Empirical Essays on Job Search Behavior, Active Labor Market Policies, and Propensity Score Balancing Methods

Inaugural-Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Wirtschafts- und Sozialwissenschaft (Dr. rer. pol.)
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Universität Potsdam

vorgelegt von

Diplom-Volkswirtin Ricarda Schmidl
geboren am 10. November 1981 in Chemnitz
wohnhaft in Mannheim

— eingereicht im Februar 2014 —

Erstgutachter:	Prof. Dr. Marco Caliendo
Zweitgutachter:	Prof. Dr. Alexander S. Kritikos
Drittgutachter:	Prof. Dr. Rainald Borck
Tag der Disputation:	15. Mai 2014

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2014/7114/
URN urn:nbn:de:kobv:517-opus-71145
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-71145

Acknowledgements

Many people have contributed to this thesis. My supervisor, Marco Caliendo was the one to encourage me to start writing, and I am very grateful for his continuous guidance and support over the whole period of completion and even beyond. The cooperation with my co-authors Steffen Künn and Arne Uhlendorff was both productive and fun, and I have benefited a lot from many fruitful discussions with them along the way.

Most part of this thesis was written during my time as a Resident Research Affiliate at the Institute for the Study of Labor (IZA) in Bonn, to which I am very much indebted for providing an outstanding and inspiring research environment. During this time, I have met many interesting and important people that made "all the difference", and some of which luckily continue to be part of my life. Gerard van den Berg gave me the opportunity to finish the last parts of the dissertation at the University of Mannheim - I am very thankful for this and for many interesting discussions on my research since the beginning.

My family and my friends have given me invaluable moral support with their love and incessant belief in my abilities, and certainly deserve most of the credit. In loving memory I dedicate this thesis to my grandmother Charlotte who probably would have been very proud to see it finished.

Contents

List of Tables

List of Figures

Abbreviations

ALMP	Active labor market policies
CART	Classification and regression trees
CIA	Conditional independence assumption
CDF	Cumulative distribution function
CMS	Cramer-von-Mises-test
DID	Difference-in-Difference
FT	Further training measures, medium to long-term
IAB	Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung)
IEB	Integrated Employment Biographies (Integrierte Erwerbsbiographien)
IPT	Inverse probability tilting
IPW	Inverse probability weighting
IV	Instrumental variable
JS	Job search measures
JCS	Job creation schemes
JUMP	Immediate action program for lowering youth unemployment (Jugend mit Perspektive)
JWS	Wage subsidies within the JUMP program
KM	Kernel matching
KS-test	Kolmogororv-Smirnov-test
LLM	Local linear matching
LRM	Local linear matching with a ridge term
LOOCV	Leave-one-out cross-validation
LPM	Local polynomial matching
NN	Nearest neighbor matching
MISE	Minimal integrated squared error
MSB	Mean standardized bias
PES	Public employment services
PS	Propensity score
PSM	Propensity score matching
PT	Preparatory training
SGB III	German social code (Sozialgesetzbuch Drittes Buch)

SB	Standardized bias
STT	Short-term trainin measures
SUTVA	Stable unit treatment value assumption
TWA	Temporary work agency
UB	Unemployment benefits
VI	Vacancy information
WS	Wage subsidies

Introduction

As many countries have made programmatic shifts from passive towards active unemployment policies, empirical labor market research aims to provide evidence-based guidelines for the optimal design of active labor market policies (ALMP). This thesis extends this literature in multiple directions. The first three chapters of the thesis are empirical analyses aimed at increasing our understanding of the determinants of unemployed job search, and the effectiveness of interventions targeted at increasing the reemployment probability of the unemployed. The fourth chapter is a methodological contribution, addressing the practical challenges in the implementation of two statistical balancing methods commonly used in empirical labor market evaluations.

Motivation

The incidence and persistence of spells of unemployment may have negative consequences on the individuals experiencing joblessness, and the society as a whole. Research on the negative private consequences of unemployment spells suggests detrimental effects on multiple areas in life. Spells of unemployment usually represent periods of financial distress, which may require unemployed to change their consumption behavior and potentially their need to increase debt (Stephens, 2001; Sullivan, 2008). Furthermore, spells of unemployment are associated with a deterioration of the psychological wellbeing of individuals (Darity and Goldsmith, 1996), and may have long-lasting negative consequences on physical health (Burgard et al., 2007; Browning and Heinesen, 2012). The experience of longer-term unemployment seems to be particularly detrimental for subsequent labor market outcomes, as they are found to develop negative duration dependence (Kroft et al., 2013), and may have persistent negative effects on subsequent wage earnings (Gregory and Jukes, 2001). Furthermore, there is evidence that cognitive skills depreciate with increasing duration of the non-employment spells (Edin and Gustavsson, 2008). The societal costs of unemployment range from increased financial burdens due to higher expenditures on unemployment benefits and social assistance, and foregone payroll taxes, but also need to take into account of potentially increased social problems as increased crime rates (Raphael and Winter-Ebmer, 2001).

Against this background, a substantial interest lies in implementing labor market policies that effectively reduce the risk of unemployment persistence and promote the take up of stable employment relationships. Remedial active labor market policies (ALMP) have played a particularly prominent role in the public policy debate, as they suggest a promising approach to quickly overcome temporary problems of mismatch between supply and demand in the labor market (Calmfors, 1994). Hence, following several decades of increasing unemployment rates in many OECD countries, the OECD Jobs Strategy in 1994 (OECD, 1993, 1994), strongly promoted a paradigm shift in unemployment policies, suggesting a stronger emphasis on active labor market policies, moving away from passive income support measures. The call for a more intensified use of ALMP also reverberated in the discussions of European Employment Strategy (EES) in the same year, which finally resulted in the promotion of active labor market schemes for vulnerable labor market groups, as youth, long-term unemployed and women, across many European countries. Today, most OECD countries have incorporated

ALMP in their standard set of labor market policies. During the recent economic crisis, ALMP were used intensively to moderate the detrimental employment impact of low economic demand. From 2007 to 2010 average public spending in OECD countries increased significantly from 0.51% to 0.65% of GDP.

In broad terms, four distinct types of barriers to entry into the labor market can be identified. First, unemployed may lack skills or experience to gather information about vacancies, send out applications, or performing job interviews, which may result in a decreased productivity of search, and hence an increase the duration of unemployment. Second, the unemployed may have acquired working skills that are not in line with the demand for skills of local or overall firms, resulting in unemployment due to geographical or structural mismatch. Third, firms may have a structurally reduced demand for labor that is lower than the supply, which may arise under bad economic condition or institutional settings that raise the cost of labor over the returns to labor (Layard and Nickell, 1986; Layard et al., 2005). Fourth, worker may have low incentives to search, due to high levels of non-pecuniary or pecuniary value of unemployment, potentially driven by high levels or extended duration of unemployment benefits (Meyer, 1990).

Different types of ALMP aim to overcome different types of barriers. Short-term job search assistance measures, intensified job matching services of the public employment services, or labor market counseling aim to improve the productivity of search. Intensive training courses, subsidized further education, but also job creation in the public sector are aimed to enhance the labor market relevant human capital endowment of the unemployed, and provide work experience for the least qualified. Temporary wage subsidies are intended to stimulate labor demand by reducing the costs of initial hiring, combined with incentives for on-the-job training as this may increase the productivity of the workers on the job. Finally, by monitoring search activities and threatening with benefit sanctions in case of non-compliance it is aimed to stimulate a higher intensity of search. While these ALMP are indisputably an important part of the unemployment policies, the question which measures work best for whom has been subject of ongoing debate.

A large empirical literature has emerged measuring the effectiveness of these various ALMP measures, taking account of varying institutional and macroeconomic settings as well as heterogenous participant characteristics. Most commonly these studies are partial impact evaluations assessing the effectiveness of different ALMP on the short and longrun employment outcomes of the participating unem-

ployed. Aggregated and condensed in meta-analyses by, e.g., Martin and Grubb (2001), Kluve (2010), Card et al. (2010) and Bergemann and van den Berg (2008), the seemingly heterogenous contributions of the literature could be merged into a rather homogenous picture of general patterns of effectiveness of the different programs, that is surprisingly stable across institutional settings, and macroeconomic conditions (Kluve, 2010). The general conclusions to emerge are that public sector job creation are not very effective, whereas job search assistance measures are, particularly on the short run. Second, longer-term training measures exhibit significant locking-in effects during programs participation; as unemployed reduce their search activity during participation the program. The long-run effects of training programs on employment rates are commonly found to be positive on the long-run, whereas the magnitude of the effect is not quite large. Furthermore, they are found to be more effective for women than for men, and to have a higher effectiveness under bad economic conditions. Third, wage subsidies and monitoring and sanctioning measures are also found to be very effective in increasing the direct labor market entry. Finally, heterogeneous effects by age of participants suggest that youth benefits less from ALMP than adult workers (Kluve, 2010). When interpreting these findings it has to be kept in mind in that these studies only represent partial evaluations ignoring substitution effects and deadweight effects of wage subsidies and job creation schemes (Calmfors, 1994). Similarly, by focussing on *ex-post* treatment effects, this literature does not account for the *ex-ante* treatment participation, although there is growing evidence of positive employment effects of anticipated ALMP entry (Bergemann et al., 2011; Rosholm and Svarer, 2008).

While the previous evaluation literature has immensely contributed to improving our understanding about the overall effectiveness of ALMP in reducing unemployment, there remain open questions about design changes of effective programs that may reduce unintended negative effects of ALMP participation. For example, it has been found that job search assistance, or imperfect monitoring may result in crowding-out of own search effort (van den Berg and van der Klaauw, 2006; Fougère et al., 2009). Also, monitoring and sanctioning may have negative effects on the labor force participation and lower the quality of accepted employment relationships (Petrongolo, 2009; Arni et al., 2013). Furthermore, given that large negative locking-in effects can be reduced by delaying intensive training programs to later stages in unemployment, the question arises how to optimally time

different program sequences over the course of the unemployment spell taking into account cost-efficiency of the activation process (Pavoni and Violante, 2007; Wunsch, 2013; Spinnewijn, 2013). Finally, the question remains how to minimize negative externalities of program participation, as substitution or displacement effects on non-participants (Gautier et al., 2012; Crépon et al., 2013).

One way to tackle the problem of search substitution is to gain a better understanding about specific factors promoting or hindering the individual-specific search productivity. A powerful theoretical framework for this is given by partial models of job search, that view the job search decisions of the unemployed as a utility maximization problem, with the level of unemployment benefits, search costs, and expected returns to search as arguments (Mortensen, 1986). If specific barriers to productive search were known, the search effort of unemployed could be targeted more efficiently. A further promising approach to improving the timing and use of ALMP is to take explicit account of the objective function of the caseworkers at the public employment services (PES), as they are important decision makers in the activation process and commonly enjoy a high degree of discretionary power. Previous research suggest that incentive structures in the PES (Heckman et al., 1996, 1997), the caseload (Hainmueller et al., 2011) as well as individual-specific characteristics and attitudes (Behncke et al., 2010a,b) affect the caseworker decision making and also affect the reemployment rates of the unemployed.

A further important avenue of research is the identification of effective labor market polices for particularly disadvantaged labor market groups. Low-educated unemployed and youth tend to be most affected by unemployment during periods of economic downturns, and are often found to exhibit the lowest levels of labor market attachment (OECD, 2013). As these labor market groups may also be farthest away from a direct and long-lasting labor market entry, it is likely that they require more intensified measures to reintegrate them into the labor market.

This thesis aims to generate further insights about the benefits and limitations of ALMP, by providing empirical evidence that contribute to the outlined open questions. The empirical work of the thesis focuses on the institutional and economic context of Germany. Similar to many other countries, the targeted activation of unemployed has become a central element of the unemployment policies in Germany. In the structural "Hartz"-reforms of the German labor market, policy makers coined the catchphrase "Promote and Demand" ("Fördern und Fordern")

to describe the reforms of activation policies. On the one hand, it is attempted to foster and strengthen the personal search efforts of unemployed via monitoring and counseling. On the other hand, structural barriers to labor market entry are to be removed by providing training measures or offering wage subsidies for disadvantaged labor market groups (Eichhorst and Ebbinghaus, 2009). At the same time, the importance of empirical program evaluation was emphasized, providing evidence-based advise for policy makers to construct, design and implement ALMP schemes in the most efficient way. While we are hence not the first to address the effectiveness of ALMP in Germany (Jacobi and Kluve, 2007), we aim to provide new insights with respect to the search behavior of the unemployed, the role of early activation in promoting or reducing further participation in ALMP, and the effectiveness of activation measures for unemployed youth. The empirical contributions are complemented by an extensive overview of the recent literature regarding the practical implementation of semi-parametric balancing methods that are very often applied in evaluation research.

Contribution of the Thesis

The empirical studies in Chapters 1 to 3 are based on data from the *IZA Evaluation Dataset S*, which consists of an administrative part and a survey part. The survey part of the *IZA Evaluation Dataset S* is based on a representative sub-sample of monthly entries in unemployment between June 2007 and May 2008. The data is constructed as a longitudinal panel study, whereby an extensive baseline interview is conducted shortly after unemployment entry, and two further interviews followed one and three years after unemployment entry respectively. The longitudinal data set up allows to construct a detailed labor market biography including spells of employment, unemployment, inactivity and ALMP participation (Caliendo et al., 2011). This data is used in the empirical studies of Chapter 1 and 2. The administrative part of the data is based on the administrative records of the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB) and consists of an random inflow sample of entries into unemployment between 2001 and 2008. The data contain about 900,000 individuals, for which detailed daily information about spells in employment, unemployment and active labor market participation is available. This data set is used in Chapter 3.

In Chapter 1 of the dissertation, the role of social networks is analyzed as an

important determinant in the search behavior of the unemployed. Based on the hypothesis that the unemployed generate information on vacancies through their social network, search theory predicts that individuals with large social networks should experience an increased productivity of informal search, and reduce their search in formal channels. Due to the higher productivity of search, unemployed with a larger network are also expected to have a higher reservation wage than unemployed with a small network. The model-theoretic predictions are tested and confirmed empirically. The regression results show that the search behavior of unemployed is significantly affected by the presence of social contacts. Larger networks imply larger substitution away from formal search channels towards informal channels. The substitution is particularly strong for passive formal search methods, i.e., search methods that generate rather non-specific types of job offer information at low relative cost. We also find small but significant positive effects of an increase of the network size on the reservation wage. These results have hence important implications on the analysis of the job search monitoring or counseling measures that are usually targeted at formal search (van den Berg and van der Klaauw, 2006). As unemployed substitute between passive formal channels and informal channels, monitoring efforts should either take into account the number of friends or focus monitoring on active search channels to avoid crowding out of search.

Chapter 2 of the dissertation addresses the labor market effects of vacancy information during the early stages of unemployment. The aim of the analysis is to measure the effects of early vacancy information on the exit rate into employment, and the effects on the quality of employment. Furthermore, the short-and medium-term effects to participate in more intensive active labor market programs are analyzed. These results show that vacancy information significantly increases the speed of entry into employment; at the same time the probability to participate in ALMP is significantly reduced. For men and West German unemployed the long-term reduction in the participation probability can be seen as a consequence of the increased employment probability. For unemployed in East Germany however, an early significant but temporary reduction in the participation probability indicates that high and low activation measures are used interchangeably from the perspective of the caseworker, which is clearly questionable from an efficiency point of view. A small negative effect is observed on the weekly number of hours worked. The results suggest that the use of early vacancy information promises a

"double dividend" with respect to the activation cost. First, the early activation reduces the duration of unemployment, and thus the necessity of subsequent measures participation. Second, the focus on early activation by vacancy information may result in lower costs of "locking-in" than other measures of early activation.

In Chapter 3, the long-term effects of participation in ALMP are assessed for unemployed youth under 25 years of age. Complementary to the results in Chapter 2, the effects of participation in time- and cost-intensive measures of active labor market policies are examined. Youth unemployment is seen especially detrimental due to long-term "scarring effects" on future labor market prospects (Ellwood, 1983; Burgess et al., 2003; Gregg and Tominey, 2005). At the time of this study, no comprehensive quantitative analysis of the effectiveness of ALMP for young unemployed in Germany existed, despite the large number of young participants in ALMP. We study the effects of job creation schemes, wage subsidies, short-and long-term training measures and measures to promote the participation in vocational training. The outcome variables of interest are the probability to be in regular employment, and participation in further education during the 60 months following program entry. Our analysis shows that all programs, except job creation schemes have positive and long-term effects on the employment probability of youth. In the short-run only short-term training measures generate positive effects, as long-term training programs and wage subsidies exhibit significant "locking-in" effects. Measures to promote vocational training are found to significantly increase the probability of attending education and training, whereas all other programs have either no or a negative effect on training participation. Effect heterogeneity with respect to the pre-treatment level education shows that young people with higher pre-treatment educational levels benefit more from participation most programs. However, for longer-term wage subsidies we also find strong positive effects for young people with low initial education levels. The relative benefit of training measures is higher in West than in East Germany.

