice in the statistical literature (see, e.g., [Rubin 2006](#)), although it is less popular among economists. This procedure can correct

⁹ The matching algorithms are implemented using the PSMATCH2 Stata ado-package by [Leuven and Sianesi \(2003\)](#).

¹⁰ See [Table 1](#) for descriptive statistics on selected variables and the notes below [Table 1](#) for additional variables that we use to estimate the propensity scores. We include information on the characteristics which have been shown to be particularly important for selection correction in matching estimation, see [Lechner and Wunsch \(2011\)](#).

Table 1 Descriptive statistics (selected variables)

	Participants		Participants		Non-participants		Non-participants	
	2002		2003		2002		2003	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
<i>Socio-demographic characteristics</i>								
Female	0.433	(0.495)	0.431	(0.495)	0.444	(0.497)	0.444	(0.497)
East	0.416	(0.493)	0.402	(0.490)	0.295	(0.456)	0.305	(0.460)
Age	37.12	(9.589)	35.66	(9.826)	39.47	(11.71)	39.10	(11.38)
German citizenship	0.907	(0.290)	0.918	(0.274)	0.854	(0.353)	0.869	(0.338)
<i>Educational attainment</i>								
No graduation	0.057	(0.232)	0.062	(0.242)	0.123	(0.329)	0.124	(0.329)
First stage of secondary level	0.321	(0.467)	0.294	(0.456)	0.479	(0.500)	0.460	(0.498)
Second stage of secondary level	0.410	(0.492)	0.436	(0.496)	0.280	(0.449)	0.294	(0.456)
Advanced technical college entrance qualification	0.068	(0.252)	0.080	(0.271)	0.038	(0.192)	0.042	(0.201)
General qualification for university entrance	0.144	(0.351)	0.128	(0.334)	0.079	(0.270)	0.079	(0.270)
<i>Vocational attainment</i>								
No vocational degree	0.192	(0.394)	0.196	(0.397)	0.366	(0.482)	0.349	(0.477)
In-plant training	0.598	(0.490)	0.601	(0.490)	0.515	(0.500)	0.519	(0.500)
Off-the-job training, vocational school, technical school	0.101	(0.301)	0.100	(0.300)	0.069	(0.254)	0.076	(0.265)
University, advanced technical college	0.110	(0.313)	0.103	(0.304)	0.050	(0.218)	0.057	(0.231)
<i>Employment history</i>								
Share of employment in 1st year before program entry	0.243	(0.301)	0.239	(0.292)	0.199	(0.289)	0.184	(0.278)
Share of employment in 2nd year before program entry	0.469	(0.434)	0.523	(0.428)	0.367	(0.415)	0.371	(0.417)
Share of employment in 3rd year before program entry	0.479	(0.445)	0.550	(0.439)	0.390	(0.431)	0.410	(0.432)
Share of employment in 4th year before program entry	0.477	(0.444)	0.545	(0.439)	0.401	(0.434)	0.431	(0.434)

Table 1 continued

	Participants 2002		Participants 2003		Non-participants 2002		Non-participants 2003	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
<i>Unemployment history</i>								
Share of unemployment in 1st year before program entry	0.583	(0.310)	0.583	(0.309)	0.608	(0.323)	0.615	(0.323)
Share of unemployment in 2nd year before program entry	0.240	(0.327)	0.226	(0.318)	0.326	(0.369)	0.334	(0.373)
Share of unemployment in 3rd year before program entry	0.197	(0.313)	0.157	(0.276)	0.287	(0.373)	0.275	(0.365)
Share of unemployment in 4th year before program entry	0.176	(0.300)	0.140	(0.263)	0.257	(0.362)	0.249	(0.358)
# Observations	25,223		1,319		592,406		604,193	

Source IEB, own calculations

SD standard deviation. Additional variables that we use to estimate the propensity scores are number of dependent children, marital status, disability, regional unemployment rate, business sector of last employment, occupation of last employment, duration of last employment, transfer income, and labor income during the 4 years prior to program entry. Note that we match exactly on region, previous unemployment duration, and quarter of program entry

a possible remaining bias (Rubin 1973; Imbens 2004; Abadie and Imbens 2006) and it is closely related to the double-robustness literature (see, e.g., Tsiatis 2006). This procedure also allows us to address the issue of potential effect heterogeneity. Furthermore, we can control for changes in the general economic situation and changes in the extent and the composition of ALMP. Such changes may be additional components of the reform effect. For instance, Lechner and Wunsch (2009) present evidence for a clear positive relation between the effectiveness of the programs and the unemployment rate over time. Although we generally argue that we control for such changes as participants and matched non-participants are subject to the same economic environment, we will explicitly address this issue in our sensitivity analysis and control for potential changes in our regressions.

For the variance of the estimated treatment effects, we base our inference on bootstrapping procedures. More specifically, we bootstrap the whole estimation process. This allows us to calculate the standard errors based on the distribution of the estimated treatment effects. The standard errors of the reform effect, the voucher effect, and the selection effect are based on the distribution of the respective differences in treatment effects across 200 replications. Using bootstrap methods to obtain the standard errors is a popular practice when applying matching estimators, but we should acknowledge that the nearest-neighbor matching does not satisfy the conditions for the standard (“ n out of n ”) bootstrap procedure which we use (Abadie and Imbens 2008).¹¹

5 Data

We use a sample of a particularly rich administrative dataset, the Integrated Employment Biographies (IEB) of the FEA. Our data contain detailed daily information on employment subject to social security contribution including occupational and sectoral information, on the receipt of transfer payments during periods of unemployment, and on participation in different programs of ALMP. Furthermore, the IEB comprises a large variety of covariates such as age, marital status, number of dependent children, disability, nationality, and education.

In Germany, training programs for the unemployed are quite heterogenous. We, therefore, concentrate on the most important program type which is occupation-related or general training.¹² Participants either learn specific skills required for a certain vocation (e.g., computer-aided design for a technician/tracer) or receive qualifications that are of general vocational use (e.g., MS Office and computer skills). The program does *not* aim to provide a certificate, i.e., no formal vocational degree is awarded. In contrast to other program types, it focuses on classroom training and is neither provided in combination with internships nor is the simulation of real operations conducted. It nevertheless aims at improving the human capital and productivity of participants. We

¹¹ Ham et al. (2011) experiment with both the standard bootstrap and “ m out of n ” bootstrap in their matching estimators, and they find that both yield similar results. However, “ m out of n ” bootstrap requires much weaker conditions.

¹² Occupation-related or general training is the most important type of training programs, which is also often referred to as further training (Fitzenberger et al. 2010). The two other broad categories of training programs are retraining and short-term training.

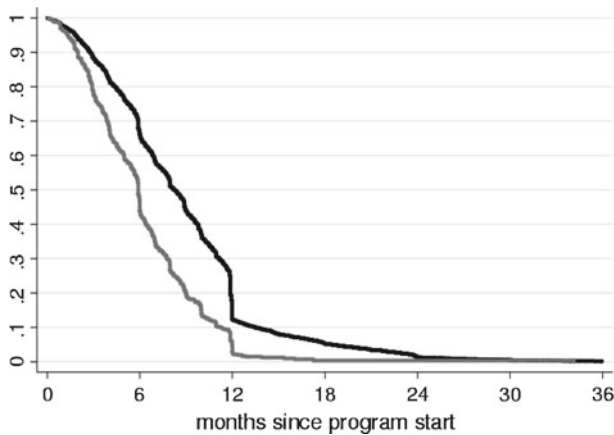


Fig. 4 Actual program duration. *Source* IEB, own calculations. *Notes* Kaplan–Meier estimates. Pre-reform period in *black*, post-reform period in *gray*

thus do not consider occupational retraining and only consider short further training programs. In the pre-reform period about 60 % of all participants in public training programs were assigned to the program type that we consider. It became even more important after the reform in 2003 as this share increased to more than 70 %. Figure 4 indicates that the program is—in comparison to other ALMP measures in Germany—a rather short measure. When we measure program duration as the time between actual program entry and exit, after 1 year more than 90 % of the participants have ended the program both in the pre-reform and in the post-reform period.¹³ However, the program duration decreased after the reform. Whereas the median program duration is about 8 months in the pre-reform period, it amounts to about 6 months after the reform. Although the shorter duration could in principle affect the program’s effectiveness, Flores et al. (2012) and Kluge et al. (2012) find that there is no substantial impact of training duration on employment outcomes. The share of shorter programs is similar in the latter paper, and the authors find that program duration has a positive impact on the training’s effectiveness only at the beginning. Therefore, changes in program duration should not affect our results.

In order to evaluate the impact of the reform and its features on the effectiveness of this training type, our data include participants as well as non-participants from the pre- and post-reform period, respectively. More specifically, we have information on: (a) participants who entered the program in 2002, (b) participants who entered the program in 2003, (c) non-participants in 2002, and (d) non-participants in 2003. We do not have information about voucher receipt. Hence, individuals who received a voucher and did not make use of it are included in our sample of non-participants in 2003.¹⁴ This is the reason why we can only identify an institutional effect, but are

¹³ It is in general possible to drop-out early from the program as well as to prolong participation. The actual duration of training is thus endogenously determined (Fitzenberger et al. 2010).