In the evaluation studies of Chapters 2 and 3 semi-parametric balancing methods of *Propensity Score Matching* (PSM) and *Inverse Probability Weighting* (IPW) are used to eliminate the effects of counfounding factors that influence both the treatment participation as well as the outcome variable of interest, and to establish a causal relation between program participation and outcome differences. While PSM and IPW are intuitive and methodologically attractive as they do not require parametric assumptions, the practical implementation may become

quite challenging due to their sensitivity to various data features. Given the importance of these methods in the evaluation literature, and the vast number of recent methodological contributions in this field, Chapter 4 aims to reduce the knowledge gap between the methodological and applied literature by summarizing new findings of the empirical and statistical literature and practical guidelines for future applied research. In contrast to previous publications (e.g., Caliendo and Kopeinig, 2008), this study does not only focus on the estimation of causal effects, but stresses that the balancing challenge can and should be discussed independent of question of causal identification of treatment effects on most empirical applications. Following a brief outline of the practical implementation steps required for PSM and IPW, these steps are presented in detail chronologically, outlining practical advice for each step. Subsequently, the topics of effect estimation, inference, sensitivity analysis and the combination with parametric estimation methods are discussed. Finally, new extensions of the methodology and avenues for future research are presented.

Chapter 1

Social Networks, Job Search Methods and Reservation Wages: Evidence for Germany*

1.1 Introduction

Social networks are an important source of information in the labor market, and many workers find jobs through friends and relatives. Seminal studies by Rees (1966) and Granovetter (1995) show that a considerable part of the working population relies on personal contacts to obtain information about job offers. According to a recent study by Franzen and Hangartner (2006), around 44% of the workers in the US and 34% of the workers in Germany found their jobs through social networks.[1] The widespread use of informal search channels has given rise to an extensive body of literature investigating the effect of networks and informal search on labor market outcomes.

One reasonable assumption is that informal job contacts reduce informational asymmetry by lowering uncertainty about the job match quality for both employees and the employers (see, e.g., Montgomery, 1991). In terms of labor market outcomes, this mechanism should lead to higher wages and longer job tenure. However, the empirical evidence is rather mixed. In particular, it has been found

*This chapter is based on the paper *Social Networks, Job Search Methods and Reservation Wages: Evidence for Germany* joint with Marco Caliendo, and Arne Uhlendorff (Caliendo et al., 2011). The research project was partly funded by the German Research Foundation (DFG).

[1]These numbers are based on the International Social Survey Program 2001.

that informal search success can be associated with a premium as well as a penalty in terms of wages and employment stability (compare, e.g., Ioannides and Loury, 2004 and Mouw, 2003 for extensive overviews). More recent studies focus on the quality of the information transmitted via the network. It is argued that the network's productivity is determined by the characteristics of individuals comprising the network, and it is expected that the employment status of individuals within a network are correlated with each other (compare Calvo-Armengol and Jackson, 2007)[2].

A related strand of literature analyzes job search outcomes by explicitly modeling the job search process. As individuals tend to use several sources of information during job search, particular attention is paid to the choice of search channels and its impact on labor market outcomes (see e.g. Holzer, 1988, van den Berg and van der Klaauw, 2006, and Weber and Mahringer, 2008). Based on theoretical job search models with differential search channels, these studies derive implications from changes in productivity or costs of search on the search channel choice, search intensity and corresponding labor market outcomes.

In this paper we link directly observable information on social networks to the job search behavior of the unemployed. In contrast to previous studies focusing on the effect of informal search on realized search outcomes, we explicitly study the effect of the extent of networks on the choices individuals make in the job search process. This approach allows us to shed some light into the "black box" of the interplay between social networks and job search choices of individuals, which has to date received little attention in the literature. If the assumption that networks convey relevant job-information holds, it is likely that well-connected individuals receive more job offers through their network than individuals with fewer social contacts. In turn, if networks do play a role in the job search process, it is expected that individuals adjust their search behavior contingent on the network they possess. For this purpose, we distinguish between two different search channels: formal and informal search. Formal search is defined as search by newspaper advertisements, internet, public employment office, etc., while informal search refers to search via friends and relatives. We discuss potential effects of network size

[2]However, the corresponding data requirements in terms of the quality of the individual network are high and usually not met in conventional survey data. Therefore, some studies approximate the network quality, e.g., with the characteristics of the neighborhood of the individuals (see, e.g., Topa, 2001, and Bayer et al., 2008). See Cappellari and Tatsiramos (2010) for a recent empirical analysis with directly observed network quality.

on job search efforts and reservation wages within a theoretical framework that is closely related to the studies by van den Berg and van der Klaauw (2006), Holzer (1988), and Weber and Mahringer (2008).

Our empirical analysis is based on the *IZA Evaluation Dataset S* (see Caliendo et al., 2011, for details). This unique data set consists of around 17,000 individuals who had become unemployed between late 2007 and early 2008. The data provide detailed information on social networks and allow us to observe the job search process of unemployed in detail, i.e., the types of search channels they use, their intensity of search, as well as their reservation wage. We link these search variables to social network indicators, measured by the number of close friends and the contact frequency to former colleagues.

The set-up of our data has several advantages which allow a direct analysis of the relation between networks and job search choices. First of all, the interviews were conducted around seven weeks after entering unemployment. The fact that each individual is interviewed at a very early point in time during the unemployment spell reduces the problem of potential reverse causality, which is a typical concern of studies on the relationship between concepts such as social networks or non-cognitive skills and labor market outcomes.[3]

Another concern is that the size of the network might be correlated with unobserved heterogeneity which simultaneously has an impact on job search behavior. In order to control for this potential omitted variable bias, we exploit a rather informative set of observable characteristics, including personality traits, previous labor market outcomes and other socio-demographic characteristics. Given this unusually rich set of individual information, exploring the relationship between networks and job search behavior conditional on observable characteristics seems to be a reasonable strategy. As mentioned above, recent research has stressed the importance of observing the quality of the network in explaining the heterogenous impact of social networks on labor market outcomes. Since we do not observe the quality of the network, i.e., we do not have any information on the labor market characteristics of friends or colleagues, we cannot deduce whether the network of friends is likely to convey helpful information for the unemployed. However, if individuals decide to use informal search channels, the assumption that larger net-

[3]Alternatively, one could model the interdependencies between network formation and employment dynamics explicitly (see, e.g., Bramoullé and Saint-Paul, 2010). For this approach, panel data is required in order to explore individual variation over time.

works convey more information is still likely to hold, independent of the network quality. For the interpretation of the results, it is important to keep in mind that the measured impact captures only one dimension—the size of a network.

Our results show that search behavior is indeed influenced by the presence of social contacts. In particular, we find evidence that individuals with larger networks substitute informal search at the cost of formal search effort. The effect is particularly pronounced for passive formal search methods, i.e., information sources that generate rather unspecific types of job offers at low relative costs. In line with the predictions of the theoretical model, we also find significantly positive effects of an increase in the network size on reservation wages. Our results further show the importance of including personality traits, e.g., openness and extraversion, in the analysis of social networks. Once we control for personality traits, the impact of our network indicators on the use of formal search model becomes stronger, while the impact on reservation wages weakens.

The outline of this chapter is as follows. Section 1.2 summarizes some related literature on job search choices of individuals. Section 1.3 presents the theoretical framework from which we derive our testable implications. Section 1.4 describes the *IZA Evaluation Dataset S* in more detail, specifies the sampling strategy for the estimation sample and motivates the choices of the network information used. Section 1.5 outlines the estimation strategy and presents the results. Section 1.6 concludes.

1.2 Previous Empirical Findings

Many studies have shown that unemployed workers use multiple channels of job search and that the majority of unemployed workers makes use of informal channels (compare for evidence from different European countries, Pellizzari, 2010). In the standard partial job search model with endogenous search effort, unemployed individuals use one general search channel and choose an optimal search effort s^* and a reservation wage ϕ in order to maximize their utility (see, e.g., Mortensen, 1986). The reservation wage defines the "stopping rule" and corresponds to the wage offer for which the present value of continued search equals the present value of accepting the wage offer, i.e., every wage offer above ϕ will be accepted. In the analysis of job search with multiple search channels, it is assumed that the choice of a particular search channel and the channel-specific search effort is determined

by the relative efficiency of that channel in generating acceptable job offers.[4]

An early example for a study on the determinants of the choice of search methods and its effectiveness is Holzer (1988). Using a sample of unemployed youths—who are interviewed at different points in time during their unemployment spell—he finds that the main determinants of search channel use are the relative costs in terms of time spent on a particular channel for generating job offers and acceptances. Blau and Robins (1990) also analyze job search choices and outcomes, emphasizing the differences between search of unemployed and employed individuals. As in Holzer (1988), they find heterogeneous job offer arrival and acceptance rates for the different channels. However, as they do not observe the channel-specific search effort, they are not able to identify whether the differential success rates are explained by differential effectiveness of these search channels or by differential use. More recently, Weber and Mahringer (2008) conducted a similar analysis in their examination of the job search choices of recently employed workers in Austria. In line with the previous studies, they find that contacting friends is one of the most commonly used search methods that is also most effective in terms of successful job offers. Furthermore, they provide evidence that the success of a search channel is indeed highly heterogenous across individual characteristics such as education and labor market attachment. However, very few of these characteristics influence the success probability of informal search, which suggests that the widespread use of informal search is driven by its high relative efficiency. It has to be noted, however, that none of these studies investigates correlations between network indicators and the choice of job search methods. An example of a study which analyzes the impact of the social networks on job search channels and search outcomes is Wahba and Zenou (2005). They use population density as a proxy for the size of social networks and find—based on cross-sectional data for Egypt—that the probability of finding a job through friends and relatives increases and is concave in population density. Mouw (2003) explicitly considers the relationship between specific network characteristics and the use of informal search channels. However, he does not find any evidence for a positive relationship between the "quality" of a network, e.g., the proportion of friends in similar jobs, and the use of informal search channels.

[4]See Mortensen and Vishwanath (1995) for a theoretical equilibrium analysis on the effects of formal and informal search on labor market outcomes. In their model they provide a rationale for workers with a higher probability of obtaining job offers through employed contacts earning more in equilibrium.

A structural analysis on the differences between formal and informal search is conducted by Koning et al. (1997). In their analysis, they find no evidence for differences in the wage offer distributions between formal and informal search channels but discover an increased exit rate from unemployment for the use of informal channels, compared to formal channels. However, they do not find any significant effect of a social network indicator—reflecting the number of friends—on the exit rate from unemployment to employment via informal channels.

Using a field experiment of randomly assigned job search assistance and search monitoring, van den Berg and van der Klaauw (2006) show that unemployed workers shift from informal search effort to formal effort if their formal search level is monitored. They find evidence that these one-sided monitoring activities may lead to inefficient substitution effects, especially for well-qualified individuals. In summary, these studies indicate that the choices of specific search channels are indeed driven by cost-benefit considerations. Accordingly, if the hypothesis that social networks give access to additional information holds, individuals with higher levels of networks should experience greater productivity from their informal search channel and thus adjust their job search behavior accordingly. In the following we discuss the theoretical implications of an exogenous increase in the size of social networks on the individual choice of search channels, corresponding search effort and the reservation wage.

1.3 Theoretical Framework

Our framework is closely related to the theoretical model of job search with endogenous search effort and two search channels by van den Berg and van der Klaauw (2006). We focus on a sequential and stationary model of job search with two search channels, formal (f) and informal (n). An unemployed worker chooses optimal levels of formal search effort s_f and informal search effort s_n, the sum of both equals the overall search effort, $s = s_f + s_n$. Each search channel has a channel-specific job offer arrival rate $\lambda_i, i = f, n$, that is a function of the search effort devoted to it. We assume that the job offer arrival rate is strictly concave in search effort for both search channels. The productivity of informal search depends positively on the size of the network. The job offer arrival rate from informal search $\lambda_n(s_n, n)$ is given by $\lambda_{n0}(s_n)f(n)$. $f(n)$ increases in the magnitude of the network n, $\frac{\partial f(n)}{\partial n} > 0$, and is multiplied with the "baseline" arrival rate, which

depends positively on search effort s_n. Furthermore, there is a cost c of search, which increases with the search effort invested. We assume that $c = c(s_n, s_f)$ is convex in s_n and s_f. An assumption that is commonly made in the literature is that the cross-partial derivative of the cost-function is greater than zero, i.e., $\partial^2 c/(\partial(s_f)\partial(s_n)) > 0$.[5] This reflects that formal and informal search are similar activities, which implies that the marginal costs for informal search are higher, the more time is invested in formal search, and vice versa.

The timing of the model is as follows. In each period of length dt, the unemployed receives a job offer with probability $(\lambda_f + \lambda_n)dt$. Each offer is characterized by a wage w, randomly drawn from the wage offer distribution $F(w)$, which is the same for both search channels. If the unemployed receives an offer, he decides whether to accept it or continue searching. If he accepts the offer, his utility will be equal to the present value $V_e(w)$ of working at wage w. His present value of continued search, given his expectations of future job offers, is V_u, which is also dependent on the utility derived being unemployed, b, and the cost incurred by searching. In order to maximize utility, the unemployed continues searching until $V_e(w) = V_u$. It can be shown that the unemployed is indifferent between either choices if the wage offer w is equal to his reservation wage $\phi = \rho V_u$, where ρ denotes the rate of discount. Hence, in each period the worker maximizes his current and expected utility by choosing a reservation wage and an optimal amount of search effort in each search channel. The maximization problem is given by:

$$\max_{s_n, s_f} \phi = b - c(s_f, s_n) + (1/\rho)(\lambda_f(s_f) + \lambda_n(s_n, n))\left[\int_{\phi}^{\infty} (w - \phi)h(w)dw\right]. \quad (1.1)$$

It follows from the first order conditions that the optimal amount of effort invested in each search channel equates the expected marginal returns and the marginal costs of search in the respective channel.

Based on these optimality conditions, we are interested in the impact of an increase in network size n on the optimal levels of the reservation wage ϕ and the search efforts s_f and s_n. We assume that the network size is determined exogenously to the unemployment spell and that it enters the optimization problem only through a change in the job arrival rate of the informal search channel. As mentioned above, the set-up of this model is very similar to the one discussed in

[5] Note that the implications of this assumption are equivalent to the implications of the assumption that the cross-partial derivative of a joint production function is negative.

van den Berg and van der Klaauw (2006). In their theoretical model, counseling by caseworkers facilitates search along the formal channel. They are interested in the effect of a change in the amount of counseling on the job search behavior and derive—under several reasonable assumptions—testable implications which can be directly adopted in our model. In particular, they assume channel substitutability and show that an increase in the amount of counseling increases the reservation wage and the effort spent on formal search, while the unemployed reduce the effort for informal search (see van den Berg and van der Klaauw, 2006, for a detailed proof). This implies in our setting that we expect individuals with a larger network to have a higher reservation wage ($\partial\phi/\partial n > 0$), a positive impact of the network size on informal search ($\partial s_n/\partial n > 0$) and a negative impact on the effort spent on formal search ($\partial s_f/\partial n < 0$).

Intuitively, an increase in network size leads to an increase in the overall search productivity, which leads—for a given amount of search effort—to an increase in the value of search. The present value of unemployment increases, which implies an increase of the reservation wage. However, as the reservation wage increases, the marginal expected benefit of search will decrease. Hence, this indirect negative effect dampens the positive effect of a productivity increase on the reservation wage—although the overall change is expected to be positive (compare van den Berg and van der Klaauw, 2006). Faced with different values of continued search depending on the size of their network, individuals optimally allocate their search effort devoted to formal and informal search. In particular for the case of substitutable or independent search channels, an increase in informal search productivity leads—for a given amount of overall search effort—to a redistribution from the effort spent on formal search to the effort spent on informal search. In the case of substitutable channels, the substitution effect is reinforced by increasing marginal costs of search with the other respective channel. If the cost functions are independent of one another, i.e., if the cross-partial derivatives are zero, this reinforcing effect is absent, which weakens the substitution effect. In both cases, however, it is expected that informal search intensity increases and formal search intensity decreases.[6]

[6]Note that in the case of channel substitutability the decision to allocate a strictly positive amount of search effort to both channels might also depend on the productivity difference of the two channels. For example, if the formal channel is much more productive than the informal one, the marginal costs of engaging in search via the informal channel might be too high at the optimal level formal search effort.

Alternatively, one could think of search channels with complementary productivity (or costs). This would imply that an increase in the search intensity in one search channel leads to an increase in the marginal productivity of the other. In this case it is more difficult to draw unambiguous conclusions about the different effects (see Holzer, 1988, for theoretical implications of varying cross-dependencies between search channel productivities). Overall, one would not expect to observe a substitution of search intensities if the productivity increase in formal search is at least as high as for informal search.[7]

1.4 Data and Descriptive Analysis

1.4.1 The IZA Evaluation Dataset

We test the hypotheses of our model empirically, using observable characteristics of the individual network as an indicator for the efficiency of search through the informal search channel. The data we use are drawn from the *IZA Evaluation Dataset S*, which consists of an inflow sample into unemployment from June 2007 to May 2008. The data set is based on two components, an administrative part which contains extensive information on past labor market experience and a survey part. The key feature of the survey data is that individuals are interviewed shortly after becoming unemployed. They are asked general questions about their socio-demographic background, their employment history, as well as a variety of non-standard questions about attitudes, expectations and personality traits (see Caliendo et al., 2011, for details).[8] The data sampling is restricted to individuals who are 17 to 54 years old and who receive, or are eligible to receive, unemployment benefits under the German Social Code III (*SGB III*). Out of the gross sample of 9% of the monthly inflow into unemployment, a representative sample of approximately 1,450 individuals was interviewed each month between June 2007 and May 2008. Altogether, 17,396 interviews were conducted, with an average time

[7] In their paper van den Berg and van der Klaauw (2006) also argue that their results only hold as long as the productivity increase induced by counseling is larger for formal search than for informal search.

[8] For those individuals who gave us their permission, we are able to link the survey data with administrative records based on the "Integrated Labour Market Biographies" of the Institute for Employment Research (IAB), which contains relevant register data from four sources: employment history, benefit recipient history, participation in active labor market programs, and job seeker history.

lag between unemployment registration and the interview of nine weeks.

In the empirical analysis we estimate the effects of social networks on the search behavior of recently unemployed workers. Hence, we restrict the sample to individuals who are still unemployed when interviewed and who are actively searching for employment. Furthermore, we exclude individuals under the age of 25 and who report that they are looking for both an apprenticeship and employment. In order to obtain comparable individuals in terms of their network composition, we further exclude individuals who reported not having colleagues from some earlier employment relationship. From this preliminary sample of about 9,400 individuals, we further exclude the lowest and highest percentile of the reported hourly reservation wage and the search intensity as well as individuals with missing values for any key variables. This leaves us with a sample of 7,953 individuals.

Table 1.1 provides descriptive statistics of the estimation sample. The average unemployed person in our sample is 36 years old and the share of females is 50%. In addition, 68% of the unemployed live in West Germany and 5% are non-German citizens. Comparing these sample figures with official unemployment data in Germany, it can be seen that the sample selection process did not affect the representativeness of our sample (compare Bundesagentur für Arbeit, 2007). Regarding the education level, the majority of individuals have a medium level high school qualification[9] and 72% have completed at most some type of professional training[10]. Before entering unemployment, the majority of individuals was in regular employment (67%). Additionally, the data contain information on personality traits such as the "locus of control" which is defined as a generalized expectation about the internal versus the external control of reinforcement (Rotter, 1966). Individuals with an internal locus of control see future outcomes as being contingent on their own decisions and behavior, while individuals whose external locus of control dominates tend to attribute life's outcomes to external factors such as luck or fate. It is generally found that individuals with a more internal locus of control do better in terms of their labor market outcomes (see, e.g., Andrisani, 1977 and Osborne Groves, 2005). Further dimensions of the individuals personality traits included in the regression are measures capturing openness, conscientiousness, extraversion and neuroticism. A large array of literature has shown

[9]The lower secondary education system in Germany is divided into three parallel tracks (dubbed "low","medium" and "high"), providing prerequisites for the post-secondary vocational system in either work- or school-based vocational training or tertiary education, respectively.

[10]This corresponds to "post-secondary non-tertiary education" at ISCED level 4.

Table 1.1: Socio-demographic characteristics and personality traits: selected descriptives of the estimation sample

Variable	Shares[1]
West Germany	0.68
Female	0.51
Age (in years)	36.12
German citizenship	0.95
Married (or cohabiting)	0.41
School leaving degree	
None, special needs, other	0.02
Lower secondary school	0.30
Middle secondary school	0.42
Specialized upper secondary school	0.26
Vocational training	
None	0.08
Internal or external professional training, others	0.72
Technical college or university degree	0.19
Employment status before unemployment	
Employed	0.67
Subsidized employment	0.07
School, apprentice, military, etc.	0.14
Maternity leave	0.05
Other	0.08
Type of employment wanted:	
Fulltime employment	0.69
Part-time employment	0.15
Full or part-time employment	0.16
Unemployment benefit recipient (yes)	0.81
Internal Locus of control	0.54
Personality traits: I see myself as a person who	
... does a thorough job.	6.45
... does things efficiently.	6.07
... is talkative.	5.83
... is outgoing, sociable.	5.50
... is reserved.	3.86
... is original, comes up with new ideas.	5.22
... has an active imagination.	4.79
... worries often.	4.88
... gets nervous easily.	3.55
... is relaxed, handles stress well.	5.10
Number of observations	7,953

Source: IZA Evaluation Dataset S, own calculations.
[1] The numbers are shares unless otherwise indicated.

that non-cognitive skills and personality traits have predictive power in models on labor market outcomes (see Borghans et al., 2008, for an overview). Therefore, it will be important to control for them later in our empirical analysis.

1.4.2 Defining Social Networks

In our analysis we are interested in exogenously determined networks that individuals might employ in order to obtain relevant information in the labor market. In particular, it is required that the network size or strength is not affected by the current unemployment spell. In general, several endogeneity issues might arise that have to be considered when using network parameters in job search equations.

In the case of a dynamic endogenous selection process, the network of the unemployed is affected by the unemployment spell or duration. First of all, it may be the case that the network of relevant social contacts is diminished during the course of unemployment, as the change in circumstances leads to the dissolution of some social ties. This implies a potential problem of reverse causality, as the unemployment spell causes a change in network size. As argued above, we expect that the set-up of the data prevents this type of selection, as individuals are all interviewed at a similar point in time relative to their entry into unemployment. Since interviews were conducted shortly after the unemployment spell commences, we also expect any effects on network composition to be rather small. Another type of dynamic endogeneity is characterized by individuals strategically increasing their social network in order to increase the probability of receiving informal job information (compare, e.g., Galeotti and Merlino, 2014). In terms of our job search model, this would imply that the measure of informal search effort should capture the effort devoted to the enlargement of the social network as well. However, as this is also linked to the magnitude of the network that the individuals had before entering unemployment, it is difficult to disentangle the effects of the pre-existing and the "new" network on the job search process. In order to avoid this problem, we restrict our analysis to networks that had already been established before individuals entered unemployment and that were presumably not altered during the course of unemployment.

In the context of job search, the most relevant information on networks contained in our data are questions regarding friends and colleagues. Clearly, these two groups of contacts are not conclusive in depicting the social network of individuals as a whole and should be seen as an approximation. We focus on these two types of networks for two reasons: first, they are very likely to convey potentially relevant job information, which makes them relevant for our analysis. Second, we are able to extract information that is unlikely to be influenced by entry into unemployment, helping to avoid the endogeneity problems previously mentioned. In

particular, we approximate the network of friends by the number of "close" friends, as it is not probable that many close friendships were formed or ended in the short time interval between unemployment entry and the interview date. With respect to the information on colleagues, we use the contact frequency to colleagues before the individual entered unemployment. As this refers to characteristics of the network established before entry in unemployment, it is by definition unaltered during the unemployment spell. Around one third of our sample did not enter unemployment directly from employment (compare Table 1.2), so some individuals might refer to colleagues they had in some other previous employment. By including information on the previous labor market state, we control for potential differences in the relevance of the colleague network for these individuals.[11]

Table 1.2: Number of close friends and former contact frequency to colleagues

Variable	N	Shares[1]
Questions in survey		
Number of close friends outside family	7,953	4.83
Contact with colleagues before UE		
never	2,349	0.30
infrequent contact	1,988	0.25
occasional contact	2,135	0.27
frequent contact	1,481	0.19
Coding in the analysis		
Number of close friends outside family		
low (0-2)	2,169	0.27
medium (3-5)	3,991	0.50
high (more than 5)	1,793	0.23
Contact with colleagues before UE		
low	2,349	0.30
medium	4,123	0.52
high	1,481	0.19
Correlation coefficient between the coded indicators		0.07***
Number of observations	7,953	

Source: IZA Evaluation Dataset S, own calculations.
[1] The numbers are shares, unless otherwise indicated.

Table 1.2 shows that the individuals in our sample have around five close friends on average, whereas the frequency of contact to colleagues is more or less evenly distributed across the different categories, with slightly fewer observations in the group with the highest contact frequency. We aggregate the information to reflect

[11]The importance of the network of colleagues might differ with the way individuals exited their last job, e.g., workers subject to mass layoffs might not use this network at all. We have no indicator for this in the data; however, a sensitivity analysis does not indicate systematic differences between individuals on layoff and individuals who lost their job for other reasons.

the individual's degree of interaction with the respective social network and thus the potential access to valuable labor market information. We use a three-level scale for both measures, differentiating between low, medium and high levels of the respective network indicator.[12]

It should be noted that we do not make any a priori assumptions about the relative effectiveness of the two observed types of networks. In previous literature the effectiveness of networks in the job search process is found to vary with several characteristics of the network, i.e., quality (compare, e.g. Cappellari and Tatsiramos, 2010) or "strength of ties" (compare, e.g. Granovetter, 1995). As mentioned before, the data do not contain any direct measure of network quality, which is why we interpret both of our network measures only along the dimensions of quantity. Regarding the strength of ties, we can readily assume close friends to be "strong ties"; however, with respect to the network of colleagues, we do not have any indication whether previously high levels of interaction lead to the formation of "strong ties" or not, which would make a categorization attempt problematic.

1.4.3 Search Behavior

The outcome of interest in our analysis is individual job search behavior, represented by the reservation wage, the choice of informal search channels and the search intensity of formal search. The survey question regarding the use of particular search channels is designed as a multiple choice question, with individuals choosing one or more different channels that were used since becoming unemployed. Ten alternatives were offered, including informal search via relatives, friends and other contacts. Table 1.3 provides a detailed list of the options given. Contacting friends and acquaintances is one of the most commonly used methods when searching for employment, with 85% of individuals using it. Other, similarly important sources of information are job advertisements in newspapers and the internet. In order to measure search intensity devoted to formal search we use the number of formal search channels used—a method proposed by Holzer (1988). Table 1.3 shows that the unemployed use on average four formal search channels. For the analysis of substitution effects between formal and informal search channels, some

[12]We obtain the three-level scale by grouping together the middle values of a quartile-decomposition of the friend distribution and the middle-values of the four-level scale of contact frequency, respectively. Regression with a linear and log-linear transformations of the number of friends show that the results are not sensitive to the definition of the network categories.

Table 1.3: Job search behavior of the unemployed

Variable	Shares[1]
Hourly reservation wage (in euros)	7.03
median	[6.60]
s.d.	(2.29)
Use of informal search channel	0.85
Use of formal search channels:	
advertisements in a newspaper	0.84
own advertisement	0.14
using the job information system (SIS)	0.60
contacting the unemployment agency	0.70
research on the internet	0.86
contacting a private agent with agency voucher	0.09
contacting a private agent without agency voucher	0.16
direct application at companies	0.67
others	0.19
Number of formal search channels used	4.25
median	[4.00]
s.d.	(1.56)
Number of active[2] formal search channels used	0.97
median	[1.00]
s.d.	(0.74)
Number of passive[2] formal search channels used	3.29
median	[3.00]
s.d.	(1.19)
Number of observations	7,953

Source: IZA Evaluation Dataset S, own calculations.
[1] The numbers are shares, unless indicated otherwise.
[2] Own advertisement, contacting a private agent without voucher and direct application at companies are considered active search. The remaining formal search channels are considered passive search.

sources of information may be considered more suitable substitutes to informal channels than others. In order to identify this, we make the additional distinction between active and passive formal search channels, where active search methods are those that individuals use if they want to solicit specific, pre-defined types of jobs, rather than react to job opportunities that appear at random. A similar distinction is made by Kahn and Low (1988), who differentiate between systematic and random search behavior. We allocate posting advertisements in newspapers, direct applications at companies, as well as using private agents without agency vouchers, to active search measures. All other formal channels are defined as passive search. Besides the fact that this distinction groups channels that generate a similar specificity of job offers, the grouping is also valid in terms of search costs associated with the two groups. While passive search channels are rather inexpensive, active search channels generally require higher investment, both in time and money. It can be seen from the descriptives in Table 1.3 that the average

individual uses three passive measures but only one active source of information.

Table 1.4 depicts the unconditional variation in job search behavior for the different categories of friends and former colleagues. Without controlling for any

Table 1.4: Job search behavior by network indicator

Outcome	Frequency			p-values of t-test		
	low	medium	high	l-m	l-h	m-h
By number of close friends						
Hourly reservation wage (in euros)	6.87	7.11	7.03	0.00	0.03	0.21
Informal search	0.81	0.86	0.87	0.00	0.00	0.43
Number of formal search channels used	4.23	4.28	4.22	0.28	0.78	0.18
Number of active formal search channels used	0.95	0.95	1.01	0.85	0.01	0.01
Number of passive formal search channels used	3.28	3.32	3.21	0.19	0.06	0.00
Number of observations	2,169	3,991	1,793			
By former contact frequency to colleagues						
Hourly reservation wage (in euros)	6.68	7.19	7.13	0.00	0.00	0.43
Informal search	0.82	0.85	0.87	0.00	0.00	0.07
Number of formal search channels used	4.19	4.30	4.22	0.01	0.63	0.09
Number of active formal search channels used	0.93	0.97	1.00	0.07	0.01	0.15
Number of passive formal search channels used	3.26	3.33	3.22	0.02	0.31	0.00
Number of observations	2,349	4,123	1,481			

Source: IZA Evaluation Dataset S, own calculations.
Note: The numbers are shares, unless otherwise indicated. The p-value refers to a two-sided t-test of mean equality between the groups.

personal characteristics, the use of informal search channels increases unambiguously with the extent of the network indicator. However, using informal search channels is also an attractive possibility for individuals with a small number of friends and a low contact frequency to colleagues. The most significant differences in usage seem to exist between low and medium levels of friends and colleagues, whereas an additional increase in network size from medium to high does not seem to be correlated with changes in job search behavior. For the other variables, the relations do not increase with network strength. The reservation wage is highest for medium levels of network indicators, and the same holds true for the search intensity invested in formal search channels. Hence, a descriptive assessment of the relationship between networks and job search behavior seems to confirm that differences do exist. The magnitude and direction of these differences need to be tested in the empirical analysis controlling for individuals' characteristics.

1.5 Estimation and Results

1.5.1 Empirical Strategy

In order to assess the impact of social networks on the job search process, we integrate the network information in a parametric regression model of the type:

$$Y_i = X_i'\alpha + \sum_{j=l,m,h} (N_{1ji}'\delta_{1j} + N_{2ji}'\delta_{2j}) + Z_i'\mu + \varepsilon_i, \tag{1.2}$$

where Y_i denotes the individual parameters of job search behavior, measured by reservation wage, use of informal sources of information and the number of formal search channels used (in total and differentiated by active and passive search). Matrix X_i includes relevant socio-demographic characteristics, extensive information of past labor market experience, together with further determinants of job search choices. N_{1ij} and N_{2ij} are dummy variables, representing the strength of the individual's network, with $j = l, m, h$ representing low, medium or high levels of the network indicators, respectively. The network types considered here are the number of friends N_{1i} as well as former contact frequency to colleagues from previous spells of employment N_{2i} (see our discussion in Section 1.4). In addition, we include a set Z_i of personality traits. By controlling for the individual's personality, e.g., the locus of control, the degree of extraversion, neuroticism, etc., we are able to remove potential bias in δ_1 and δ_2 arising from omitted personality traits that simultaneously affect job search behavior and network formation. In particular, if we assume that these factors affect labor market success and network formation in the same way, neglecting them leads to an upward bias in δ_1 and δ_2 and thus an overestimation of the effects of networks in the regressions for the reservation wage. Furthermore, if individuals with a higher locus of control tend to search more intensely while possessing a larger network of friends, as suggested by Caliendo et al. (2014), we would obtain upwardly biased coefficients in the regression of formal search. If our model correctly predicts a reduction in formal search, however, omission of Z_i would lead to an underestimation of the true effect of networks. In order to assess the magnitude and sign of the potential bias, we conduct the regression with and without the Z_i and compare the results.

1.5.2 Results

Table 1.5 depicts the marginal effects of the logistic and least squares regression analyses, simultaneously incorporating the low-medium-high scaled network indicators of friends and colleagues.[13] The upper part of the table displays the results of a regression model omitting information on personality traits. Before turning to the analysis of the model with personality traits we begin by examining the general findings obtained by the former model. Column (1) reports the effects of an increase in the number of friends or the frequency of contact to former colleagues on the use of informal search channels. We find that the magnitude of the effect of a medium or high level of the network are very similar for both network measures used, increasing the probability of using informal search channels by around 5% on average, compared to individuals who have a low number of friends. These findings confirm the relevance of the network indicators used in the analysis and show that there is a significant positive relationship between the extent of the network, and hence its productivity, and the use of informal search channels. The comparably low magnitude of the effect is to be interpreted in the context of the relatively little variation in the use of informal search channels (compare Table 1.4), as most individuals with low levels of networks also consult their contacts for job information. Moreover, as this only considers the extensive margin of informal search, it provides rather limited insight into whether individuals increase the intensity of informal search when they have a greater number of friends or have more frequent contact to former colleagues. Based on our theoretical predictions in Section 1.2, we are able to deduce further insight by examining the reservation wage and the intensity of search devoted to formal channels.

From our theoretical model we deduce that the informal network only increases the productivity of the informal channel. As a consequence, we expect that individuals who decide not to use informal search channels do not reduce their formal search effort when their network increases. In order to explore the more common case in which individuals use formal and informal channels simultaneously, we exclude individuals in our sample who report not using informal sources of information. This reduces the sample by 14%, which leaves us with a sample of around 6,750 individuals. Further reference to this issue will be made in the sensitivity analysis in the next section.

[13] We include all network indicators simultaneously—separate analyses do not change the results significantly.

Table 1.5: Effect of friends and colleagues on the use of informal search channels and other job search behavior, using only individuals who use both, formal *and* informal channels.