¹⁴ The take-up rate is 86 % from 2003 to 2006 (Kruppe 2009). A simple correction would be to weight the treatment effect by the inverse take-up rate, see Bloom (1984). Assuming a positive impact of voucher

not able to directly identify the voucher effect. Moreover, we cannot rule out that an intention-to-treat effect exists that is based on voucher receipt and potentially biases our results. In order to at least partly circumvent this caveat of our data, as part of our main results we, therefore, also present results when we exclude the first quarter in 2003 (i.e., when we exclude the period when the transitional arrangement was in place).

Our sample of participants who entered the program in 2003 consists of more than 1,300 individuals. In order to apply the two-step matching approach, roughly 20 participants from the period before the reform were drawn per participant in 2003. Therefore, we have information on about 25,000 participants who entered the program in 2002. Beyond matching of the post-reform participants with the pre-reform participants, we need to match participants with non-participants. In both years (2002 and 2003) our sample of non-participants—i.e., potential controls—consists of roughly 600,000 individuals. Non-participants are required to not have participated in the given type of training before and in the quarter of the participant's program entry, but we do not condition on future non-participation.

Table 1 displays descriptive statistics of the selected variables for the samples of participants and non-participants in 2002 and 2003, respectively. First, we find evidence for a change in the composition of participants in training between the pre- and post-reform period in our data. The most remarkable change can be observed with respect to previous employment histories. Considering a period of 4 years prior to program entry, participants who entered after the reform show a substantially higher labor market attachment in terms of lower unemployment rates and higher employment rates. The average age of a participant dropped by more than 1 year between 2002 and 2003, while other characteristics remain on average rather stable between the 2 years. In particular, differences with respect to the educational or vocational attainment do not appear to be substantial.¹⁵ On the other hand, the groups of non-participants are very different from the groups of participants in both years. They are on average older and less educated. Moreover, their employment histories reveal a higher incidence of unemployment as well as a lower incidence of employment when compared to participants.

The success of program participation is evaluated by looking at (a) the employment probability, and (b) earnings. Our observation period—i.e., the period in which outcomes are observed—starts at program entry and ranges over 18 months. This period is based on the facts that we focus on program participation in the years 2002 and 2003, and that we can observe reliable data for all employment states until December 31, 2004. Individuals are regarded as employed if they hold a job in the primary labor market. For instance, participation in job creation schemes is not included in this outcome measure. Moreover, the administrative dataset only includes employment that

Footnote 4 continued

receipt, our results can, therefore, be seen as a lower bound of the institutional effect (which includes the voucher effect).

¹⁵ We implement an imputation procedure for the educational and vocational attainment variables that is similar to the imputation procedure “IP1” in [Fitzenberger et al. \(2006\)](#). However, we only apply this procedure in case of missing information at the (fictitious) program entry.

Table 2 Economic and labor market variables

	2002 Mean (SD)	2003 Mean (SD)	2004 Mean (SD)
(1) Job seekers	0.1238 (0.0630)	0.1329 (0.0603)	0.1441 (0.0617)
(2) Unemployment rate	0.0991 (0.0499)	0.1057 (0.0491)	0.1065 (0.0490)
(3) Vacancies	0.0107 (0.0050)	0.0082 (0.0044)	0.0065 (0.0038)
<i>Participants in ...</i>			
(4) Public training programs	0.0085 (0.0052)	0.0065 (0.0037)	0.0045 (0.0023)
(5) Subsidized employment	0.0037 (0.0043)	0.0040 (0.0049)	0.0029 (0.0037)
(6) Job creation schemes	0.0050 (0.0076)	0.0037 (0.0059)	0.0029 (0.0047)
(7) GDP growth rate	1.5191 (1.3170)	1.0258 (0.6455)	2.2837 (0.7575)

Source FEA; Statistical Offices of the Federal States.

(1)–(6) are monthly shares in the civilian labor force in 178 FEA districts. (7) are annual GDP growth rates for the 16 Federal States

is subject to social security contributions. This implies, for example, that we do not observe self-employment in our data. Additionally, we evaluate the effect of program participation on monthly earnings in the primary labor market. We apply the described definition of employment and consider nominal remunerations associated with these spells in terms of monthly earnings. If individuals are not employed, we assume zero earnings.

In order to control for changes in the general economic situation, which may constitute another component of the reform effect (Lechner and Wunsch 2009), we consider a number of economic and labor market characteristics available for each labor market district in our sensitivity analysis. We use monthly information on the share of unemployed, vacancies, and participants in various ALMP measures (including training) as well as on GDP growth rates.¹⁶ Table 2 reports the mean values of these variables between 2002 and 2004. For example, the unemployment rate slightly increased on average from around 10 % in 2002 to around 10.7 % in 2004, while the share of unemployed individuals participating in training programs decreased during this period.

Furthermore, the implementation of the reform may have varied across local FEA districts. The strategies of how to implement the reform may influence the effectiveness

¹⁶ We include annual GDP growth rates for the 16 federal states since more disaggregated data is not available.

Table 3 Rating of the Hartz reforms by FEA districts

How did the reforms affect the...	Negative		Neutral 0	Positive		Missing
	-2	-1		+1	+2	
...effectiveness of job placement	1	8	81	66	3	4
	0.6	4.9	49.7	40.5	1.8	2.5
...process of job placement	1	9	61	82	6	4
	0.6	5.5	37.4	50.3	3.7	2.5
...efficiency of job placement	1	7	64	83	5	3
	0.6	4.3	39.3	50.9	3.1	1.8
...process of benefit granting	2	14	67	64	13	3
	1.2	8.6	41.1	39.3	8.0	1.8
...co-operation with third parties	2	34	108	15	1	3
	1.2	20.9	66.3	9.2	0.6	1.8
...administration effort	23	60	56	18	1	5
	14.1	36.8	34.4	11.0	0.6	3.1
...matching accuracy of job placement	1	6	71	75	6	4
	0.6	3.7	43.6	46.0	3.7	2.5

Source Survey in 163 FEA districts conducted in the beginning of 2005

First row: frequencies; second row: percentages

of labor market policies such as training programs, and they could, therefore, be actually part of the treatment. We address this issue by additionally using information on the administrators’ subjective judgement of the Hartz reforms in our sensitivity analysis. This information is obtained through a survey conducted in the beginning of 2005 in the management departments of the local FEA districts. The respondents are asked about the change of the job placement, the benefit granting, the administrative effort, and the co-operation with the third parties like training providers and employers. The subjective judgements appear on average rather positive. However, we observe heterogeneity in the judgements, and we will control for this in our regressions. The included items are reported in Table 3.

6 Matching quality

We apply different strategies to evaluate the balancing of observable characteristics between the different groups after the matching.

One way to assess the matching quality is to compare the standardized difference before matching, SD^b , to the standardized difference after matching, SD^a . The standardized differences are defined as

$$SD^b = \frac{(\bar{X}_1 - \bar{X}_0)}{\sqrt{0.5 \cdot (V_1(X) + V_0(X))}} ; \quad SD^a = \frac{(\bar{X}_{1M} - \bar{X}_{0M})}{\sqrt{0.5 \cdot (V_1(X) + V_0(X))}}, \quad (5)$$

where $X_1(V_1)$ is the mean (variance) in the treated group before matching and $X_0(V_0)$ is the analog for the comparison group. X_{1M} and X_{0M} are the corresponding means after matching (Rosenbaum and Rubin 1985). The mean standardized difference should be reduced after matching.

Following the suggestion of Sianesi (2004), we also re-estimate the propensity score on the matched sample to compute the pseudo- R^2 before and after matching. The pseudo- R^2 indicates how well the observable characteristics X explain the probability of being treated. After matching the pseudo- R^2 should be low because there should be no systematic differences between the treated and untreated individuals.

In a third approach we test the balancing following a suggestion by Smith and Todd (2005b) and estimate the following regression for each observable characteristic x included in our preferred specification:

$$x_k = \beta_0 + \beta_1 \widehat{\text{PS}}(X) + \beta_2 \widehat{\text{PS}}(X)^2 + \beta_3 \widehat{\text{PS}}(X)^3 + \alpha_0 D + \alpha_1 D \widehat{\text{PS}}(X) + \alpha_2 D \widehat{\text{PS}}(X)^2 + \alpha_3 D \widehat{\text{PS}}(X)^3 + \varepsilon_k, \quad (6)$$

where D is the treatment indicator, $\widehat{\text{PS}}(X)$ the estimated propensity score, and x_k is the observable characteristic k . For each x we perform an F test of the joint null hypothesis that all coefficients on terms involving D equal zero. If the balancing score satisfies the balancing condition, D should not provide any information about x_k .

Table 4 summarizes the results from the three balancing tests for the different subsamples of women and men in East and West Germany, respectively. Altogether, we perform five matching procedures: (a) the pre-reform participants are matched with the post-reform participants, (b) the *unmatched* pre-reform participants are matched with the pre-reform non-participants, (c) the *unmatched* post-reform participants are matched with the post-reform non-participants, (d) the *matched* pre-reform participants are matched with the pre-reform non-participants, and (e) the *matched* post-reform participants are matched with the post-reform non-participants.

Note that *unmatched* and *matched* participants may differ because we do not find for every participant after the reform a match from the period before the reform, i.e., the *matched* participants are a subset of the *unmatched* participants due to lack of common support. In the one-step matching approach, we exclude 547 participants in 2002 (or 2.2 % of our original sample) and 314 participants in 2003 (23.8 %). The two-step matching approach, however, is more demanding. Therefore, we exclude a larger fraction of participants because either we do not find a match among the participants in the first step or, for those with a matched participant, we do not find a match among the non-participants in the second step. Altogether, we find matches for about 52 % of the potential number of matched pairs in the pre- and post-reform period. Our two-step matched sample thus consists of about 700 matched pairs in each 2002 and 2003.¹⁷ The relatively low number of matched pairs may appear problematic, but this should mainly affect the interpretation of our results as they are based only on this sub-group of individuals.

¹⁷ Note that the potential number of matched pairs in the two-step matching approach is determined by the sample size of participants in 2003, which is 1,319 individuals.

Table 4 Matching quality

Sex	Region	Before matching			After matching			
		# obs.	Mean SD	R^2	# obs.	Mean SD	R^2	# sign.
<i>(a) Two-step matching: Participants 2002–Participants 2003</i>								
Female	East Germany	4,390	8.359	0.2001	210	8.652	0.1329	0
Female	West Germany	7,094	10.211	0.2125	320	6.760	0.1054	0
Male	East Germany	6,624	7.493	0.0990	422	5.135	0.0595	0
Male	West Germany	8,434	8.184	0.1158	470	5.972	0.0769	0
<i>(b) One-step matching: Participants 2002–Non-participants 2002</i>								
Female	East Germany	80,915	11.621	0.0536	8,116	1.059	0.0017	4
Female	West Germany	193,096	11.991	0.0693	13,190	0.982	0.0013	2
Male	East Germany	104,265	11.730	0.0562	12,328	1.050	0.0015	0
Male	West Germany	239,353	11.725	0.0584	15,718	0.895	0.0009	0
<i>(c) One-step matching: Participants 2003–Non-participants 2003</i>								
Female	East Germany	81,406	13.461	0.0486	328	4.329	0.0337	0
Female	West Germany	187,288	15.383	0.0558	508	3.714	0.0278	0
Male	East Germany	103,110	14.231	0.0535	510	4.064	0.0271	1
Male	West Germany	233,708	13.358	0.0480	664	2.241	0.0106	0
<i>(d) Two-step matching: Matched Participants 2002–Non-participants 2002</i>								
Female	East Germany	76,845	24.504	0.1092	202	5.811	0.0385	0
Female	West Germany	186,516	25.836	0.1068	310	3.402	0.0247	0
Male	East Germany	98,167	20.709	0.0881	410	2.969	0.0214	0
Male	West Germany	231,589	22.185	0.0724	446	2.629	0.0204	0
<i>(e) Two-step matching: Matched Participants 2003–Non-participants 2003</i>								
Female	East Germany	81,296	21.841	0.0859	206	3.924	0.0333	0
Female	West Germany	187,094	23.526	0.0946	310	2.581	0.0121	0
Male	East Germany	103,006	18.842	0.0664	412	2.916	0.0219	0
Male	West Germany	233,508	18.242	0.0714	464	2.329	0.0151	0

obs. number of observations, mean SD mean standardized difference, R^2 pseudo- R^2 of propensity score estimation, # sign. number of observable characteristics for which F test rejects the joint null. Further details are given in the text

Overall, the balancing of the different matching procedures is quite satisfactory. The mean standardized differences in the matched samples are—with one exception—noticeably smaller than in the unmatched samples and are mostly below 5 % after matching. Likewise, the pseudo- R^2 after matching are fairly low and decrease substantially compared to before matching. Moreover, in most of the matching procedures our third test indicates that D does not provide any information about the observable characteristics. However, some of our matching procedures perform better than others. We get the worst performance for our matching of participants before the reform with participants after the reform, especially for females in East Germany—although the third test indicates no problems for the participant–participant matching for any of our sub-samples. Therefore, we will check the sensitivity of our results to the inclusion

of observed characteristics in our regressions based on the matched samples, and we have to be careful when interpreting our results for females in East Germany.¹⁸

7 Results

In this section, we present the effects on employment probabilities and earnings for a period of 1.5 years after program entry. First, we present results for a restricted sample where we exclude the period after the reform during which the transitional arrangement was in place. Second, we assess the effects for the full sample. Third, we investigate whether effects differ across gender, region, and skill groups. Finally, we conduct a sensitivity analysis in which we account for additional economic and labor market characteristics.

7.1 Restricted sample

We mentioned above that there has been a transitional arrangement in place until March 2003 (see Sect. 3). Unfortunately, our data do not allow us to identify those participants who actually received and redeemed a training voucher. According to [Schneider et al. \(2007\)](#) who analyze survey data, the fraction of participants in public training programs actually receiving a voucher was about 30 % in the first quarter of 2003. We thus perform our analysis first in a restricted sample where we exclude participants who entered training in the first quarter of 2003.¹⁹

The estimates of the pre- and post-reform average treatment effects on employment probabilities in the restricted sample are reported in Fig. 5. We observe that participants before and after the reform face a substantial locking-in effect as both treatment effects are significantly negative in the first months.²⁰ After around 6 months of training, both treatment effects diverge and the treatment effect for participants after the reform constantly lies above the treatment for participants before the reform. At the end of our observation period, i.e., 1.5 years after program entry, the point estimates of the treatment effects amount to about 3 % points before the reform and more than 10 % points after the reform.

The difference between the two treatment effects corresponds to the reform effect. We thus find a positive impact of the reform, which may be due to the institutional effect or due to the change in the composition of participants. Figure 6 displays the decomposition of the reform effect and reveals insights about the extent and magnitude

¹⁸ Additionally, we test whether our matching procedure generates comparable samples following the framework of [Heckman and Hotz \(1989\)](#). [Mueser et al. \(2007\)](#) apply a similar approach in the context of binary treatment. Focusing on the pre-training employment status, we see that both one-step and two-step matching procedures generate comparison groups with employment probabilities prior to participation which are very close to those of the treatment groups.

¹⁹ Of course excluding participants who entered public training programs in the first quarter of 2003 implies that we also exclude participants who entered public training programs in the first quarter of 2002 as well as corresponding non-participants based on our matching algorithm.

²⁰ While participating—or being “locked-in” in the program—individuals probably reduce their search activities for new jobs ([van Ours 2004](#)).

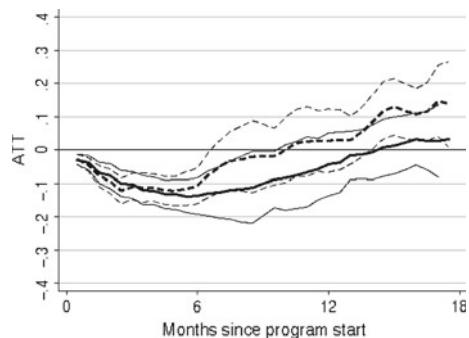
of reform effect, institutional effect, and selection effect. The upper part reports the point estimates of the three effects, while the effects with corresponding confidence intervals are reported in the lower part.

The decomposition shows that the positive reform effect seems to be almost entirely based on the institutional effect. Similar to the reform effect, the institutional effect becomes substantially positive after around 6 months. Despite substantial point estimates, both effects are mostly not significantly different from zero (which could be due to the reduced sample size). The point estimates of the selection effect are almost always virtually zero and never significantly different from zero. This indicates that there is no evidence for a positive impact of a stricter selection of participants on the average treatment effect. Our results, therefore, suggest that the overall reform effect is hardly affected by the change in the composition of participants.

The latter finding is consistent with Heckman et al. (1997, 2002), among others, who find that the impact of creaming on the effectiveness of training programs is modest. On the other hand, our finding of a positive institutional effect is in line with Lechner and Smith (2007) who present evidence that caseworkers are not the best choice to allocate unemployed individuals into programs. Although their results are based on Swiss data, the situation in which caseworkers select the training providers (and programs) on behalf of the unemployed precisely describes the pre-reform situation in Germany. This changed after the reform as job seekers are free to choose their provider on their own by means of training vouchers. As in the case of the effects on employment probabilities, we present the average treatments effects on monthly earnings before and after the reform in Fig. 7. Again, we observe substantial locking-in effects for both periods and obtain larger point estimates for the post-reform period after around 6 months of treatment. 18 months after entering the program, the point estimate of the treatment effect is about € 50 in the pre-reform period, and roughly € 160 per month in the post-reform period.

Figure 8 displays the decomposition of the reform effect in terms of earnings. Similar to the employment probabilities, the positive reform effect seems to be almost entirely based on the institutional effect. We find no significant selection effect and a positive institutional effect, although the latter is mostly not significant. The similarity

Fig. 5 Reform effect, employment (excl. first quarter). Notes Pre-reform period as solid lines, post-reform as dashed lines. Thick lines point estimates, thin lines 95 % confidence intervals (95 % CI).