	Informal search	Reservation wage	Formal search channels	Active formal search	Passive formal search
	(1)	(2)	(3)	(4)	(5)
No personality traits					
Number of close friends outside family (ref. low)					
medium	0.042***	0.017**	-.049	-.027	-.022
	(0.009)	(0.007)	(0.045)	(0.022)	(0.034)
high	0.051***	0.009	-.018	0.037	-.055
	(0.009)	(0.008)	(0.055)	(0.026)	(0.042)
Contact frequency with colleagues before UE (ref: low)					
medium	0.032***	0.017**	0.015	0.015	-.0006
	(0.01)	(0.008)	(0.05)	(0.025)	(0.038)
high	0.049***	0.018*	-.046	0.041	-.087*
	(0.011)	(0.009)	(0.062)	(0.03)	(0.048)
Adjusted/Pseudo R^2	0.037	0.413	0.061	0.028	0.065
Including personality traits					
Number of good friends outside family (ref. low)					
medium	0.039***	0.012*	-.070	-.036*	-.034
	(0.009)	(0.007)	(0.045)	(0.022)	(0.034)
high	0.044***	0.002	-.074	0.014	-.087**
	(0.01)	(0.008)	(0.055)	(0.026)	(0.042)
Contact frequency with colleagues before UE (ref: low)					
medium	0.03***	0.013*	-.005	0.005	-.010
	(0.01)	(0.008)	(0.05)	(0.025)	(0.038)
high	0.044***	0.012	-.092	0.019	-.111**
	(0.011)	(0.009)	(0.062)	(0.03)	(0.048)
Adjusted/Pseudo R^2	0.043	0.417	0.077	0.041	0.074
Number of observations	7,953	6,748	6,748	6,748	6,748
Unconditional mean			4.378	1.003	3.375

Standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level. All effects are marginal effects. The coefficients of informal search channel use are estimated using a logit; for the other variables we conducted LS regressions. Poisson regression results for number of search channels used yielded very similar results and are available from the authors upon request. Additional control variables used in the estimation: local UE rate, standard socio-demographic characteristics, UB recipient, months in unemployment, available communication, employment status before UE, time of entry into UE. Furthermore, the bottom regressions include measure for internal Locus of Control and personality traits.

Theory further predicts that in the case of substitutable or independent search channels, a sufficiently productive network will lead to a substitution from formal to informal channels, thereby increasing the reservation wage. Columns (3) to (5) in Table 1.5 refer to the network effects on the intensive margin of formal search channels, measured by the total number of search channels used. When

considering the total sum of formal search channels used in column (3), the negative coefficients indicate that an increase in network measures does indeed lead to the predicted substitution effect. The effects are strongest for a medium number of friends and a high contact frequency to former colleagues, resulting in a reduction of formal search by around 1% ($= \delta_{1j}/\bar{Y}$). The effects on the aggregate formal channel use are not statistically significant. Splitting up formal search into active and passive formal search, however, yields an improvement in statistical and economic significance, but only for the case of passive search channels. For active search intensity, the regression coefficients are predominantly positive, except for a medium number of friends, but not statistically significant. For the case of passive search channels, a high number of friends leads to a reduction of passive formal search effort by 2%, and for high frequency of contact to colleagues of 2.6%. This suggests that informal search is perceived as a substitute for formal search channels that generate rather unspecific types of job offers at a low cost. Our theoretical model further predicts that, given the productivity increase in informal search, an increase in networks should lead, ceteris paribus, to an increase in reservation wages. Column (2) shows the effect of the networks on reservation wages. Indeed, we find a small but significant increase in reservation wages for medium levels of friends and medium and high contact frequency with colleagues of around 1.7%.

Continuing the line of thought from Section 1.5.1, we now proceed to assess the bias arising from omitting personality traits in the above analysis. The lower part of Table 1.5 contains the same regression as above, with the inclusion of the personality traits. Comparison of the respective coefficients in column (1) shows that the inclusion of personality traits does not significantly affect the results on the use of informal search channels for either network indicator. However, when comparing the results in column (2), we find a decrease in the effects observed for the reservation wage, from 1.7% to 1.3% for medium levels of both network indicators. The significance of the effects is reduced. This confirms our hypothesis that certain personality traits affect networks and labor market outcomes in the same manner, e.g., outgoing individuals are simultaneously more likely to have more friends and be more successful in their career, resulting in higher reservation wages. For formal search channels in column (3), we observe the opposite. The effect becomes more pronounced and significant once personality traits are included. In particular, we observe a 1.7% reduction in the aggregate number of formal search channels for high numbers of close friends and a 2% decrease for

a high frequency of contact to colleagues. Once again, we find that the negative coefficients are mainly driven by the passive formal search intensity. In column (5) we observe a reduction in passive formal search of 2.6% and 3.3% for high levels of the respective network indicators. For active formal search, we only find a significant reduction in the search intensity for a medium level of friends. The high responsiveness of our results on the inclusion of personality traits underscores the problem of unobserved heterogeneity in the context of analyzing social network effects without any source of exogenous variation. Although the overall model fit does not increase significantly once these variables are included, the marked change of the results in the expected directions indicates that personality traits have an important effect on job search choices as well as on the individual's network.

In summary, our model predictions are largely confirmed by the empirical analysis. First, we are able to confirm the relevance of our chosen network indicators, due to their high significance in predicting the probability of using informal search channels. As our theoretical framework derives predictions for search intensity on the intensive margin, we proceed by examining the effect these network indicators have on formal search channels. The finding that unemployed significantly reduce their passive formal search effort as the network increases is in line with the notion that passive formal and informal channels are considered substitutes in terms of their productivity in generating job offers. Similar findings on search channel substitution have already appeared in previous literature, e.g., van den Berg and van der Klaauw (2006). Our analysis takes the additional step of establishing an explicit link between passive formal search and the size of an individual's network. Regarding active formal search, we find very little evidence for a substitution, which indicates that informal search and active formal search are rather seen as information complements. Within the framework of Kahn and Low (1988), who distinguish between systematic (active) and random (passive) search, depending on the previous level of labor market information, our findings indicate that the social network is used as source of random information that can be used to extend the knowledge about labor market opportunities, rather than produce specific types of offers. In terms of the reservation wage, theory predicts that in the case of search channel substitutability, reservation wages will increase. We are also able to observe this in the data, although the size of the effect is rather small. This is to be expected, however, as it is generally found that informal search channels lead to an increase in the exit rate out of unemployment, which would

be counterintuitive if the effect of networks on reservation wages were too strong.

1.5.3 Sensitivity Analysis

We test the sensitivity of our results with respect to systematic variation of the estimation sample. In order to assess the robustness of our model to the assumption that the use of informal search channels is the only way in which networks influence the job search process, we expand the sample to individuals who do not use informal search channels. Table 1.6 reports the corresponding regression results after including individuals who do not use informal channels. Overall, the results are quite stable. The point estimates for the impact of networks on the use of formal search channels are smaller, but have the same sign. Similar to the previous results, we do not find a significant impact on the use of active search channels, but do find a negative impact of high levels of our network indicators on the use of passive formal search channels. Furthermore, we find that the impact on the reservation wage is the same, indicating that networks might affect the job offer arrival rate of individuals who do not actively search via their network. Another point of interest centers around the question of whether the importance of network information varies with the type of job searched. In Table 1.7 we distinguish between individuals searching for both full-time and part-time employment, or part-time employment only. As the expected income stream resulting from the type of job searched is most likely to differ between the two groups, individuals might differ with respect to their search behavior. In particular, as the expected return from part-time work is lower than from full-time work, it could be that individuals search less intensely, as e.g. found by Weber and Mahringer (2008), and are hence more likely to rely on less costly search methods. We find that the results obtained in the main analysis only persist for individuals searching for full-time work or both types of jobs. However, we find no significant effect of networks on the search choices of individuals looking for part-time work only. These results seem to indicate that the reasoning of our model does not apply to this subgroup of unemployed workers. As this group is predominantly female, we further investigate the sensitivity of our previous results to gender effects. Stratification of the regression analysis, however, does not provide evidence of differential effects on the job search behavior of men and women. Thus, the question why part-time workers differ in their behavior requires further investigation.

As a further sensitivity test, we assess the dependence of our results on

Table 1.6: Effect of friends and colleagues on the use of informal search channels and other job search choices, including individuals who do not use informal search channels.

	Informal search	Reservation wage	Formal search channels	Active formal search	Passive formal search
	(1)	(2)	(3)	(4)	(5)
Number of close friends outside family (ref. low)					
medium	0.039***	0.015**	-.004	-.011	0.007
	(0.009)	(0.006)	(0.041)	(0.02)	(0.032)
high	0.044***	0.004	-.037	0.024	-.061
	(0.01)	(0.008)	(0.051)	(0.024)	(0.039)
Contact frequency with colleagues before UE (ref: low)					
medium	0.03***	0.013*	0.011	0.011	-.0006
	(0.01)	(0.007)	(0.047)	(0.023)	(0.036)
high	0.044***	0.011	-.070	0.022	-.091**
	(0.011)	(0.009)	(0.058)	(0.028)	(0.045)
Adjusted/Pseudo R^2	0.043	0.42	0.075	0.04	0.074
Observations	7,953	7,953	7,953	7,953	7,953
Unconditional Mean			4.253	.9652	3.288

Standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level. All effects are marginal effects. The coefficients of informal search channel use are estimated using a logit, for the other variables we conducted LS regressions. Poisson regression results for number of search channels used yielded very similar results and are available from the authors upon request. Additional control variables used in the estimation: local UE rate, standard socio-demographic characteristics, UB recipient, months in unemployment, available communication, employment status before UE, time of entry into UE, a measure for internal Locus of Control and personality traits.

the estimation method used. In the previous analysis the regression coefficients on search intensity were obtained by least squares regression analysis. However, given the non-normal distribution of the number of search channels used, this might not have been appropriate. Therefore, we also estimate poisson regressions, which are a better fit to the properties of count data; but find very similar results.

1.6 Conclusions

In this chapter we analyze the influence of social networks on job search behavior of unemployed individuals. Using the extensive survey data on recently unemployed workers in Germany collected in the *IZA Evaluation Dataset S*, we test hypotheses derived from a theoretical model of job search with two distinct search channels and endogenous search effort. In contrast to many previous studies, the data allow us to analyze directly the relationship between social contacts and the job search behavior of unemployed individuals. Our findings underscore the established importance of networks in the job search process. In particular, we find

Table 1.7: Effect of friends and colleagues on the use of informal search channels and other job search choices, stratified by type of employment searched.

	Informal search	Reservation wage	Formal search channels	Active formal search	Passive formal search
	(1)	(2)	(3)	(4)	(5)
Individual looking for fulltime or part-time employment					
Number of close friends outside family (ref. low)					
medium	0.04***	0.013*	-.065	-.031	-.034
	(0.01)	(0.007)	(0.049)	(0.024)	(0.038)
high	0.041***	0.0002	-.084	0.016	-.101**
	(0.01)	(0.009)	(0.06)	(0.029)	(0.046)
Contact frequency with colleagues before UE (ref: low)					
medium	0.023**	0.017**	-.012	0.004	-.016
	(0.011)	(0.008)	(0.055)	(0.028)	(0.042)
high	0.04***	0.016	-.122*	0.014	-.136***
	(0.012)	(0.01)	(0.067)	(0.033)	(0.051)
Adjusted/Pseudo R^2	0.045	0.443	0.012	0.01	0.01
Observations	6,768	5,746	5,746	5,746	5,746
Unconditional Mean			4.418	1.034	3.384
Individual looking for part-time employment only					
Number of close friends outside family (ref. low)					
medium	0.029	-.003	-.050	-.051	0.002
	(0.022)	(0.021)	(0.106)	(0.054)	(0.082)
high	0.058**	0.005	0.041	0.02	0.021
	(0.023)	(0.026)	(0.139)	(0.068)	(0.106)
Contact frequency with colleagues before UE (ref: low)					
medium	0.053**	-.0007	0.011	0.013	-.002
	(0.025)	(0.026)	(0.123)	(0.063)	(0.096)
high	0.055**	-.004	0.128	0.08	0.048
	(0.025)	(0.031)	(0.153)	(0.078)	(0.123)
Adjusted/Pseudo R^2	0.093	0.306	0.112	0.013	0.136
Observations	1,185	1,002	1,002	1,002	1,002
Unconditional Mean			4.149	.824	3.244

Standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level. All effects are marginal effects. The coefficients of informal search channel use are estimated using a logit, for the other variables we conducted LS regressions. Poisson regression results for number of search channels used yielded very similar results and are available from the authors upon request. Additional control variables used in the estimation: local UE rate, standard socio-demographic characteristics, UB recipient, months in unemployment, available communication, employment status before UE, time of entry into UE, a measure for internal Locus of Control and personality traits.

that individuals with larger networks substitute informal sources of information for formal ones. We also find that the substitution effect is strongest for formal search channels that are considered to generate job offers with rather unspecific job characteristics at lower costs. Moreover, we find evidence that larger networks lead to a statistically significant increase in reservation wages of around 1%. Hence, our analysis confirms the wide-spread belief that social contacts constitute rele-

vant sources of information in the job search process. These results advance our understanding of the role an individual's network plays in the job search process.

However, our analysis relies on several assumptions that require further testing. For instance, we assume that the network indicators used are unchanged by the incidence of unemployment. Given that further data points will be available in the data set in the future, we will be able to test the stability of networks with respect to labor market changes. In addition, a drawback of our data set is that we do not have qualitative information about the networks, e.g., the share of friends with a job and the type of occupations they have. Exploring such information directly or approximating the quality of the network indirectly with the help of neighborhood characteristics would shed more light on the relationship between job search behavior and social networks.

Further research is required to validate our findings in an analysis of subsequent labor market success of the unemployed. For example, we expect that individuals who experience increased productivity of search also leave unemployment earlier than otherwise similar unemployed without relevant contacts. Furthermore, we should find that the observed increase in reservation wages results in higher wages, irrespective of the successful search channel. Examining outcomes should also provide insight into the (relative) efficiency of the different types of job search channels. Since this efficiency might differ across specific types of individuals and jobs, it is also important to investigate differential effects, e.g., with respect to the skill level and previous wages of the unemployed.

Chapter 2

Competing Policies? The Effectiveness of Early Vacancy Information and its Effects on ALMP Use

2.1 Introduction

Labor market policies for the unemployed usually comprise a diverse set of activation measures; the optimal type and intensity of activation depends on the characteristics of the unemployed and the elapsed unemployment duration. During the initial unemployment spell, it is common to focus on 'job broking' activities: by transmitting information about open vacancies to the unemployed, the matching between firms and job seekers is to be facilitated. If unemployment continues, more extensive active labor market programs (ALMP) may be used, including job search coaching, training schemes, public sector job creation, and hiring incentives for firms. The gradual increase in activation intensity over time takes account of the higher costliness of more intensive measures: in European countries, public spending on placement services is with 0.07% of GDP about seven times lower than the spending on active labor market schemes.[1] Taking further into account

[1]The data are drawn from the OECD Statistics database and pertain to average spending levels between 2004 and 2011. Only countries with valid and comparable reporting for placement services and ALMP measures are included, i.e., Austria, Belgium, Czech Republic, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, and Sweden.

that the returns to search may be decreasing over time and that participation in activation programs tends to stifle own search effort, delaying high intensity interventions to later stages in unemployment may also be more efficient (Spinnewijn, 2013; Wunsch, 2013).

As a consequence of this progressive activation, the provision of high quality vacancy information during the early unemployment spell may not only improve individual labor market outcomes, but also the efficiency of the overall activation process. By lowering the risk of longer-term unemployment, the use of costly and time-intensive activation schemes later may be reduced. Also, in practice, caseworkers may be required to offer immediate activation rather than to choose the optimal timing for a program, so that high intensity schemes may be used as substitutes (rather than complements) to compensate for lacking or ineffective vacancy information. Here, the availability of high quality vacancies during early unemployment may also increase the efficiency of the early activation process, as caseworkers are less likely to resort to alternative, more intensive measures. In contrast, in the presence of monitoring schemes, the availability of bad quality information may also result in an increased use of ALMP on the long run, as unemployed are more likely to enter instable jobs by being incentivized to apply to and accept low quality employment relationships.

In this context, our paper assesses the effectiveness of receiving vacancy information early in the unemployment spell in Germany, taking into account potential interactions between the provision of vacancy information and the use of alternative activation schemes. In line with the previous literature on the labor market effects of vacancy information, we examine the effects of receiving vacancy information on the exit rate from unemployment, and the quality of accepted employment relationships. Two previous studies for France and Germany confirm that the provision of vacancy information may increase the speed of employment entry for unemployed job seekers (Fougère et al., 2009; van den Berg et al., 2013). The German study suggests that in the presence of monitoring more instable jobs and jobs with lower wages are accepted. Neither study accounts for the simultaneous presence of alternative activation schemes. We hence extend the previous literature by further addressing the effects of early vacancy information on the probability to participate in ALMP.

Upon unemployment registration, unemployment entrants in Germany are assigned to a caseworker at the local public employment services (PES). Follow-

ing a mandatory labor market profiling, the caseworker decides on the activation approach, including, e.g., the provision of vacancies and participation in ALMP. While caseworkers have discretionary power to select the optimal strategy, they may face availability restrictions with respect to the provision of vacancy information. Qualitative studies show that caseworkers spend very little time on the acquisition of vacancies, but consider vacancy registrations as exogenously given (Boockmann et al., 2013). Between 2005 and 2008, the average vacancy coverage rate ranged between 30% to 50%, varying locally, by business cycle and by sector of work.[2] Whether or not unemployed receive vacancy information hence depends on the labor market characteristics of the unemployed, differences in caseworker activity, as well as the availability of vacancy information.