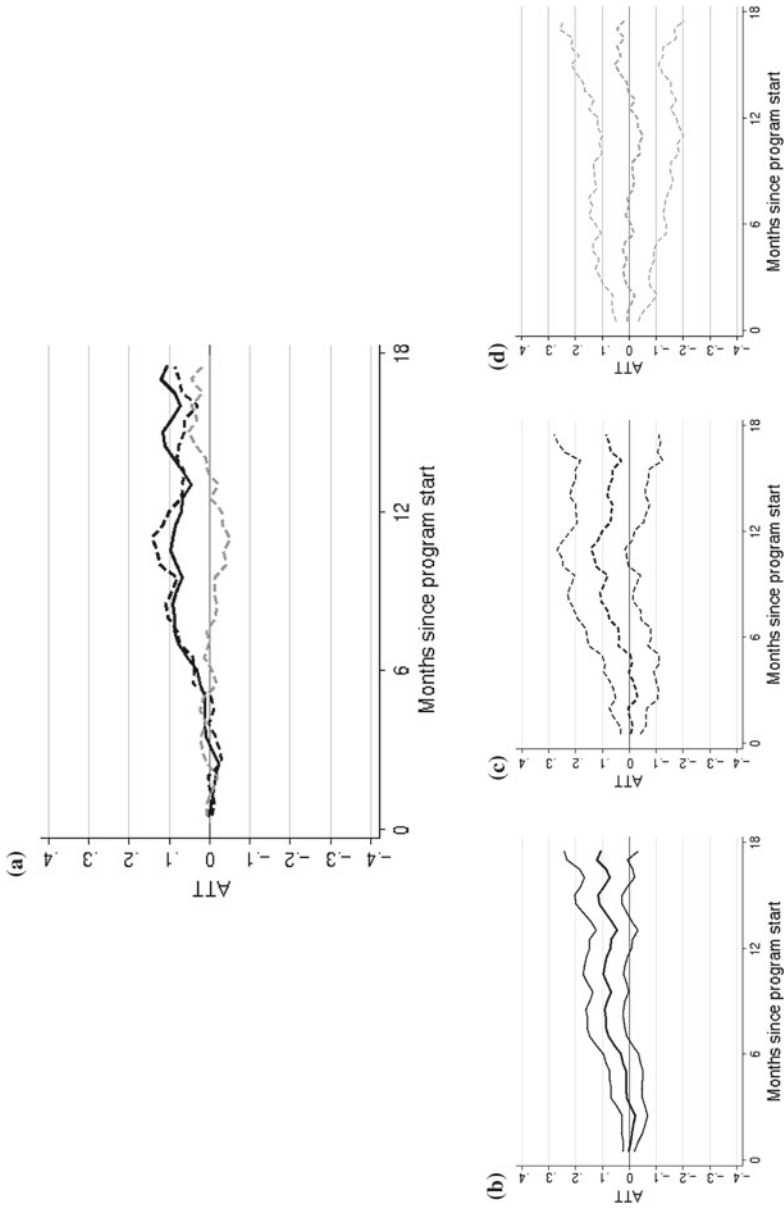
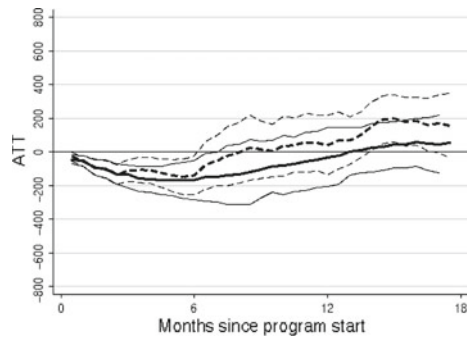


Fig. 6 Decomposition, employment (excl. first quarter). **a** Point estimates of effects. **b** Reform effect. **c** Institutional effect. **d** Selection effect. *Notes:* Total reform effect in black (solid), institutional effect in black (dashed), and selection effect in gray (dashed). Thick lines point estimates, thin lines 95 % confidence intervals

Fig. 7 Reform effect, earnings (excl. first quarter). *Notes* Effects in terms of monthly earnings where no earnings are treated as zero. Pre-reform period as *solid lines*, post-reform as *dashed lines*. *Thick lines* point estimates, *thin lines* 95 % confidence intervals



to the effects on employment probabilities is not surprising given that the positive earnings effects reflect, at least partly, increased employment probabilities.²¹

7.2 Full sample

The results for the full sample on employment probabilities are depicted in Fig. 9. When we do not account for the transitional arrangement in the beginning of 2003, we still observe the main result of a positive impact of the voucher and related institutional changes. Our point estimates are in general of similar magnitude as in the restricted sample, but in the full sample the reform effect and the institutional effect also exhibit statistical significance. The institutional effect is significantly positive from month 7 until month 13 after entering the program. Although this effect is still positive in the last 5 months of our observation period, it is not significantly different from zero anymore. The selection effect is almost always negative, but never significantly different from zero. Results are very similar for earnings, as reported in Fig. 10.

It thus appears that the reform leads to an increase in the training's effectiveness. This finding results both in terms of employment probabilities and earnings, and also both in the restricted sample and in the full sample. We moreover find that the introduction of the voucher and related institutional changes appear responsible for the increased effectiveness. The observed changes in the composition of participants have virtually no impact.

7.3 Effect heterogeneity

So far, our results focus on the average impacts of the reform. In a next step, we investigate and decompose the reform effect for different subgroups to assess whether impacts are heterogeneous across gender, skill group, and region.

²¹ When considering realized earnings, i.e., earnings conditional on being employed, we find that the introduction of the voucher and related institutional changes—next to an increased employment probability—also lead to better job matches for the participants, measured by on average higher monthly earnings in the new job.

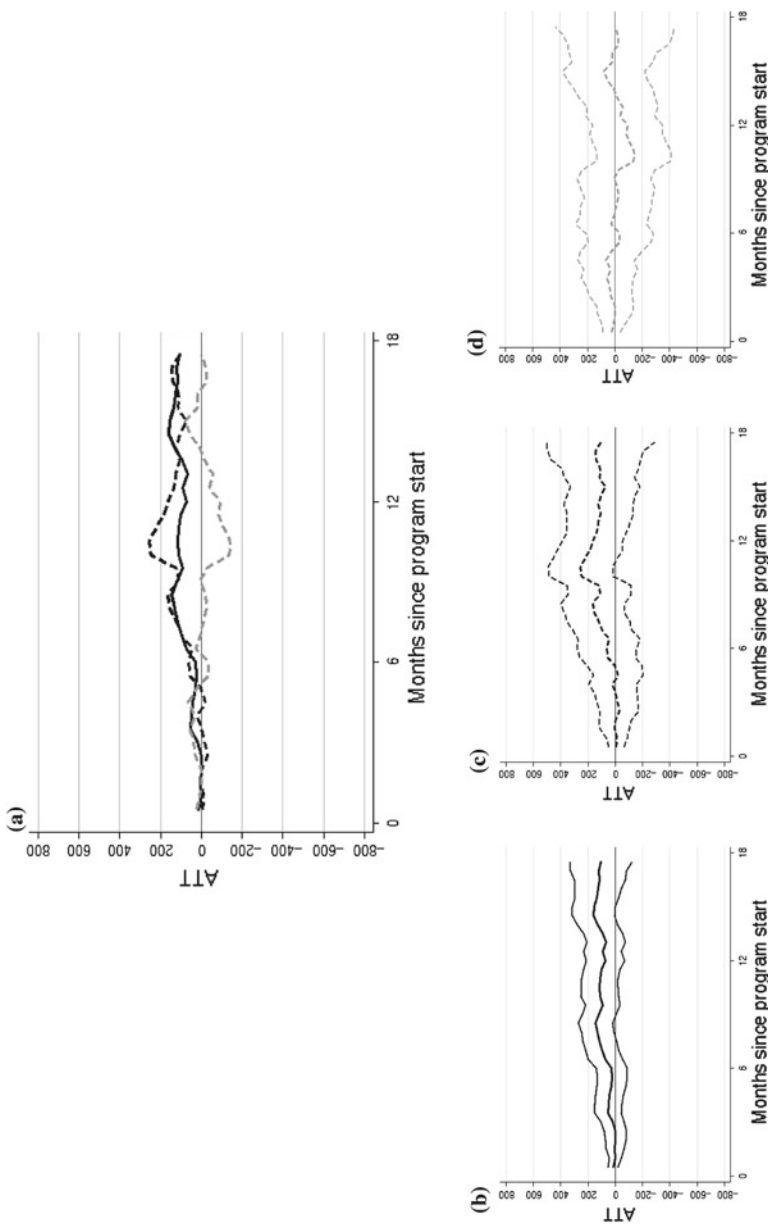


Fig. 8 Decomposition, earnings (excl. first quarter). **a** Point estimates of effects. **b** Reform effect. **c** Institutional effect. **d** Selection effect. *Notes:* Effects in terms of monthly earnings where no earnings are treated as zero. Total reform effect in black (solid), institutional effect in black (dashed), and selection effect in gray (dashed). Thick lines point estimates, thin lines 95 % confidence intervals

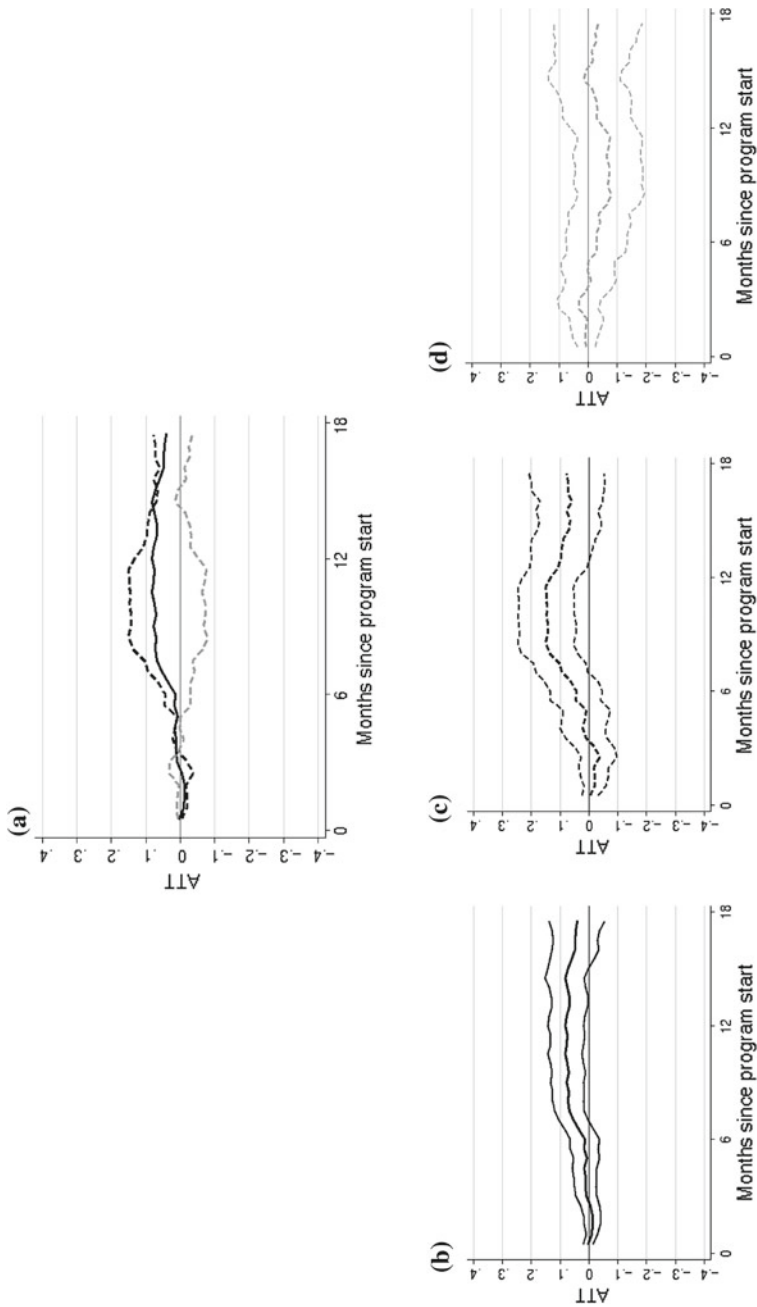


Fig. 9 Decomposition, employment, Point estimates of effects. **a** Point estimates of effects. **b** Reform effect. **c** Institutional effect. **d** Selection effect. *Notes* Total reform effect in black (solid), institutional effect in black (dashed), and selection effect in gray (dashed). Thick lines point estimates, thin lines 95 % confidence intervals

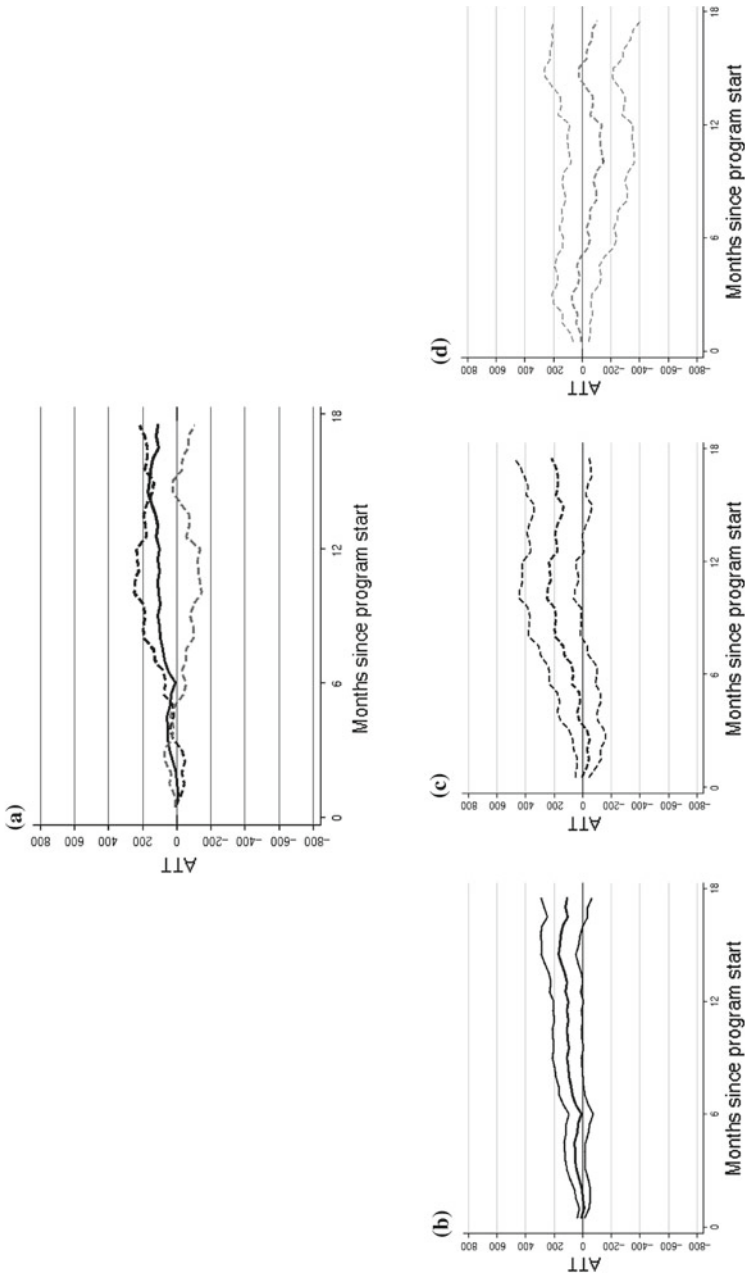


Fig. 10 Decomposition, earnings, **a** Point estimates of effects, **b** Reform effect, **c** Institutional effect, **d** Selection effect. *Notes:* Effects in terms of monthly earnings where no earnings are treated as zero. Total reform effect in black (solid), institutional effect in black (dashed), and selection effect in gray (dashed). Thick lines point estimates, thin lines 95 % confidence intervals

7.3.1 Gender

Although empirical evidence suggests that average treatment effects are of relatively similar magnitude for male and female participants (see, e.g., Rinne et al. 2011), there could be important differences with respect to the impact of the reform on the training's effectiveness across gender. We thus analyze and decompose the reform effect separately for men and women.

Table 5 displays the reform effect, institutional effect, and selection effect in terms of employment probabilities and earnings for male and female individuals at 3, 6, 9, 12, and 15 months after program entry. In terms of both outcomes the reform effect is relatively similar across gender, albeit slightly larger for men. The institutional effect is large and positive for both men and women, but slightly larger for women. The selection effect is, if at all, negative for both male and female participants, although it is relatively more pronounced for women.

It thus appears that the reform had a similar impact across gender. Although female participants could take slightly more advantage of the voucher scheme and related institutional changes, the selection effect appears to be somewhat more negative for this subgroup.

7.3.2 Skill group

Preliminary evidence suggests that low-skilled job-seekers may lack the abilities to navigate the training market and to take an active role in searching for an appropriate course and provider (Kruppe 2009). If this is the case, the advantages of the introduction of vouchers and related institutional changes will only partly occur in this group and mainly occur among skilled individuals. In order to assess this issue in more detail, we differentiate between two skill groups—skilled and unskilled individuals—and analyze the reform effect, institutional effect, and selection effect separately for these groups.

Our classification of skilled and unskilled individuals is based on whether or not an individual has received a formal vocational degree before entering the program.²² This distinction closely follows Dustmann and Meghir (2005) who define skill groups similarly. The importance of this distinction is emphasized as the authors find substantial differences between the two groups in terms of job mobility, wage growth, and returns to experience. In their view, these differences have important implications, e.g., for the design of ALMP.²³

Table 6 displays the results for the subgroups of unskilled and skilled individuals in terms of employment probabilities and earnings. Importantly, we do not find any significant impacts of the reform for the unskilled. All three effects (reform effect, institutional effect, and selection effect) are not significantly different from zero for

²² We consider completed in-plant training and off-the-job training as well as degrees from a vocational school, a technical school, a university, or a university of applied sciences as vocational degrees.

²³ For a broader and more general overview about the German system of secondary school tracks, the apprenticeship system, and vocational degrees which can be obtained, see for example Winkelmann (1996) and Dustmann (2004).