A causal analysis of vacancy effectiveness needs to take account of this non-random treatment assignment. In the absence of exogenous variation, informative data is required that allows to assume exogeneity of the receipt of vacancy information, conditional on observable confounders. A suitable data base for this purpose is given by the *IZA Evaluation Dataset S* (Caliendo et al., 2011), a representative survey on unemployment entrants between 2007 and 2008 in Germany. The data consist of a baseline interview shortly after unemployment entry, followed by two subsequent interviews capturing labor market outcomes up to three years later.

In particular, the data comprise a detailed assessment of the type of activation measures offered to the unemployed. From this we retrieve the information on the receipt of vacancy information as our treatment of interest. In our main analysis we focus on information on fulltime regular employment to have a more precise definition of the quality of vacancies. Unfortunately, our treatment indicator has the caveat that vacancies that were received after the first survey interview are not observed, which may bias our estimates toward zero as some non-receivers may have received vacancy information later in the unemployment spell. The data capture very detailed information that allows us to control for the endogenous. As unemployed job seekers are likely to use multiple search channels, we need to account for systematic differences in the overall search productivity that is the same across all channels. We hence control for socio-demographic characteristics, past labor market experience, personality traits and information on current job search behavior, as well as local labor market conditions. PES-specific heterogeneity is

[2]An overview from the OECD (2001) suggests that the vacancy to total hirings-rate was between 7% to 50% in OECD countries during the mid-1990's. To our knowledge, no later information is available.

accounted for by indicators on vacancy coverage rates, the frequency of sanctions, and the intensity of ALMP use. Furthermore, we control for the contact frequency with the caseworker, as well as complementary activation measures offered, to account for potential interaction between being offered vacancy information and more intensive types of activation.

We employ semi-parametric matching on the propensity score to estimate the treatment effects, and focus on a homogenous group of unemployment entrants that is subject to the same legal framework and the same activation process. In particular, we restrict the estimation to unemployment entrants that actively search for fulltime employment, and focus on those who receive, or are eligible to receive unemployment benefits. Within these restrictions, we conduct the analysis separately by gender and by region of work, i.e., East Germany vs. West Germany, respectively. As our main outcome of interest we investigate the effects on the speed of exit into employment during the 13 months following unemployment entry, differentiating between jobs found through the PES channel and jobs found through all other ("non-PES") channels. The quality of the first employment spell is measured by the hourly wage, weekly hours worked and the acceptance of 'unstable' jobs, i.e., short-term employment and employment in a temporary work agency. Finally, we assess the effects of early vacancy information on the probability to participate in training programs and job creation schemes.

Our results paint a diversified picture of the effectiveness of vacancy information in Germany. For all labor market groups we find that the transition rate into regular employment is significantly increased. This effect is mainly driven by an increased early exit rate through the PES channel, which has a long-lasting positive effect on the reemployment probability. At the end of the observation period receivers of vacancy information are about 150% more likely to have exited unemployment through the PES channel. Regarding the exit from non-PES channels, we observe an initial drop in exit rates, which is consistent with a substitution of search effort outlined in the job search literature, e.g., Fougère et al. (2009). Subsequently, however, the sign of the effect changes, becoming positive or zero. As we do not observe whether vacancy receivers are subject to more intensive monitoring, we cannot rule out that this is driven by monitoring (see, e.g., Abbring et al., 2005; Cockx et al., 2011). However, we find convincing evidence that this is explained from a lowered participation probability in ALMP, which results in a locking-in of non-receivers. For East Germans, the reduction of ALMP

participation occurs temporarily between the second and the fourth month after unemployment entry. For men and West Germans the effect occurs after the fourth month and onwards, the size of the reduction is about 25% for both groups. The early timing for East Germans suggests that ALMP may serve as substitute for vacancy information during the initial activation in East Germany, but are used as complementary activation if unemployment persist for unemployed males and in West Germany. With respect to the characteristics of accepted jobs we find small and weakly significant negative effects on the number of hours worked for men and East Germans of 2% and 3% respectively, which entails a reduction in overall income, as the hourly wages are not affected. These findings are consistent with those of van den Berg et al. (2013).

Overall, our analysis hence supports the hypothesis that early vacancy information has a 'double-dividend' in that it increases the exit rates from unemployment and lowers the use of more extensive ALMP. Our analysis further suggests that an increase in the quality of matched vacancy information is crucial to optimize the effectiveness of early job broking activities, by increasing the matching quality and the duration of the subsequent employment spell.

The chapter is structured as follows. In Section 2.2, the related literature is summarized. Section 2.3 outlines the institutional setup of unemployment benefit eligibility in Germany, characterizing the provision of vacancy information in more detail. Section 2.4 sets the stage for the empirical analysis, presenting the data and a descriptive assessment of the activation schemes used in practice. In Section 2.5, we describe the econometric approach and its empirical implementation. Section 2.6 presents the empirical results; Section 2.7 concludes.

2.2 Related Literature and Job Search Theory

Only few studies address the relative effectiveness of vacancy information in the overall activation process. Indirect evidence is given by studies assessing the effectiveness of caseworkers at the PES. As caseworkers select the activation measures, differences in effectiveness between caseworkers may be explained by different preferences for specific types of activation. Behncke et al. (2008) investigate the importance of caseworker connections to local firms, under the hypothesis that a higher connectedness results in a higher counseling quality and a higher availability of vacancies. They find that caseworker who are better connected with firms increase

the reemployment probability of unemployed by two to three percentage-points, which seems to confirm this hypothesis. Similarly, Lagerström (2011) finds a significant "caseworker effect" in the reintegration success of unemployed in Sweden, and provides evidence that more successful caseworker put a higher emphasis on providing job contacts rather than putting individuals into labor market training programs.

Pavoni et al. (2013) provide a normative analysis of the optimal use of job broking services within a principal-agent framework of the labor market, assuming that the caseworkers or the PES (the principal) cannot observe the search effort of the unemployed (the agent) and that human capital of the unemployed is decreasing over time. Aiming to balance incentive and insurance motives while considering budget constraints, the caseworker can either decide to let search be unassisted, or to provide vacancy information, which may however reduce incentive of own search. They show that a social planer should implement assisted search only when work-fare schemes are not feasible, and then only after an initial period of unassisted search, to reduce crowding out of own search intensity during a period of high productivity of own search.

More commonly, the effects of vacancy information from the PES is addressed in the context of unemployed job search behavior. Based on the observation that unemployed rely on multiple channels to gain information about employment opportunities, the effect of an exogenous arrival of vacancy information on the exit from unemployment, and the subsequent employment quality is assessed. The predications of this literature can be used as guideline for our empirical analysis. As we cannot distinguish between monitored and non-monitored vacancy information, we outline the expected results for either type of vacancy information, and outline expected differences in effects.[3]

Vacancy information without monitoring Assuming that job seekers choose their search effort endogenously and that job search is costly in terms of time and money, job search models posit that unemployed maximize the returns to search by focussing on search channels that exhibit the highest returns to search. The

[3]We only briefly and informally outline these search theoretic predictions and refer the interested reader to the papers of Holzer (1988), van den Berg and van der Klaauw (2006) and Fougère et al. (2009) for more details. The following section is a condensed summary of their findings. Note, that our subsequent discussion implicitly assumes that both the application to vacancies from the PES channel, as well as the application to jobs from non-PES channels require costly search effort.

productivity of each search channel is defined by the vacancy arrival rate, and the probability that a vacancy is turned into a job offer. Provided that the unemployed search via the PES channel, the arrival of vacancy information hence constitutes a direct increase in search productivity. To illustrate the predictions of the effect of this channel-specific productivity shock on labor market outcomes, we use previous search-theoretic results of Holzer (1988), van den Berg and van der Klaauw (2006) and Fougère et al. (2009). We assume that search effort can be directed to two stylized search channels, the PES channel and non-PES channels; the latter typically includes search via the network of friends, the internet, newspapers, posting own ads, etc.

Assuming substitutability between PES and non-PES channels, the arrival of vacancy information from the PES unambiguously increases the exit rate from the PES channel. In contrast, the exit rate from non-PES channels is likely to be reduced, due to the substitution of search effort away from non-PES channels towards PES channels, and due to an increase in productivity of search that is expected to increase the selectivity (reservation value) in the acceptance of job offers. The reduction in the exit rate from non-PES channels is however not expected to be sufficiently strong to fully counteract the positive effect on the exit rate from the PES channels[4] (van den Berg and van der Klaauw, 2006). In a study of the effectiveness of job contacts of the PES in France, Fougère et al. (2009) confirm the outlined theoretical predictions, as they find that job contacts have a significantly positive effect on the exit rate from unemployment. Dividing the analysis by labor market subgroups, they further find that the acceptance rate of jobs generated through job contacts of the PES is highest for low-skilled and low-educated unemployed.

The notion that the PES serves as information source of last resort for unemployed who are not able to find employment otherwise is also supported by previous literature comparing the use and returns to specific search channels (see, Holzer, 1988; Osberg, 1993; Gregg and Wadsworth, 1996; Addison and Portugal, 2002; Weber and Mahringer, 2008). Here, it is commonly found that unemployed using the PES search channel are negatively selected in terms of characteristics

[4] A positive total effect on the exit rate requires that the rate with which acceptable job offers are increased in the PES channel needs to compensate the rate with which acceptable job offers are decreased in the non-PES channels. In practice, this depends on the relative distribution of wages in the PES and the non-PES search channel. While the exact location of wages are difficult to verify, the wages in the PES channel are expected to be lower or similar than the wages offers in non-PES channels (see Section 2.3), so that this condition is likely to hold.

that positively affect labor market success. The ceteris paribus probability of finding a job through the PES channel is usually higher for individuals with relatively worse labor market characteristics. As these studies do not observe the receipt of vacancy information, the quality of vacancy information cannot be observed. Hence, whether negative employment effects are driven by negative selection or low quality vacancy information cannot be assessed (e.g., Addison and Portugal, 2002). From a search-theoretic perspective, the quality of accepted jobs is expected to be similar or increased due to the arrival of vacancy information, as the unemployed can always rely on non-PES channels to generate job offers. However, when the PES channel is the only channel of search available, the arrival of low quality vacancy information from the PES may cause a reduction in the quality of employment.

Vacancy information with monitoring In context of universal policy reforms entailing an intensified monitoring of unemployed job search efforts, a more recent literature addresses the effects of monitored vacancy information. It is common in many countries that the application to a proposed vacancy is monitored, whereas compliance is enforced with the threat of sanctioning (OECD, 2007). A straightforward effect of monitoring the application to vacancies is that the exit rate from the PES channel is further increased relative to no monitoring, as some unemployed are forced to invest a sub-optimal high level of effort to PES search. The distortion from optimality is also expected to lower the reservation value of employment, thus lowering the quality of accepted jobs (Abbring et al., 2005; van den Berg and van der Klaauw, 2006).

The effect on non-PES channels depends on whether the search effort invested in non-PES channels is also monitored ("job search monitoring"). In case where both types of effort are monitored, the exit rate from non-PES channels is not likely to react very strongly, as the unemployed are already operating at a sub-optimally high level of search. In case where only the application to vacancies is monitored, the exit rate from non-PES channels may be lowered, as incomplete monitoring of only one channel (the PES channel) allows substitution away from the non-monitored search channels (non-PES channel), thus counteracting the positive effect on the exit rate (van den Berg and van der Klaauw, 2006). In this case, the same predictions hold as in the non-monitoring case: the arrival of vacancy information increases the exit rate from the PES channel, but decreases the exit rate from non-PES channels. In contrast to before, it is not evident that

the positive effect on the PES channel outweighs the negative effect on the non-PES channels, and a lower quality and stability of the subsequent employment relationships may be expected.

Under certain institutional settings, monitoring of vacancy information may also increase the exit rate from non-PES channels. For example, when monitoring does not occur instantly but later during unemployment spell, and unemployed anticipate the date of monitoring, they might find it optimal to increase their search effort in all channels to exit unemployment before monitoring occurs. A corresponding model and empirical evidence for this 'front-loading' effect is given by Cockx et al. (2011) and Cockx and Dejemeppe (2012).

Two papers assess the effect of monitored vacancy information empirically. Van den Berg et al. (2013) study the effect of monitored vacancy information for unemployment benefit recipients in Germany. They find that the receipt of vacancy information has a significant positive effect on the exit rate out of unemployment, but a significant negative effect on the accepted quality of jobs in terms of wages and employment stability. The negative employment effects are likely to be driven by the threat of sanctioning, as it is also found that the vacancies increase the risk of sanctioning and the rate of sickness registrations — a status that allows the unemployed to suspend their search effort. Engström et al. (2012) show that in Sweden about one third of vacancy information from the PES does not result in an application, and study the scope of improving the application rate by announcing an increased monitoring of vacancy information. They find a significant but small positive impact on the application rate, but do not find any effect on the exit rate from unemployment, suggesting that the announced increase was not sufficiently credible.

In summary, the job search literature suggests that vacancy information increases the exit rate from unemployment by increasing the probability to find a job via the PES channel. In the presence of simultaneous monitoring or expected higher monitoring in the future, the exit rate from non-PES channels may also be increased. Otherwise the exit rate from non-PES channels is reduced to the substitution of search effort. With respect to the quality of accepted employment, no significant changes are expected in the absence of monitoring, provided that unemployed have alternative search channels at their disposal. However, if all other types of search are unproductive, or in the presence of simultaneous search monitoring, job acceptance monitoring and sanctioning, the arrival of vacancy in-

formation may lower the reservation wage of the unemployed, which results in the acceptance of lower quality employment. Assuming that vacancy information may be used as alternative to more intensive activation schemes, it is important to take account of potentially negative effects of job broking on employment quality, as this may increase the probability to reenter unemployment and hence the probability to participate in ALMP on the long-run.

2.3 Institutional Background

The German system of passive and active unemployment support is heavily centralized and provides all services from the PES. Prerequisite for redeeming unemployment benefit claims is the unemployment registration at the local PES office in charge, which is determined by the place of residence of the unemployed. The statutory framework for the unemployment benefit (UB) entitlements, the rights and duties of the unemployed, and the activities of the caseworkers are given by the Social Code III (*SGB III*). In the following, the regulations pertaining to the receipt and role of vacancy information during the job search process are outlined. Note, that we focus exclusively on the regulations related to the receipt of unemployment benefits. Unemployed who are not entitled to UB receive means-tested unemployment assistance, and are subject to different regulations. As our empirical analysis focusses on unemployed who are eligible to UB, we do no address these regulations here.

2.3.1 Entitlement to Unemployment Benefits

During the period of observation (between May 2007 and June 2009), unemployed individuals are entitled to UB when they have been employed subject to social security contributions for at least 12 months during the 24 month-long reference period preceding unemployment entry. The replacement rate is at 60% (67% for unemployed with children), and based on the average gross wage earned during the previous twelve months. The duration of UB entitlement depends on the duration of employment during the reference period, and the age of the individual. For individuals below the age of 50, the maximal duration of UB entitlement is 12 months, older individuals may be entitled to benefit receipt for 24 months.

An additional prerequisite for UB entitlement is the readiness to work, which

is defined as an active search for employment, the willingness to accept "reasonable" job offers (see definition below), and the availability for participation in active labor market programs (ALMP). As a consequence, UB can be temporarily withdrawn if the behavior of unemployed does not reflect the willingness to work.

Unemployed receiving job offers are generally required to accept them, unless the job offers are not considered "reasonable"; then the rejection does not entail a sanction. The *SGB III* outlines reasonability criteria, which are aimed to ensure that the unemployed can maintain a certain level of matching quality compared to their previous job. "Unreasonable" jobs include jobs with a wage less than 80% of the wage earned in the last employment, jobs requiring more than 2.5 hours commuting time, or jobs necessitating a change of occupation. Furthermore, short-term employment and employment requiring transitory separation from the family are not considered reasonable. The reasonableness criteria are tightened over the duration of the unemployment spell: after three and six months in unemployment, larger wage cuts (70% or 60% , respectively) and longer commuting times are considered "reasonable".

2.3.2 Vacancy Information and ALMP Use

Upon registering unemployed, all unemployment benefit claimants are required to attend a personal meeting with the caseworker, during which a detailed profiling of the unemployed is conducted.[5] The labor market profiling serves to summarize the labor market characteristics, work motivation, abilities and deficits of the unemployed, and to define a strategy for the subsequent activation process. The caseworker may categorize the unemployed into one of four activation groups reflecting the relative labor market closeness of the unemployed. While the intensity of contact is found to increase with the severity of labor market barriers, the activation strategy may be highly individualized and is continuously adapted over the course of the unemployment spell.

The labor market profiling also serves to define the types of employment sought, and to assess whether vacancies can be offered to the unemployed. Vacancy

[5]The interaction between caseworkers and unemployed is usually a "black box" — the following section relies heavily on the information gathered by structured caseworker interviews and evaluations of caseworker meetings at selected public employment services. Three different sources are used complementarily: Hielscher and Ochs (2009), Schütz et al. (2011a), and Boockmann et al. (2013).

information is generated by a computerized matching of the characteristics of the unemployed with the characteristics of vacancies registered at the PES. The matching hence requires little cost in terms of time and money. According to self-assessed use of working time, caseworkers devote most of their time to career counseling, but only very little time to the acquisition of vacancies (Boockmann et al., 2013). The availability of matching vacancies is hence driven by exogenous determinants rather than the effort of the caseworker. The discussion of vacancy information is an important component of the early meetings with the caseworker. All vacancies are "open" vacancies, so that they may be offered to more than one unemployed, and the unemployed still have to apply to get in contact with the employer. The caseworker may decide to monitor the application process — in practice, both monitored and non-monitored vacancies are observed (Schütz et al., 2011a).