Table 5 Decomposition by gender

Month	Employment			Earnings (in €)		
	Reform effect	Institutional effect	Selection effect	Reform effect	Institutional effect	Selection effect
	<i>Panel A1: Men</i>			<i>Panel A2: Men</i>		
3	0.006 (0.021)	-0.021 (0.037)	0.027 (0.042)	42.97 (39.17)	-16.55 (69.13)	59.52 (79.36)
6	0.002 (0.029)	0.050 (0.051)	-0.048 (0.059)	-10.35 (51.66)	75.29 (96.97)	85.64 (108.72)
9	0.080 (0.034)	0.126 (0.056)	-0.046 (0.069)	103.60 (62.26)	136.47 (108.20)	-32.87 (132.59)
12	0.085 (0.038)	0.123 (0.057)	-0.039 (0.067)	107.41 (67.59)	224.41 (110.70)	-117.00 (134.34)
15	0.084 (0.041)	0.053 (0.062)	0.031 (0.070)	187.71 (77.58)	105.36 (121.73)	82.34 (141.12)
	<i>Panel B1: Women</i>			<i>Panel B2: Women</i>		
3	0.013 (0.020)	-0.022 (0.038)	0.035 (0.041)	40.91 (34.17)	-47.77 (65.81)	88.67 (74.08)
6	0.025 (0.030)	0.032 (0.053)	-0.007 (0.064)	27.69 (48.53)	45.48 (91.57)	-17.79 (105.50)
9	0.062 (0.035)	0.169 (0.060)	-0.107 (0.068)	102.43 (58.36)	263.15 (105.82)	-160.73 (121.06)
12	0.067 (0.038)	0.146 (0.069)	-0.080 (0.081)	63.67 (62.82)	240.74 (113.71)	-177.06 (130.65)
15	0.047 (0.046)	0.073 (0.077)	-0.026 (0.090)	84.11 (76.89)	155.60 (128.54)	-71.49 (156.21)

Bootstrapped standard deviation in brackets

this subgroup during our observation period. On the other hand, we find a significantly positive reform effect and a significantly positive institutional effect for skilled individuals. The selection effect, although not significantly different from zero, is in general negative. This overall pattern holds for both outcomes.

When we differentiate between two skill groups, we, therefore, find that the positive reform effect and the positive institutional effect only arise for skilled individuals. There are a number potential explanations for this finding. For example, it may be the case that skilled participants can take advantage from an increased consumer sovereignty, whereas unskilled individuals have problems in adequately using the newly introduced voucher. However, it should be noted that unskilled individuals who participate in the program are not worse off after the reform, but the institutional changes do not improve the program's effectiveness for this subgroup. Whether unskilled job seekers are free to select the training provider in the market or caseworkers make this choice, therefore, appears not to make a difference.

7.3.3 Region

More than a decade after German reunification, there were still important differences in the economic and labor market conditions between East and West Germany when the vouchers had been introduced in 2003. Moreover, the literature on the effectiveness of training generally finds heterogenous results for the two German regions, at least to some extent (see, e.g., [Lechner et al. 2007, 2011](#)). Therefore, it seems appropriate to analyze the effects of the introduction of the vouchers separately for East and West Germany.

Table 7 shows that the reform effect is larger in East Germany, both in terms of employment probabilities and earnings. The reforms effect is not significantly different from zero in West Germany. We find a positive institutional effect in both German regions, although the effect is again larger in East Germany. The selection effect is, if at all, slightly negative in both German regions.

The differences in the reform impacts between the two German regions are related to our findings on skill groups as the skill distributions for participants in the two German regions differ quite substantially. Whereas about 25 % of the participants are unskilled in West Germany, this share amounts to only 10 % in East Germany. Our finding of a larger reform effect and a larger institutional effect in East Germany is, therefore, in line with our previous result that positive reform impacts only arise for skilled individuals.

7.4 Sensitivity analysis

We address the robustness of our previous results in this section. For this purpose, we perform a sensitivity analysis in which we assess the robustness of our results with respect to the inclusion of additional control variables.

One may argue that changes in the general economic situation constitute another component of the reform effect. Therefore, we additionally control for a number of economic and labor market characteristics which are available for each local FEA

Table 6 Decomposition by skill group

Month	Employment			Earnings (in €)		
	Reform effect	Institutional effect	Selection effect	Reform effect	Institutional effect	Selection effect
	<i>Panel A1: Unskilled</i>			<i>Panel A2: Unskilled</i>		
3	-0.009 (0.024)	-0.036 (0.046)	0.027 (0.052)	12.48 (39.18)	-69.82 (78.21)	82.30 (88.59)
6	0.005 (0.038)	0.020 (0.079)	-0.016 (0.090)	-31.90 (59.26)	-12.74 (133.18)	-19.16 (148.57)
9	-0.019 (0.043)	0.031 (0.091)	-0.049 (0.103)	-82.45 (67.82)	29.54 (146.27)	-111.98 (165.00)
12	-0.019 (0.050)	0.039 (0.089)	-0.058 (0.098)	-58.31 (78.87)	100.23 (148.30)	-158.55 (161.48)
15	-0.017 (0.053)	0.005 (0.108)	-0.022 (0.121)	-63.06 (91.00)	-11.96 (172.43)	-51.10 (193.89)
	<i>Panel B1: Skilled</i>			<i>Panel B2: Skilled</i>		
3	0.014 (0.019)	-0.018 (0.035)	0.031 (0.039)	51.59 (34.97)	-17.43 (64.02)	69.02 (73.56)
6	0.016 (0.027)	-0.050 (0.047)	-0.033 (0.055)	21.16 (45.96)	83.11 (87.34)	-61.96 (98.80)
9	0.097 (0.031)	0.168 (0.052)	-0.071 (0.063)	154.96 (55.61)	223.96 (98.79)	-69.00 (121.05)
12	0.103 (0.034)	0.154 (0.053)	-0.051 (0.063)	135.15 (57.97)	266.30 (101.87)	-131.15 (124.84)
15	0.093 (0.038)	0.079 (0.058)	0.014 (0.067)	207.52 (70.24)	171.34 (110.77)	36.18 (130.67)

Boostrapped standard deviation in brackets

Table 7 Decomposition by region

Month	Employment			Earnings (in €)		
	Reform effect	Institutional effect	Selection effect	Reform effect	Institutional effect	Selection effect
	<i>Panel A1: West Germany</i>			<i>Panel A2: West Germany</i>		
3	0.013 (0.020)	-0.010 (0.036)	0.023 (0.039)	49.26 (37.08)	-11.68 (67.43)	60.94 (75.22)
6	0.002 (0.028)	0.022 (0.054)	-0.020 (0.063)	-7.09 (50.28)	54.95 (100.36)	-62.04 (111.58)
9	0.048 (0.033)	0.112 (0.062)	-0.064 (0.074)	100.25 (63.47)	168.82 (120.35)	-68.57 (144.47)
12	0.029 (0.038)	0.117 (0.060)	-0.088 (0.075)	46.53 (70.39)	209.89 (125.76)	-163.35 (156.42)
15	0.025 (0.042)	0.020 (0.065)	0.004 (0.077)	77.72 (79.95)	75.79 (129.30)	1.93 (157.03)
	<i>Panel B1: East Germany</i>			<i>Panel B2: East Germany</i>		
3	0.005 (0.020)	-0.035 (0.040)	0.040 (0.045)	34.90 (36.12)	-47.95 (70.36)	82.85 (80.29)
6	0.026 (0.032)	0.070 (0.053)	-0.044 (0.061)	23.30 (49.74)	76.54 (92.40)	-53.24 (105.95)
9	0.104 (0.036)	0.179 (0.057)	-0.075 (0.067)	105.75 (59.35)	208.89 (101.35)	-103.14 (119.74)
12	0.137 (0.038)	0.153 (0.063)	-0.016 (0.073)	135.58 (61.26)	263.89 (108.09)	-128.32 (125.40)
15	0.119 (0.044)	0.112 (0.070)	0.007 (0.078)	220.53 (75.71)	199.32 (124.76)	21.21 (139.79)