While the "reasonableness" criteria protect the unemployed from being offered vacancies that are too far away from the previous job, few overall evidence is available on the selectivity and the quality of the registered, or matched, vacancies. The vacancies registered at the PES only represent a subset of all vacancies in the labor market. The average coverage rate (*Meldequote*), i.e., the share of all vacancies registered at the PES, ranged between 30% and 50% between 2007 and 2008 (IAB, 2008). Exemplary evidence for specific occupations and qualifications levels suggests that the coverage rate is lower for higher qualified jobs (Christensen, 2003; Koppel, 2008). Furthermore, jobs in the temporary work sector are highly over-represented at the PES relative to their importance in the labor market. During the period of observations, the stock of vacancies in temporary work agencies accounted for 30% to 40% of all registered vacancies, whereas the stock of employees in temporary work accounted only for 10% of all employed.[6]

Similarly, only little evidence is available regarding the selection of more intensive active labor market programs. While unemployed are entitled to placement vouchers — an instrument that subsidizes the training or intermediation services of private placement agencies — after having been unemployed for six weeks, the

[6]The high incidence of temporary work vacancies is likely to be a remnant of regulations in place between 2003 and 2006 that required that every PES had to install a Temporary Work Agency (TWA) who would then be in charge of training the unemployed or lease them in fixed-term employment. Now, the TWA are able to select their clients from the pool of registered unemployed themselves, which may have led to the strong over-representation of temporary vacancy postings at the PES (compare for more details of the regulation WZB and infas, 2006, pp. 266-303).

provision of all other measures is at the discretion of the caseworker. Based on selected caseworker interviews, the long-term labor market integration is considered the most important objective of the activation measures selected (Boockmann et al., 2013). However, also alternative factors, like PES-specific regulations, the availability of activation programs, and the specific request of unemployed are also stated to influence the decision making.

2.4 Data and Descriptive Statistics

To assess the effect of early vacancy information on labor market integration and ALMP participation, we use a representative survey of unemployment entrants in Germany. The sampling of the *IZA Evaluation Dataset S* (Caliendo et al., 2011) focussed on monthly unemployment entrants between June 2007 and May 2008, who were between 17 and 54 years of age, and who received, or were eligible to receive, unemployment benefits, and were hence subject to the regulations of the *SGB III*. Based on an initial interview conducted shortly after unemployment registration, two further interviews followed, after one and three years, respectively. In the baseline interview, the respondents were given an extensive questionnaire capturing information on general socio-demographic characteristics, the previous employment history, job search efforts, expectations, and personality traits. The questionnaire also included questions regarding the frequency of interaction with the PES, and what types of services the unemployed were offered during unemployment. The follow-up interviews were used to construct the subsequent labor market biography, capturing timing, duration and type of spells in employment, unemployment, education, active labor market programs and inactivity on a monthly level. Unfortunately, the data collection was marked by significant sample attrition, resulting in a reduction of the sample by about 50% from the first to the second interview.[7] In the third questionnaire, only participants of the second wave were contacted, a comparable attrition rate lead to a further reduction in sample size. To ensure a sufficient size of our estimation sample, we restrict the estimation to information collected up to the second questionnaire.

To ensure that all unemployed are subject to the regulations of the *SGB III*,

[7]Whereas the initial sample consisted of 17,396 unemployed, only 8,915 individuals participated in the second interview. Comparisons by age, gender and migration status do not provide evidence for selective attrition, except for the youngest age cohorts of less than 25 year-olds that were slightly less likely to participate in the second wave (Caliendo et al., 2011).

we restrict the empirical analysis to unemployed who stated to actively search for full-time employment, and further exclude all unemployed who could be identified to receive means-tested benefits[8]. Finally, all observations with missing values in any of the relevant variables were excluded, resulting to a total sample size of 5,126 unemployed. Table 2.1 summarizes the sample selection. Note, that our sample selection criteria are likely to create a positively selected sub-sample of unemployed relative to all unemployment entrants. Within this estimation sample,

Table 2.1: Sample selection criteria and number of observations

Reduction criterion	N	%-reduction
Respondents of the 2nd wave	8,915	
actively searched for employment	7,088	20.49
searching full-time employment and not SGB II	5,534	37.94
non-missings in relevant variables	5,126	41.49

Source: IZA Evaluation Dataset S, own calculations.

the subsequent analysis is constructed within gender-groups and within region-of-work groups, i.e., East and West Germany. This is done to assess potential effect heterogeneity with respect to overall search productivity and differences in activation strategies. Due to the structurally worse labor market conditions in East Germany[9], East Germans are expected to experience more difficulties in finding employment through alternative channels than West Germans, and may also face activation efforts that are more focussed on removing demand-sided barriers (employment subsidies) than West Germans. We separate by gender based on previous findings suggesting that women are less likely to be assigned to ALMP, irrespective of expected program success (Müller and Kurtz, 2003). Also, women may be more constrained by familial issues than men, something that is difficult to observe in the data but may be observed by the caseworker, and hence affect both activation and labor market outcomes. The respective sub-samples contain 2,898 observations for men, 2,228 observations for women, 1,647 unemployed in East Germany, and 3,479 West German unemployed, respectively.

The average timing of the initial interview was nine weeks after unemployment registration. Due to this delay, about 25% of unemployment entrants had

[8]For some unemployed the receipt of UB or social assistance could not be identified, as the question in the survey asked about their *current* receipt of financial support. Hence, eligible unemployed who had not yet received the payment were categorized as non-receivers.

[9]During the time of observation, the unemployment rate in East Germany was with on average 16% still twice the size of the unemployment rate in West Germany.

already left unemployment at point of the initial interview. As their quick exit may have been the result of receiving vacancy information, their omittance would lead to biased effect estimates. We hence include them in the sample, and control for the timing of the interview in our empirical analysis to account for potentially differential response behavior.[10]

We construct the treatment indicator using a question on the services offered from the PES during the unemployment spell, listing 14 alternative activation schemes. The respondents could select multiple answers; amongst others, they were asked whether they were offered vacancy information for regular fulltime jobs, part-time jobs and jobs in TWA. To facilitate the interpretability of the quality of vacancies received, we use fulltime vacancy information as our main treatment indicator, and assess the sensitivity of our estimates to extending our treatment definition in Section 2.6. The last row of Table 2.2 provides an overview of the treatment probabilities for the respective labor market groups. While about 43% of men and West German unemployed received fulltime vacancy information, only 36% of women and 32% of East Germans received information. The differential treatment propensity may be indicative of differences in overall labor market conditions or differences in overall activation strategies. In the following we descriptively explore these differences.

2.4.1 Alternative Activation Offers

To assess whether receivers and non-receivers of fulltime vacancy information[11] are subject to differential activation, we cross-tabulate our treatment indicator with the responses to being offered alternative activation schemes. The columns of Table 2.2 depict the unconditional treatment difference for gender and region of work groups. For better readability, we differentiate between measures aimed at a direct labor market entry, i.e., job offers, and subsidized self-employment, and measures aimed at an intermediate integration into other types of measures, e.g., job creation schemes, training programs, or private intermediation services.

Across all labor market groups, the receivers of fulltime vacancy information are significantly more likely to receive alternatives types of vacancy information,

[10]Note, that both groups were given the same baseline questions, whereas the outcome questions were the same way in both questionnaires. Hence, no systematic bias in responses is expected by getting outcome information from either the first or the second interview.

[11]Also referred to as 'treated' and 'controls' in the subsequent analysis.

Table 2.2: Activation services offered by the PES by treatment indicator

Type of service offered	Women			Men			East Germany			West Germany		
	no VI	VI	*t*-test	no VI	VI	*t*-test	no VI	VI	*t*-test	no VI	VI	*t*-test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Direct labor market entry												
Job in TWA	0.03	0.09	0.00	0.04	0.14	0.00	0.02	0.07	0.00	0.04	0.14	0.00
Self-Employment subsidy	0.04	0.05	0.28	0.06	0.06	0.91	0.04	0.05	0.36	0.06	0.06	0.97
Marginal employment	0.02	0.01	0.71	0.01	0.02	0.09	0.02	0.03	0.53	0.01	0.01	0.21
Regular full-time job	0.00	1.00		0.00	1.00		0.00	1.00		0.00	1.00	
Regular part-time job	0.06	0.26	0.00	0.02	0.07	0.00	0.04	0.14	0.00	0.04	0.15	0.00
Apprenticeship place	0.01	0.02	0.71	0.01	0.02	0.66	0.01	0.01	0.45	0.01	0.02	0.41
Intermediate integration in training or job creation scheme												
One-Euro-Job	0.01	0.01	0.75	0.01	0.03	0.00	0.01	0.02	0.12	0.01	0.02	0.04
Job Creation Scheme	0.01	0.02	0.64	0.01	0.03	0.01	0.01	0.01	0.64	0.02	0.03	0.04
Work-training	0.15	0.19	0.02	0.11	0.18	0.00	0.11	0.15	0.01	0.14	0.19	0.00
Employability training	0.10	0.14	0.00	0.07	0.11	0.00	0.06	0.08	0.12	0.10	0.13	0.00
German language course	0.00	0.00	0.57	0.00	0.00	0.74	0.00	0.00	0.04	0.00	0.00	0.28
English language course	0.03	0.05	0.02	0.02	0.01	0.27	0.02	0.01	0.08	0.02	0.03	0.10
Training voucher	0.08	0.07	0.84	0.04	0.06	0.05	0.04	0.05	0.37	0.07	0.07	0.80
Placement voucher	0.08	0.11	0.05	0.07	0.11	0.00	0.12	0.16	0.03	0.05	0.09	0.00
N	1,420	808		1,654	1,244		1,118	529		1,956	1,523	
%	0.64	0.36		0.57	0.43		0.68	0.32		0.56	0.44	

Source: IZA Evaluation Dataset S, own calculations. All numbers are shares. The unemployed were asked the question: "Since you entered unemployment in (date), have you ever been offered one of the following from the local employment agency or the jobcenters?" Multiple answers were possible.

which may be explained by general differences in the overall availability of vacancies at the PES. The receipt of further vacancy information also differs systematically across labor market groups. While women are much more likely to receive part-time offers then men, men and West Germans are more likely to receive TWA job offers.

The unconditional probabilities further suggest a positive relation between more intensive measures and vacancy receipt. The most common types of activation offered are employability training, work place training and placement vouchers, with 10%, 15% and 9%, respectively in the overall sample. Across all labor market groups we observe a significant positive relation with fulltime vacancy receipt. The probability to be offered job creation schemes and alternative training measures, e.g., language courses, is rather low (1% to 5%); the differences between treatment groups are not significant or vary in sign. Across labor market groups, we find that women have a higher propensity to receive training-related activation offers than men. East Germans are less likely to receive training-related offers, and work-training schemes than West Germans, but have a higher probability to be offered placement vouchers.

The positive relation between the higher intensity schemes and the receipt

of vacancy information goes against the notion that caseworkers dichotomize their activation strategies to either direct labor market entry or intermediate integration into activation schemes, as that would suggest a negative relation between these two measures. Instead, these patterns suggest that caseworkers who offer vacancies are generally more likely to offer more intensive activation schemes. This may be driven by the labor market characteristics of the unemployed, or differences in caseworker behavior.

2.4.2 Characteristics of the Unemployed

Selected descriptives on the socio-demographic characteristics, employment history and job search behavior of treated and controls are depicted in Table 2.3. With respect to heterogeneity across labor market groups, we find that women in our sample are similarly educated as men, but are less likely to have children, and invest more effort into search than men. However, they also have less labor market experience on average, and are less willing to move for employment than men, so that they might still face higher labor market restrictions than men. East and West German unemployed differ in that East Germans are on average older, are more likely to have a professional degree, rather than a university degree or no degree, and their labor market history is characterized by more and longer spells in unemployment.

Across treatment groups we find that un-married, younger unemployed and unemployed with a professional training degree, rather than no vocational education or a university degree, are significantly more likely to receive vacancy information. Furthermore, unemployed who entered unemployment from employment, and unemployed with less previous unemployment experience are more likely to receive vacancy information. Receivers of vacancy information also have a significantly higher contact frequency with the caseworker, and a higher probability to receive unemployment benefits.

Overall, this suggests a positive selection with respect to overall labor market chances, whereas the lower availability of vacancy information for university graduates is suggestive of selective vacancy registration at the PES. The differences by caseworker contact frequency provide additional evidence that receivers of vacancy information may be subject to a higher overall activation intensity. We also find that receivers of vacancy information search more intensively, in that they sent out

more applications and use a higher number of search channels. This may result from the higher overall activation intensity, but may also be due to the receipt of vacancy information. As outlined in Chapter 2.2, the receipt of vacancies may result in a higher intensity of search, if it is accompanied by a higher overall monitoring level and sanctioning risk. Unfortunately we do not have direct information on this in the data. Note, that we do not find any differences with respect to the probability to have exited unemployment at point of the initial interview, which suggests that the observed treatment indicator is not influenced by the timing of the interview.

2.4.3 Outcomes of Interest

As one of the main outcomes of interest, we consider the speed of transiting into the employment subject to social security contributions. The timing of the first employment entry is constructed using information on monthly labor market states between unemployment registration and the second interview, covering a period of 13 months duration. Exits from unemployment that occur later than 13 months are assumed to be subject to random censoring. To analyze the specific mechanism by which vacancy information affects the exit rate, we consider the joint hazard rate as well as the channel-specific hazards, differentiating between exits through the PES channel, and non-PES channels, respectively. The definition of non-PES channels takes into account that some channels may be considered complements to the PES, as the online information system (SIS) of the PES, and the search through placement officers with a voucher. For these channels an increase in exits could be observed due to spill-over effects of information. To simplify the interpretation of the effects, we hence exclude exits through these two channels from the set of non-PES channels.[12]

The first rows of Table 2.4 show the unconditional probability to enter regular employment during the period of observation. The average probability to find employment is at 62%. For women (men) it is slightly lower (higher) and 60% (64%), no significant differences emerge for East and West Germany. Across treatment groups we observe a substantially higher reemployment probability for receivers of vacancy information. The differences lie between 5%-points (men) and 12%-points (women).

[12]Note, that this only makes up for about 6% of all exits. Sensitivity checks show that this restriction does not affect our estimates.

Table 2.3: Labor market characteristics and job search information by treatment indicator

	Women			Men			East Germany			West Germany		
	no VI	VI	t-test	no VI	VI	t-test	no VI	VI	t-test	no VI	VI	t-test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Socio-demographic characteristics												
West Germamy	0.63	0.74	0.00	0.64	0.74	0.00	0.00	0.00		1.00	1.00	
Female	1.00	1.00		0.00	0.00		0.47	0.39	0.00	0.46	0.39	0.00
Has a child	0.37	0.34	0.35	0.43	0.40	0.14	0.38	0.34	0.23	0.42	0.39	0.19
Married	0.41	0.33	0.00	0.38	0.33	0.01	0.44	0.33	0.00	0.37	0.33	0.03
Age of the respondent												
Less than 24 years	0.21	0.25	0.01	0.23	0.24	0.25	0.21	0.28	0.00	0.22	0.24	0.19
Between 25 and 34 years	0.25	0.29	0.02	0.26	0.26	0.93	0.22	0.24	0.32	0.27	0.29	0.43
Between 35 and 44 years	0.26	0.23	0.11	0.25	0.27	0.12	0.25	0.24	0.46	0.25	0.26	0.50
Between 45 and 54 years	0.29	0.23	0.00	0.27	0.22	0.01	0.31	0.25	0.00	0.26	0.21	0.00
Vocational education												
None	0.08	0.06	0.10	0.09	0.06	0.00	0.04	0.03	0.21	0.11	0.07	0.00
Professional training	0.68	0.74	0.00	0.66	0.70	0.02	0.74	0.77	0.22	0.63	0.70	0.00
Technical college/university	0.25	0.21	0.03	0.24	0.23	0.53	0.21	0.20	0.48	0.26	0.23	0.03
Employment history												
Employment status before unemployment												
Employed	0.72	0.73	0.86	0.75	0.78	0.02	0.75	0.77	0.33	0.72	0.76	0.04
School, apprenticeship, military	0.17	0.18	0.60	0.18	0.14	0.01	0.17	0.16	0.70	0.18	0.16	0.09
Maternity Leave	0.03	0.03	0.86	0.00	0.00	0.44	0.02	0.01	0.21	0.01	0.02	0.25
Other	0.08	0.06	0.23	0.08	0.07	0.51	0.06	0.05	0.68	0.09	0.07	0.12
Reason for terminating last employment												
Did not have a job before	0.16	0.15	0.60	0.13	0.11	0.12	0.14	0.14	0.87	0.14	0.12	0.04
Quit personally	0.11	0.12	0.72	0.07	0.08	0.17	0.05	0.07	0.21	0.11	0.10	0.73
Laid-off by employer	0.68	0.69	0.54	0.73	0.75	0.20	0.76	0.73	0.26	0.68	0.73	0.00
Further reasons	0.05	0.05	0.35	0.07	0.06	0.10	0.05	0.06	0.52	0.07	0.05	0.01
Share of adulthood in...												
unemployment	0.06	0.06	0.23	0.06	0.06	0.09	0.08	0.07	0.11	0.05	0.05	0.85
employment	0.59	0.59	0.69	0.70	0.72	0.12	0.66	0.65	0.55	0.64	0.68	0.00
Previous unemployment spells												
Number of spells	1.73	1.52	0.05	2.23	1.92	0.00	2.49	2.17	0.04	1.73	1.62	0.23
Long-term unemployed	0.23	0.16	0.00	0.18	0.18	0.82	0.27	0.22	0.01	0.16	0.16	0.85
Search behavior and interaction with the PES during unemployment												
Number of visits to the PES												
0 to 2	0.50	0.41	0.00	0.46	0.36	0.00	0.46	0.33	0.00	0.49	0.39	0.00
3 to 5	0.45	0.49	0.08	0.45	0.52	0.00	0.46	0.54	0.00	0.44	0.49	0.00
≥ 6	0.05	0.11	0.00	0.09	0.12	0.01	0.09	0.13	0.01	0.07	0.11	0.00
Unemployment benefit receipt												
Current receipt (yes/no)	0.73	0.79	0.00	0.74	0.80	0.00	0.78	0.80	0.35	0.72	0.80	0.00
Level of UB (Euro)	443.2	492.2	0.01	584.4	605.9	0.26	463.5	489.3	0.20	551.0	586.1	0.05
Search intensity												
Number of own applications	15.38	16.87	0.20	14.46	15.80	0.07	14.99	16.19	0.42	14.82	16.24	0.04
Zero applications	0.06	0.03	0.00	0.06	0.04	0.03	0.07	0.05	0.10	0.06	0.03	0.00
Number of search channels[1]	4.70	5.31	0.00	4.55	5.17	0.00	4.64	5.01	0.00	4.60	5.30	0.00
Willingness to move	0.26	0.28	0.18	0.30	0.34	0.03	0.27	0.35	0.00	0.28	0.31	0.14
Unemployed at first interview	0.74	0.72	0.43	0.70	0.70	0.96	0.75	0.73	0.35	0.70	0.71	0.86
N	1,420	808		1,654	1,244		1,118	529		1,956	1,523	

Source: IZA Evaluation Dataset S, own calculations. All numbers are shares, unless indicated otherwise.

As outlined before, we are additionally interested in the effects on the participation probability in active labor market programs. As information on subsidized employment was not recorded consistently across the two interview periods, we only consider job creation schemes and training programs in our definition of ALMP. The training programs include subsidized participation in further schooling and training, publicly sponsored retraining measures, short-term training measures, and job search courses.[13] The first two rows of Table 2.4 show the unconditional probability to enter regular employment and ALMP during the period of observation. The average probability to participate in ALMP is at 20%. However, no differences are found across labor market, or treatment groups.

For unemployed who entered regular employment during the period of observation, we assess the effect of vacancy information on the quality of the first accepted job. We observe information on the hourly wage levels, the weekly number of hours worked, and whether the accepted jobs were temporary, i.e., limited in their duration to less than one year, and whether the job was at a temporary work agency (TWA). As an additional validation of our treatment indicator we also assess the unconditional differences to exit via the PES channel. The lower rows of Table 2.4 present descriptive statistics on the quality indicators considered.

It can be seen that receivers of vacancy information have a substantially increased probability to find employment via the PES channel. The probability to exit via the PES channel is at 8% for non-receivers of all subgroups. For receivers, the exit rate is between 8%-points to 10%-points higher. Furthermore, in all labor market groups, except women, we see that treated have a significantly elevated probability to enter TWA employment. For the remaining indicators we find substantial variation across labor market groups, but no differences by treatment status. The probability to enter short-term employment is 49% for women, but only 35% for men; 44% for East Germans and 39% of West Germans. The hourly wages also differ significantly, and in the expected direction: women earn less than men (€7.4 vs. €8.3) and East Germans less than West Germans (€7.0 vs. €8.4). As before no difference emerge across treatment groups. With respect to weekly hours worked, we find that women work less hours than men.

[13]Regions with low labor demand are more likely to use demand-stimulating wage subsidies as activation measure, rather than supply-targeted training programs. Our estimates may thus understate the true effect on ALMP participation in these regions. However, as the employment subsidy may also have been part of the vacancy offer, our non-employment definition of ALMP also rules out an understatement of the true employment effect.

Table 2.4: Successful channel and quality of first employment by treatment indicator

	Women			Men			East Germany			West Germany		
	no VI	VI	*t*-test	no VI	VI	*t*-test	no VI	VI	*t*-test	no VI	VI	*t*-test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Labor Market Outcomes												
Reemployment probability	0.56	0.68	0.00	0.62	0.67	0.00	0.61	0.67	0.00	0.58	0.68	0.00
Probability to enter ALMP	0.22	0.23	0.67	0.20	0.20	0.86	0.21	0.22	0.66	0.21	0.21	0.97
N	1,420	808		1,654	1,244		1,118	529		1,956	1,523	
Characteristics of first employment spell												
Successful PES channel	0.08	0.18	0.00	0.08	0.16	0.00	0.08	0.18	0.00	0.08	0.16	0.00
Temporary work agency	0.12	0.14	0.17	0.16	0.20	0.06	0.14	0.18	0.06	0.15	0.17	0.08
Temporary employment	0.49	0.48	0.75	0.35	0.35	0.94	0.44	0.44	0.97	0.39	0.39	0.85
Hourly wage (Euro)	7.47	7.48	0.93	8.30	8.34	0.77	7.09	6.95	0.41	8.44	8.36	0.52
Hours worked (log)	3.60	3.62	0.19	3.76	3.75	0.26	3.71	3.71	0.92	3.68	3.70	0.13
N	790	549		1,027	834		678	353		1,139	1,030	

Source: IZA Evaluation Dataset S, own calculations. $N = 3,200$. Sample of unemployed who entered regular employment within 13 month of their initial unemployment registrations.

2.5 Econometric Analysis

To formalize the evaluation problem, let D denote a binary treatment indicator with $D = 1$ when a vacancy was received and $D = 0$ when no vacancy was received, and let Y_1 and Y_0 denote the outcomes realized after treatment participation or after non-participation, respectively. Note, that the unconditional differences shown in the previous section can hence be represented by $\Delta = E(Y_1|D = 1) - E(Y_0|D = 0)$.

Following the potential outcome framework developed by Roy (1951) and Rubin (1974), a causal interpretation of Δ requires that the observed non-treatment outcome of controls $E(Y_0|D = 0)$ can be used as approximation for the hypothetical and unobserved non-treatment outcome of the treated $E(Y_0|D = 1)$. In the absence of random variation in treatment assignment this is not likely to hold, as non-random selection into treatment results in systematic differences in outcomes even in the absence of the treatment. An assumption commonly invoked in this context, is the assumption of conditional independence (CIA), which states that all systematic differences in the control outcomes can be eliminated by controlling for pre-treatment characteristics X, so that conditioning on these characteristics renders treatment status and outcomes conditionally independent,

$$D \perp\!\!\!\perp Y_0|X. \tag{2.1}$$

A weaker assumption is independence of conditional means, $E_{X|D=1}(Y_0|D = 0, X) =$

$E_{X|D=1}(Y_0|D=1,X)$, so that the causal average treatment effect on the treated (ATT) can be calculated as,

$$\Delta_X^{ATT} = E_{X|D=1}(Y_1|D=1,X) - E_{X|D=1}(Y_0|D=0,X). \tag{2.2}$$

Whether the CIA assumption can be used for identification in an empirical analysis needs to be justified case-by-case, and depends on the availability of a sufficiently informative set of X that are known to affect outcomes, and are distributed differently in treatment and the control group. We discuss the plausibility of this assumption in our analysis in detail later.

An additional assumption required for a causal interpretation of Δ_X^{ATT}, is the "stable unit treatment assumption" (SUTVA), which states that the treatment only affects the treated, ruling out spill-over effects, peer effects, or general equilibrium effects. In our context, the number of jobs in the labor market may be limited so that the unemployed compete for vacancies. While the provision of vacancy information may create additional competition for other treated, we are consider "open" vacancies that are also posted online and hence technically available to everyone. The additional competition that arises from informing one additional unemployed is hence expected to be small, so that the SUTVA is assumed to hold. Finally, to ensure that Δ_X^{ATT} is not based on extrapolation, we need to make a common support assumption. By calculating the conditional outcomes over the distribution of characteristics amongst the treated, the counterfactual last term of equation (4.5) can only be constructed for characteristic values appearing in both the treatment and the control group. Formally, this condition is given by the set of characteristics S_X for which $S_X = \{X|P(D=1|X) < 1\}$ is fulfilled. In contrast to the CIA, the common support assumption can be assessed in the empirical analysis, and violations can be fixed by eliminating characteristic combinations that lie outside of the common support.

2.5.1 Empirical Strategy

As the exit rate from unemployment is a function of time t since entry into unemployment, a standard approach to estimate treatment effects is the estimation of a parametric mixed proportional hazard model (MPH) as outlined in van den Berg (2001). By specifying an unobserved heterogeneity distribution, MPH models bear the advantage that the hazard function can be consistently estimated; in the

estimation of competing risk models, the specification of unobserved heterogeneity distribution may further allow to identify the marginal hazard rates, which are otherwise not identified. At the same time, the assumption of proportional hazard may be rather restrictive as it is difficult to justify by economic theory. Furthermore, in a single-spell model, the correct specification of the hazard function is not expected to assist identification of our static treatment parameter (Nicoletti and Rondinelli, 2010). Against this background it may be preferred to follow a more flexible semi-parametric estimation approach as suggested by Fredriksson and Johansson (2008) and Crepon et al. (2009). Here, semi-parametric matching on the propensity score (Rosenbaum and Rubin, 1983b) is conducted in a first step, followed by the non-parametric estimation of hazard rates of treated and controls in the matched sample. The estimation of the cause-specific hazard rates proceeds in a similar fashion, taking the exits from the respective other channels as censored. As we are mainly interested in the mechanism by which the exit into employment is achieved, the marginal effect of vacancies on counterfactual exit rates is not of particular interest. We hence adopt the more tractable semi-parametric approach to estimate the hazard functions. A similar propensity score matching approach can be straightforwardly adapted to estimate the effect on the probability to participate in ALMP, and the quality of employment outcomes. Here, however, the focus is on the conditional means in outcome measures.

As outlined before, our definition of treatment and control group is based on the implicit assumption that the receipt of vacancy information before the first interview reflects general differences in the exposure to vacancy information over the course of the whole unemployment spell. Recall, that the provision of vacancy information is one of the main tasks discussed during the early meetings with the caseworker (see Section 2.3). In our sample, 98% of unemployed have had at least one meeting with the caseworker, so that it is reasonable to assume that the differences in early receipt of vacancy information also reflect differences in the receipt later during the unemployment spell. Clearly, however this may not hold true for all controls, so that part of our control group may receive vacancy information later, which might bias our estimates towards zero. A number of studies discuss the dynamic treatment assignment in combination with dynamic selection out of unemployment, and show that an adequate handling usually requires knowledge of the timing of treatment, which unfortunately cannot be observed (see, Sianesi, 2004; Fredriksson and Johansson, 2008; Crepon et al., 2009; Vikström et al., 2012).

Hence, the potential attenuation bias of our estimates needs to be kept in mind when interpreting the results.

2.5.2 Conditional Independence Assumption

The institutional setup and the descriptive analysis suggest that the receipt of vacancy information may be subject to various types of selection, which need to be addressed by the empirical analysis. First, the registration of vacancies at the PES is likely to be indicative of the overall demand in the labor market, so that the availability of matching vacancies at the PES is positively correlated with overall availabilities of employment opportunities. To account for this positive relation, we control for individual and labor market characteristics capturing differences in expected labor market success, e.g., demographic characteristics, schooling and vocational education indicators, indicators of past labor market performance, as well as a measure of personality traits (see Goldberg, 1993, for the "Five Factor" model). Furthermore, local labor market indicators are controlled for, capturing the unemployment rate, the vacancy rate and the share of long-term unemployed. To capture systematic differences in the relative productivity of other search channels we control for the availability of internet and the number of friends and colleagues. Finally, we also control for regional differences in the coverage rate of vacancies across over all sectors, and for the coverage of the PSA sector separately. To capture regional differences in the use of certain policies, we control for regional difference on the use of active labor market polices and the occurrence of sanctions, in either specification.

A second endogeneity issue to address is the simultaneous receipt of vacancy information and other treatments and services. The descriptive analysis shows a strong positive relation between the receipt of vacancy information and being offered alternative, more extensive measures, which seem to suggest that treated and controls differ with respect to their overall propensity to receive assistance from the caseworker. As caseworkers seem to use the activation strategies interchangeably, the absence of vacancy information may influence the probability to offer alternative measures and hence affect the probability to be treated. An unbiased estimation of our treatment effects hence requires that differences in treatment offers are controlled for. We also control for differences in the contact frequency with the caseworker, as this may capture further differences in the assistance or monitoring the unemployed is subject to.

While it is well-established that it is important to control for extensive information on past labor market history, to account for unobserved heterogeneity with respect to labor market attachment (Lechner and Wunsch, 2013), so far little previous evidence is available on the interplay between multiple types of activation offers. To make the effect of these control variables transparent, we present two types of effect estimates, one based on the baseline specification, excluding information on the treatment offers and the frequency of contact, and a 'full' specification including these indicators.

2.5.3 Implementation of the Matching Estimator

In a seminal paper, Rosenbaum and Rubin (1983b) show, that instead of conditioning on all confounders X directly, one can also condition on a single summary measure, the propensity score, to render two treatment groups conditionally independent. The propensity score is estimated as the conditional probability to receive treatment, while controlling for all characteristics X that are assumed to be important to fulfill the CIA condition. Hence, propensity score matching (PSM) requires the estimation of the propensity score in a first step; we use a probit regression model to bound the predicted values between zero and one, including the control variables outlined in the previous section.[14] As the propensity score only represents a summary score of the confounders, a consistent estimation of the treatment probability is not required (Zhao, 2008; Waernbaum, 2012). To rule out that outliers in the predicted probabilities get too much weight in the matching analysis, we impose a common support condition excluding treated observations with propensity score values (smaller) larger than the (minimal) maximal value of the controls - and vice versa for controls (Dehejia and Wahba, 2002). The elimination of the extreme values resulted in a deletion of only very few treated observations (five men, five East Germans, one in each of the remaining groups), and is hence not expected to alter the representativeness of our estimation sample.

Matching is conducted using kernel matching with an Epanechnikov kernel, which has the feature of weighing more distant observations downward, and only considering control observations within a particular radius as defined by the selected bandwidth parameter. The estimator has been shown to produce reliable estimates under a number of data settings, and is quite robust to the choice of the

[14] Additionally we include information on month of entry into unemployment, and elapsed unemployment duration at point of the interview.

bandwidth (Huber et al., 2013). As the matching estimator is intended to maximize the balance in covariates X across treatment groups, we test the sensitivity of the balancing quality to the choice of the bandwidth, adopting an approach proposed by Huber et al. (2012). Here, the bandwidth is selected as a multiple of the largest distance in propensity scores obtained from pair-matching with replacement. This allows a data-driven selection of the optimal bandwidth by taking into account the relative position of treated and controls. We conduct a grid search for various multiples of this value, comparing the level of balance achieved for each value, ultimately selecting the bandwidth that maximizes the balance within each subgroup.[15]

Table 2.5: Summary of balancing quality: t-test and standardized bias

Two-sample t-test	Women		Men		East Germany		West Germany	
	Unmatched Sample	Matched Sample	Unmatched Sample	Matched Sample	Unmatched Sample	Matched Sample	Unmatched Sample	Matched Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of characteristics with p-value								
less than 0.01	15	0	16	0	15	0	14	0
less than 0.05	24	0	27	0	23	0	25	0
less than 0.10	31	0	36	0	31	0	30	0
less than 0.20	38	0	51	0	37	0	40	0
less than 0.30	48	0	56	0	46	0	48	0
less than 0.40	55	1	58	0	53	1	55	0
less than 0.50	61	2	63	1	59	2	66	1
less than 0.60	66	6	68	5	71	12	69	4
less than 0.70	77	14	74	16	78	29	75	17
less than 0.80	84	39	81	33	81	49	82	31
less than 0.90	87	63	85	63	91	70	87	60
less than 1.00	94	94	93	93	94	94	94	94
Mean Standardized bias	6.46	1.04	5.9	0.85	6.86	1.51	4.88	0.73

Source: IZA Evaluation Dataset S, own calculations. Matching was done using kernel matching with an Epanechnikov kernel and optimal bandwidth that was selected to minimize the difference in characteristics in the matched sample. Varying numbers of variables arise due to differences in the specification of the propensity score model. Results are based on the 'full' specification.

The balancing quality is tested using the mean standardized bias (see, e.g., Caliendo and Kopeinig, 2008) and the t-test, calculated for the unbalanced sample and the balanced sample, whereby in the later the matching weights $\omega(P(X))$ are used to reweigh the characteristic values. As it has been found that the p-values of standard statistical tests are not very reliable in the matched sample (Lee, 2013), it is advised to reduce imbalance as much as possible, i.e., maximizing the minimum p-values over all variables and reducing the standardized bias. Table 2.5 presents the distribution of p-values for the t-test before and after matching, as well as the

[15]The values chosen in the grid search were (0.25, .5, 1, 2, 2.5, 3, 4, 5) and the optimally selected bandwidth value were 2.5 times the maximal bandwidth for women and West Germans, and 4 times the maximal bandwidth for East Germans and men.

average standardized biases. Reweighting the sample with the matching weights results in a substantial reduction in imbalance across treatment groups in all samples; none of the characteristics exhibits a significant difference in characteristics by conventional significance levels.