Bootstrapped standard deviation in brackets

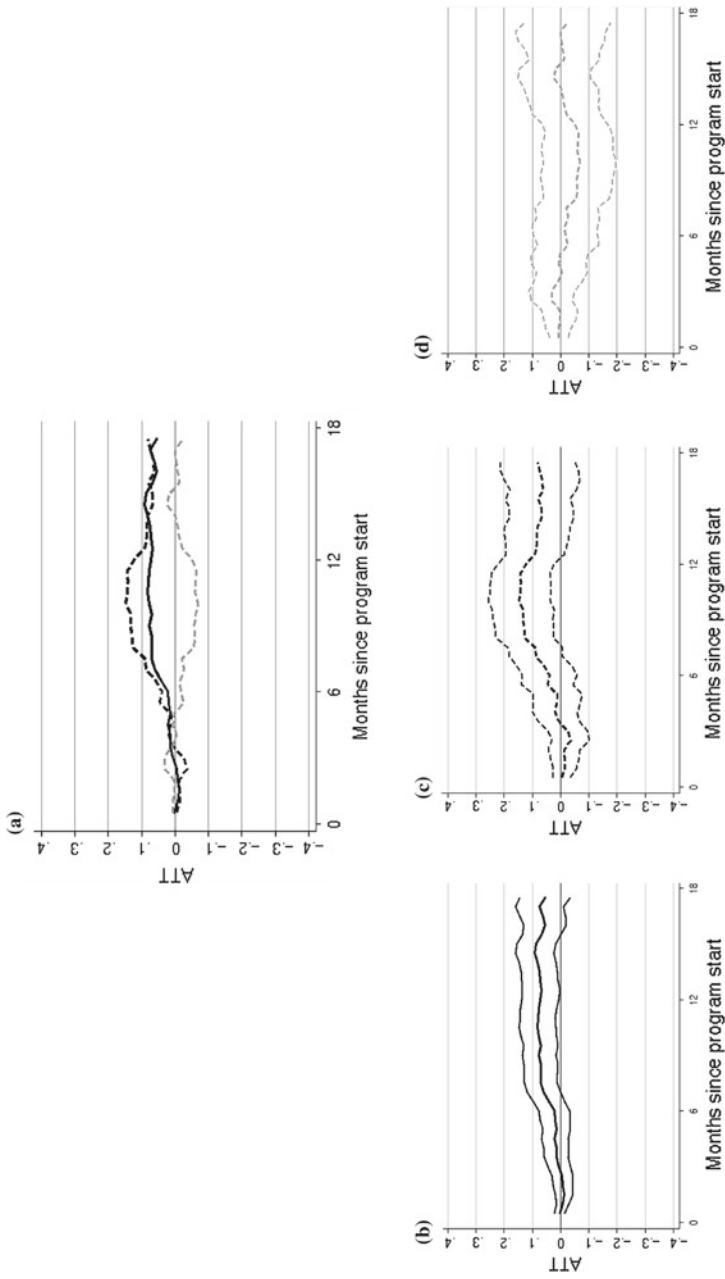


Fig. 11 Decomposition, employment (incl. additional control variables). **a** Point estimates of effects. **b** Reform effect. **c** Institutional effect. **d** Selection effect. *Notes* Total reform effect in *black (solid)*, institutional effect in *black (dashed)*, and selection effect in *gray (dotted)*. *Thick lines* point estimates, *thin lines* 95 % CI

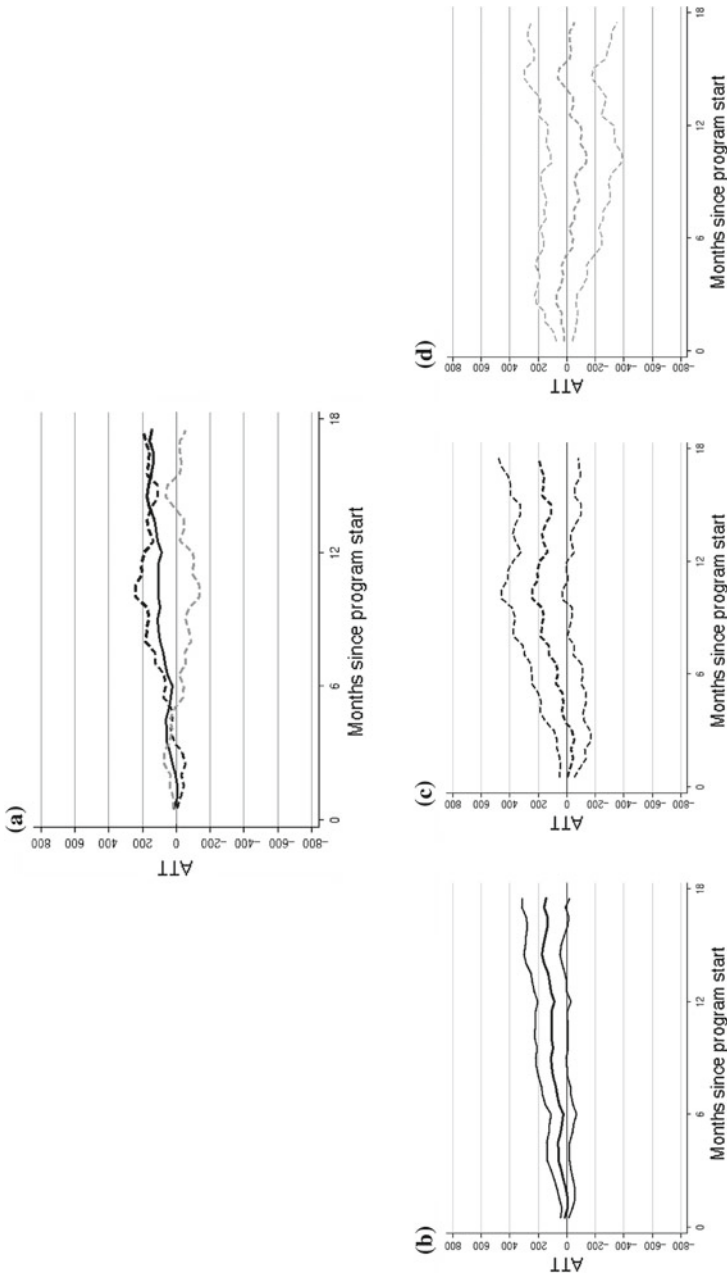


Fig. 12 Decomposition, earnings (incl. additional control variables). **a** Point estimates of effects. **b** Reform effect. **c** Institutional effect. **d** Selection effect. *Notes* Effects in terms of monthly earnings where no earnings are treated as zero. Total reform effect in *black (solid)*, institutional effect in *black (dashed)*, and selection effect in *gray (dashed)*. *Thick lines* point estimates, *thin lines* 95 % CI.

district. These variables are changing over time.²⁴ In addition to that, we include observable individual characteristics measured before entering the treatment and also include—only for the post-reform period—indicators describing the implementation of the Hartz reform on the FEA district level.

The results with respect to employment probabilities are presented in Fig. 11. In general, the picture is very similar to the results presented above. The point estimates of the institutional effect are slightly lower, while the selection effect is slightly less negative. However, the institutional effect is still significantly positive between month 7 and month 13 after entering the program, and the selection effect is still almost always negative.

The results are also very similar for earnings, as reported in Fig. 12. The institutional effect is only marginally lower and the selection effect slightly increases. Our results thus appear to be robust to the inclusion of additional control variables.

8 Conclusions

This paper analyzes the impact of the labor market reform in 2003 on the effectiveness of the most important type of public training program in Germany. The reform had two main features: (a) the introduction of training vouchers and (b) the application of more selective criteria on participants. Next to estimating the overall impact, we decompose the reform effect into an institutional effect and a selection effect.

We find a slightly positive impact of the reform. The decomposition of this overall effect shows that the selection effect is, if at all, slightly negative. This finding is in line with the literature on performance standards that typically reports the modest impacts, if at all. The introduction of the training voucher and related institutional changes, on the other hand, increased both the employment probability and earnings of participants. The institutional effect becomes substantially positive after around 6 months of training, and it decreases slightly at the end of our observation period (1.5 years after program entry). However, this seems to be only the case for skilled participants as we do not find any significant institutional effect (or reform effect) for the unskilled.

We study the effects of the introduction of vouchers on participants who entered training in 2003, i.e., the first year after the vouchers were introduced. This means that we do not consider long-run impacts of their introduction, which were, however, interesting to study. It may, for example, take some time for the market to react. Both demand and supply of training may only adjust over a longer period. Furthermore, our results are based only on the sub-group of individuals for which we find suitable matches in our data. Nevertheless, our findings appear promising with respect to the idea of vouchers being able to raise the effectiveness of training. Despite this important first result, an interesting question is left to be answered: How do vouchers exactly impact the effectiveness of public training? It may be due to changes in training

²⁴ The business cycle might affect the composition of the unemployed who are eligible for training. While our sensitivity analysis explicitly accounts for differences in the business cycle in different regions of Germany, it may not be able to control for changes in the global trend.

demand, due to changes in training supply, or even due to a combination of both. Future research may pursue this avenue. Future research may additionally investigate whether training vouchers generate potentially important intention-to-treat effects. Our data do not contain information about voucher receipt, and, therefore, the question if issuing a voucher already has an impact on the recipients is beyond the scope of this paper.

Acknowledgments We would like to thank the anonymous referees, Marton Csillag, Bernd Fitzenberger, Steffen Künn, Peter R. Mueser, Núria Rodríguez-Planas, Jeffrey A. Smith, Hilmar Schneider, Marc Schneider, and conference participants in Bonn, Kiel, London, Mannheim, Munich, Nuremberg, and Oslo for valuable discussions and helpful comments. We also thank *infas* and especially Doris Hess and Helmut Schröder for providing supplementary information about the reform under study. Arne Uhlendorff thanks DIW DC where part of this research was pursued. Zhong Zhao acknowledges financial support from the Capacity Building Project for the Economics Discipline Group at Beijing Municipal Universities (2010–2012). All remaining errors are our own.

References

- Abadie A, Imbens GW (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267
- Abadie A, Imbens GW (2008) On the failure of the bootstrap for matching estimators. *Econometrica* 76(6):1537–1557
- Barnow BS (2000) Vouchers for federal targeted training programs. In: Peterson GE, Reischauer RD, Steuerle CE, Doorn Ooms V (eds) *Vouchers and the provision of public services*. Brookings Institution Press, Washington, DC
- Barnow BS (2009) Vouchers in U.S. Vocational Training Programs: An Overview of What We Have Learned. *J Labour Market Res* 42(1):71–84
- Bell S, Orr L (2002) Screening (and creaming?) Applicants to job training programs: the AFDC homemaker-home health aide demonstration. *Labour Econ* 9(2):279–302
- Biewen M, Fitzenberger B, Osikominu A, Waller M (2007) Which program for whom? Evidence on the comparative effectiveness of public sponsored training programs in Germany. IZA discussion paper 2885, Institute for the Study of Labor (IZA), Bonn
- Bloom H (1984) Accounting for no-shows in experimental evaluation designs. *Eval Rev* 8(2):225–246
- Bruttel O (2005) Delivering active labor market policy through vouchers: experiences with training vouchers in Germany. *Int Rev Adm Sci* 71(3):391–404
- Caliendo M (2009) Income support systems, labor market policies and labor supply: the German experience. IZA discussion paper 4665, Institute for the Study of Labor (IZA), Bonn
- Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *J Econ Surv* 22(1):31–72
- Caliendo M, Hujer R, Thomsen SL (2008) The employment effects of job creation schemes in Germany: a microeconomic evaluation. In: Millimet DL, Smith JA, Vytlačil EJ (eds) *Estimating and evaluating treatment effects in econometrics, advances in econometrics*, vol 21. Emerald Group Publishing, Bingley, pp 383–430
- Card D, Hyslop DR (2005) Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica* 73(6):1723–1770
- Dustmann C (2004) Parental background, secondary school track choice, and wages. *Oxf Econ Papers* 56(2):209–230
- Dustmann C, Meghir C (2005) Wages, experience and seniority. *Rev Econ Stud* 72(1):77–108
- Fitzenberger B, Osikominu A, Völter R (2006) Imputation rules to improve the education variable in the IAB employment subsample. *Schmollers Jahrbuch* 126(3):405–436
- Fitzenberger B, Osikominu A, Völter R (2008) Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany. *Annales d’Economie et de Statistique* 91/92:321–355
- Fitzenberger B, Osikominu A, Paul M (2010) The heterogeneous effects of training incidence and duration on labor market transitions. IZA discussion paper 5269, Institute for the Study of Labor (IZA), Bonn

- Flores C, Flores-Lagunes A, Gonzalez A, Neuman T (2012) Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Rev Econ Stat* 94(1):153–171
- Fortin NM, Lemieux T, Firpo S (2011) Decomposition methods in economics. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 4A, chap 1. Elsevier, Amsterdam, pp 1–102
- Friedman M (1962) The role of government in education. In: *Capitalism and freedom*, chap 6. University of Chicago Press, Chicago
- Görlitz K (2010) The effect of subsidizing continuous training investments—evidence from German establishment data. *Labour Econ* 17(5):789–798
- Ham JC, Li X, Reagan PB (2011) Matching and semi-parametric IV estimation, a distance-based measure of migration, and the wages of young men. *J Econ* 161(2):208–227
- Heckman JJ, Hotz VJ (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J Am Stat Assoc* 84(408):862–874
- Heckman JJ, Heinrich C, Smith JA (1997) Assessing the performance of performance standards in public bureaucracies. *Am Econ Rev* 87(2):389–395
- Heckman JJ, Ichimura H, Smith JA, Todd PE (1998) Characterizing selection bias using experimental data. *Econometrica* 66(5):1017–1098
- Heckman JJ, Heinrich C, Smith JA (2002) The performance of performance standards. *J Human Resour* 37(4):778–811
- Hipp L, Warner ME (2008) Market forces for the unemployed? Training vouchers in Germany and the USA. *Social Policy Adm* 42(1):77–101
- Hujer R, Thomsen SL, Zeiss C (2006) The effects of vocational training programmes on the duration of unemployment in Eastern Germany. *All Stat Arch* 90(2):299–322
- Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat* 86(1):4–29
- Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47(1):1–81
- Jacobi L, Kluge J (2007) Before and after the Hartz reforms: the performance of active labour market policy in Germany. *J Labour Market Res* 40(1):45–64
- Kluge J (2010) The effectiveness of European active labor market policy. *Labour Econ* 17(6):904–918
- Kluge J, Schneider H, Uhlendorff A, Zhao Z (2012) Evaluating continuous training programs using the generalized propensity score. *J R Stat Soc Ser A* 175(2):587–617
- Kruppe T (2009) Bildungsgutscheine in der aktiven Arbeitsmarktpolitik. *Sozialer Fortschritt* 58(1):9–19
- Ladd HF (2002) School vouchers: a critical review. *J Econ Perspect* 16(4):3–24
- LaLonde RJ (2003) Employment and training programs. In: Moffitt R, Feldstien M (eds) *Means tested transfer programs in the U.S.* University of Chicago Press for the National Bureau of Economic Research, Chicago
- Lechner M, Smith JA (2007) What is the value added by caseworkers? *Labour Econ* 14(2):135–151
- Lechner M, Wunsch C (2008) What did all the money do? On the general ineffectiveness of recent West German labour market programmes. *Kyklos* 61(1):134–174
- Lechner M, Wunsch C (2009) Are training programs more effective when unemployment is high? *J Labor Econ* 27(4):653–692
- Lechner M, Wunsch C (2011) Sensitivity of matching-based program evaluations to the availability of control variables. IZA discussion paper 5553, Institute for the Study of Labor (IZA), Bonn
- Lechner M, Miquel R, Wunsch C (2007) The curse and blessing of training the unemployed in a changing economy: the case of East Germany after unification. *Ger Econ Rev* 8(4):468–509
- Lechner M, Miquel R, Wunsch C (2011) Long-run effects of public sector sponsored training in West Germany. *J Eur Econ Assoc* 9(4):742–784
- Leuven E, Sianesi B (2003) PSMATCH2: stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical software components. Boston College Department of Economics, available at <http://ideas.repec.org/c/boc/bocode/s432001.html>
- Mueser PR, Troske KR, Gorislavsky A (2007) Using state administrative data to measure program performance. *Rev Econ Stat* 89(4):761–783
- Neal D (2002) How vouchers could change the market for education. *J Econ Perspect* 16(4):25–44
- Neyman JS (1923) On the application of probability theory to agriculture experiments. *Essay on principles*. Section 9. *Roczniki Nauk Rolniczych Tom X*:1–51 (*Ann Agric Sci*), translated in: Neyman JS, Dabrowska DM, Speed TP (1990) *Stat Sci* 5(4):465–472

- Rinne U, Schneider M, Uhlendorff A (2011) Do the skilled and prime-aged unemployed benefit more from training? Effect heterogeneity of public training programs in Germany. *Appl Econ* 43(25):3465–3494
- Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
- Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 39(1):33–38
- Roy AD (1951) Some thoughts on the distribution of earnings. *Oxf Econ Papers* 3(2):135–146
- Rubin DB (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29:185–203
- Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66(5):688–701
- Rubin DB (2006) *Matched sampling for causal effects*, 1st edn. Cambridge University Press, Cambridge
- Schneider H, Uhlendorff A (2006) Die Wirkung der Hartz-Reform im Bereich der beruflichen Weiterbildung. *J Labour Market Res* 39(3–4):477–490
- Schneider H, Brenke K, Jesske B, Kaiser L, Rinne U, Schneider M, Steinwede J, Uhlendorff A (2007) Evaluation der Maßnahmen zur Umsetzung der Vorschläge der Hartz-Kommission—Bericht 2006. IZA Research Report 10, Institute for the Study of Labor (IZA), Bonn
- Sianesi B (2004) An evaluation of the active labor market programmes in Sweden. *Rev Econ Stat* 86(1):133–155
- Smith JA, Todd PE (2005a) Does matching overcome LaLonde’s critique of non-experimental estimators? *J Econ* 125(1–2):305–353
- Smith JA, Todd PE (2005b) Rejoinder. *J Econ* 125(1–2):365–375
- Stephan G (2008) The effects of active labor market programs in Germany: an investigation using different definitions of non-treatment. *Jahrbücher für Nationalökonomie und Statistik* 228(5–6):586–611
- Steuerle CE (2000) Common issues for voucher programs. In: Peterson GE, Reischauer RD, Steuerle CE, Doorn Ooms V (eds) *Vouchers and the provision of public services*. Brookings Institution Press, Washington DC
- Tsiatis A (2006) *Semiparametric theory and missing data*, 1st edn. Springer, Berlin
- van Ours JC (2004) The locking-in effect of subsidized jobs. *J Comp Econ* 32(1):37–55
- Winkelmann R (1996) Employment prospects and skill acquisition of apprenticeship-trained workers in Germany. *Ind Labor Relat Rev* 49(4):658–672
- Winterhager H, Heinze A, Spermann A (2006) Deregulating job placement in Europe: a microeconomic evaluation of an innovative voucher scheme in Germany. *Labour Econ* 13(4):505–517
- Zhao Z (2008) Sensitivity of propensity score methods to the specifications. *Econ Lett* 98(3):309–319