Following the calculation of $\omega(P(X))$, the hazard rates of treated and controls are estimated on the balanced sample. To estimate the treatment effect on the exit rate from unemployment, the Kaplan-Meier (Kaplan and Meier, 1958) survival functions $\hat{S}(t)$ are estimated for treated and controls separately. Let $h(t) = e(t)/R(t)$ denote the hazard at time t, defined as the fraction of unemployment exits $e(t)$ of all unemployed still at risk $R(t)$. The survival function is given by $\hat{S}(t) = \sum_{i:t_i<t}(1-\frac{e(t)}{R(t)})$, and the treatment effect estimate is calculated as $\Delta^{TT}(t) = \hat{S}_1(t) - \hat{S}_0^{\omega}(t)$. Note that a positive (negative) effect on the exit rate from unemployment is given by a negative (positive) difference in survival rates. As it is more intuitive to think in terms of hazards, we focus on the interpretation of $-\Delta^{TT}(t)$, which is approximately similar to the treatment effect on the cumulative hazard function.[16]

To estimate the channels-specific hazard rates $h_j(t)$, only exits from the same channel $e_j(t)$ are considered and all other exits are taken as censored. As individuals are only at risk provided that they survive all competing risks until t, standard Kaplan-Meier estimates were found to inflate the true exit rates in case of competing risks (Gaynor et al., 1993). We hence focus on cumulative incidence functions, defined as $\hat{F}_j(t) = \sum_{i:t_i<t} \hat{h}_j(t) \cdot \hat{S}(t-1)$, and take the difference $\Delta_j^{TT}(t) = \hat{F}_{1j}1(t) - \hat{F}_{0j}^{\omega}(t)$ as the treatment effect estimates.

To estimate the effect of vacancy information on the expected probability to participate in ALMP, the ATT is now calculated as the difference in average treatment probabilities $Y_D(t)$ at each point in time t, $\Delta^{TT} = E[Y_1(t)] - E[Y_0(t)|P(X)]$.

Similarly, for the estimation of the treatment effect on the employment quality, the matching procedure is repeated for the subgroup of unemployed who exited unemployed during the period of observation. The propensity score uses a similar specification as before, except that we now additionally include an indicator of the duration of the unemployment spell. The imposition of the common support condition and the selection of the optimal bandwidth was also done as before. The

[16]The cumulative hazard is equivalent to the negative log of survival functions, i.e., $\hat{\Lambda(t)} \approx -\log(\hat{S}(t))$, so that the negative difference in survival rates is approximately equal to the effect on the cumulative hazard rate. In the following, we will hence refer more loosely to the effect on "exit rates" when interpreting the effects of $-\Delta^{TT}(t)$.

lower part of Table 2.6 summarizes the balancing quality in the matched sample, and the number of treated individuals deleted by the common support condition. For all effect estimates, the standard errors are obtained via bootstrapping using 300 replications. Based on random draws from the estimation samples, the bootstrap procedure replicates the whole matching procedure, including the estimation of the propensity score. All estimations are conducted separately for the respective labor market groups.

2.6 Results

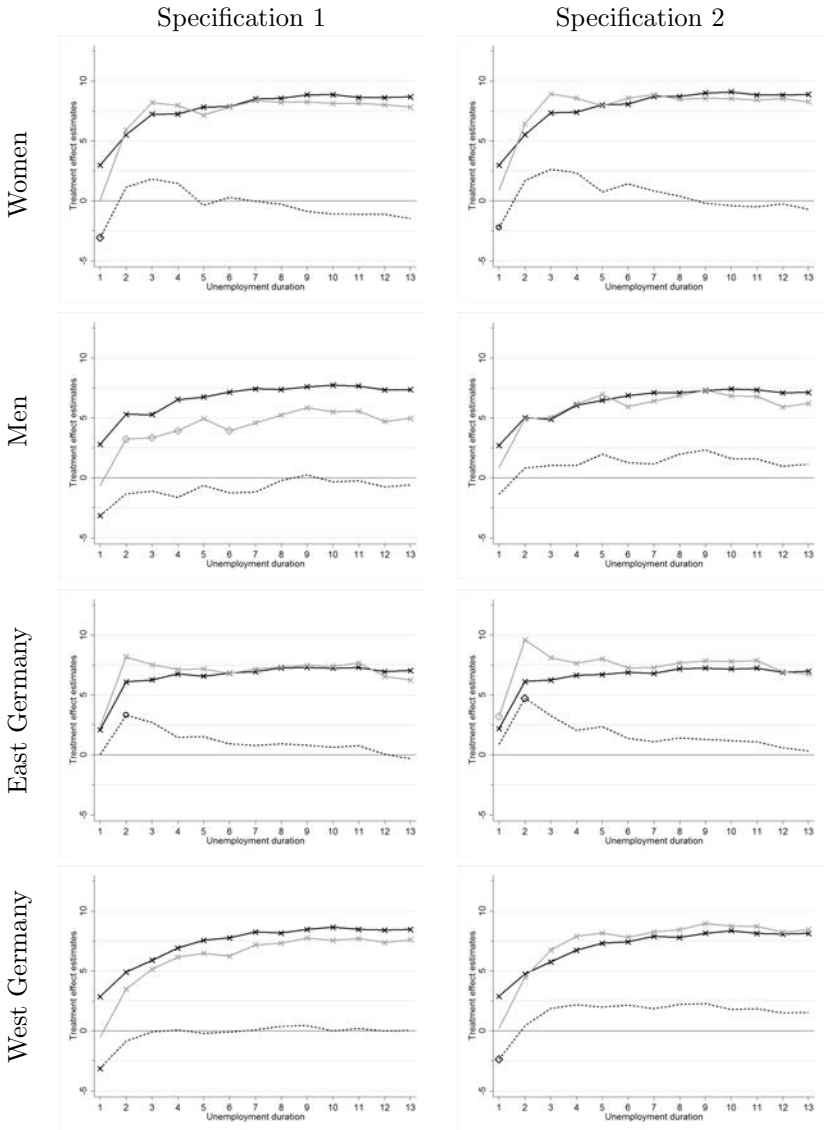
In the following, the effect estimates are presented. We start by outlining the effects for the exit rate into regular employment, and then complement these findings with the results on the ALMP participation rates. Finally, we discuss the effects on employment quality. As outlined in Section 2.5, all estimates are obtained by balancing the relevant control variables across treatment groups, so that their influence on the outcomes of interest is eliminated. To emphasize the impact of controlling for difference in the simultaneous activation strategy, we present the results of two different specifications, Specification 1 does not include information on alternative treatment offers, whereas Specification 2 does.

2.6.1 Exit Rates from Unemployment

Figure 2.1 presents the ATT estimates on the transition rate to employment during the thirteen months following unemployment entry, by specification and labor market group. Next to depicting the effects on the overall transition rates, the effects on the channel-specific exit rate for the PES channel and non-PES channels are shown. As we look at cumulative hazards, the effect of vacancy information on the hazard rate is reflected by the slope of the curves. The level of the curves represent the cumulative probability to have exited unemployment at each point in time.

Focussing on the overall exit rates we find that the receipt of vacancy information has a significant positive effect for all labor market groups. The increase in exit rates is strongest during the first three months in unemployment, afterwards the slope of the hazard function becomes zero and remains zero until the end of the observation period. This suggests that the early receipt of vacancy informa-

Figure 2.1: Overall and channel-specific exit rates from unemployment.



Note: The gray solid line depicts the effect on the negative survival function for all unemployment entrants; the black solid (dashed) lines depict the effect on cumulative incidence functions considering only exits from the PES channels (non-PES channels). Specification 1 does not include information on the alternative activation offers, Specification 2 includes this information. X's indicate significance at the 1%-level, diamonds indicate significance at the 5%-level and O's indicates significance at the 10%-level.

tion creates a head-start in terms of unemployment exits that non-receivers are unable to catch up to even later on. While these patterns are very similar across subgroups and specifications, the magnitude of the effect may vary. For better comparability of the effect estimates across groups, we translate the percentage-point changes observed in the graphs to percentage changes, and focus on the long-term effect sizes observed after 13 months. From the results of Specification 1 we find that women and West Germans experience a 23% increase in the exit

rate after 13 months, men experience a 16% increase, and East Germans a 20% increase in the probability to have exited unemployment. When controlling for activation intensity, the effect estimates are slightly increased. While the increase is negligible for most subgroups, it is quite strong for men, increasing the long-term effect estimate to 19%. The omittance of information on the overall propensity to receive treatment hence results in a downward bias of the effect estimates for men.

The mechanisms by which vacancy information affect unemployment exit rates are explored further by looking at the channel-specific hazard rates. As expected from job search theory, we find that the effect of vacancy information is strongest for the exit rate from the PES channel. The effect patterns over time are very similar to the ones observed on the overall exit rates, with a strong increase during the first three months, after which the effects become gradually lower, and zero towards the end of the observation period. Focussing again on the long-run effects at the end of the observation period in Specification 1, we find that the effect on the PES exit rate is stronger for women than for men, and stronger for West Germans than for East Germans. While the PES exit rate is increased by 160% for women, it is increased by 100% for men; the effect is 150% for West Germans, and 85% for East Germans. In contrast to before, the effect estimates are not changed when controlling for general activation differences.

Assuming that the quality of proposed vacancy information is the same across all subgroups, the differential effectiveness across labor market groups may be explained by differences in the relative effectiveness of the PES channel. As women are likely to experience a lower productivity of non-PES search productivity than men, the additional information from the PES becomes more valuable. In contrast, in East Germany, both the non-PES channels and the PES channel are expected to be less productive, due to a high competition for jobs. As a consequence, it is less likely that vacancy information is turned into a job offer.

Turning to the effects on the exit rate from non-PES channels, and considering Specification 1, we find that the exit rate during the first month is distinctly and sometimes even significantly negative for all labor market groups except East Germans. This confirms previous job search theoretic findings on channel substitution (Fougère et al., 2009), and also seems to confirm the notion that the East German productivity of own search is relatively low. Following this initial dip, the effect increases, but the patterns are somewhat different by subgroup. While the effect becomes zero for men and West Germans, women and East Germans

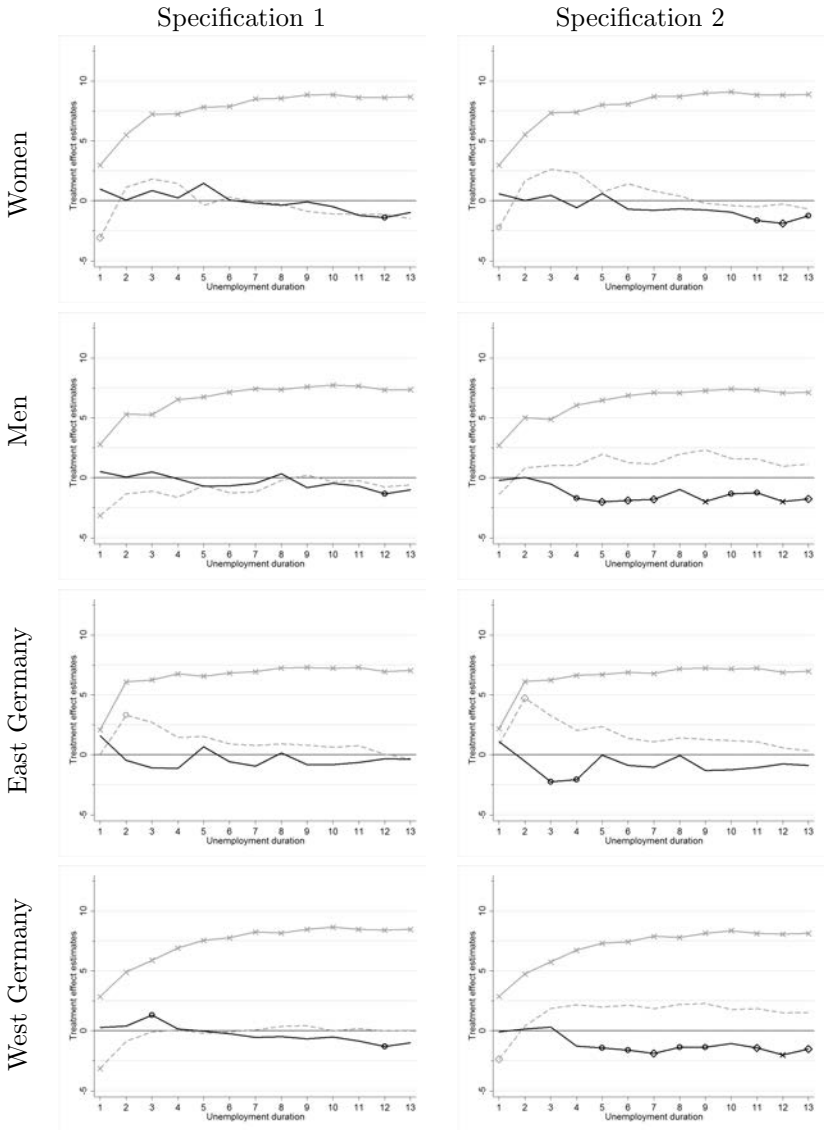
experience a short positive effect on non-PES exit rates during months two and three, before the effect is reduced again and becomes zero for the rest of the observation period. Controlling for differences in overall activation exposure, the effects on non-PES exit rates are shifted upwards in all groups. For women, the upward shift is rather small; the temporary increase in non-PES exit rates is still largest during month two and three at around 9%, but the long-run effects are zero. Similarly, for East Germans, the upward shift in Specification 2 is strongest during months two and three. This raises the size of effect estimates on the initial peak to 24%; the long-term effect however is zero. For men and West Germans, the monthly effect estimates are increased more strongly and more persistently by about 2%-points over whole of the observation period. Overall, this results in a continuously increased exit rate from unemployment after the fourth month in unemployment, with a long-run effect estimate of 2%.

2.6.2 ALMP Participation

The variability of the non-PES exit rates to the inclusion of activation confounders suggests that the observed positive correlation between activation schemes and the receipt of vacancy information also affects the exit rate from unemployment. In the following we asses the effects on the actual participation rate in ALMP. Figure 2.2 depicts the effect estimates on the monthly average participation rate in ALMP programs; for better comparability we also depict the effect estimates on the channel-specific exit rates discussed in the previous Section.

Comparing effect estimates across specifications, we find a substantial sensitivity of ALMP participation to the inclusion of the activation confounders. In contrast to the observed upward shift in non-PES exit rates, we now find that the inclusion of these confounders results in a downward shift of estimates. Whereas the effects on ALMP participation are approximately zero in Specification 1, the effects are significantly reduced in Specification 2. For women, the downward shift is very minor, and we only observe a significant reduction in ALMP participation rates towards the very end of the observation period. Accumulating the effect estimates over the whole duration of the observation period, the cumulative negative effect on ALMP participation is at 7%-points. For men and West Germans, the reduction in ALMP participation rates is stronger and more persistent. While the participation probability remains zero during the first three months in unemployment, it becomes significantly negative from month four onwards. The monthly

Figure 2.2: ALMP participation and channel-specific exit rates.



Note: The black solid lines depict the treatment effect on the participation probability in ALMP. The gray solid (dashed) lines depict the effect the cumulative incidence function considering only exits from the PES channels (all non-PES channels). Specification 1 does not include information on the alternative activation offers, Specification 2 includes this information. The treatment indicator is the receipt of fulltime vacancy information. Standard errors are obtained by bootstrapping with 300 replications. X's indicate significance at the 1%-level, diamonds indicate significance at the 5%-level and O's indicates significance at the 10%-level.

effects oscillate around 20% to 25% in both groups, resulting in a cumulative reduction of the ALMP participation of 14%-points for West Germans and 17%-points for men. For East Germans, the ALMP participation rate is significantly decreased during month two to four after unemployment entry, later on, the effect becomes zero. The magnitude of the effect is fairly strong for East Germans - in the third month after unemployment entry, ALMP participation is reduced by 36%. By

the short duration of this increase, the cumulative reduction is at 10%-points and lower than for West Germans.

Comparing the timing of the decreased participation and the increased exit rates from non-PES channels, it is interesting to see that they follow very similar patterns over time. East Germans experience their strongest reduction in ALMP participation in month three and four, whereas their exit rate from non-PES channels is increased strongest during months two and three. Similarly, for West Germans and males we observe a significant reduction in ALMP participation after the fourth month in unemployment, while the exit rate from non-PES channels is visibly increased during this time. Overall, this suggests that the positive effect on non-PES exit rates is at least partially explained by foregone entry into ALMP, which results in a locking-in of control observations in treatment.

As outlined in the introduction, two different sets of activation strategies might exist. On the one hand, caseworkers may concentrate their activation efforts on particularly disadvantaged subgroups, e.g., unmotivated unemployed, or those facing particularly high barriers to entry, considering high and low intensity activation schemes as substitutes. On the other hand, caseworkers might treat all unemployed similarly, considering higher intensity activation as complementary to a failed initial low intensity activation. Alternatively, the caseworkers might use a mix of the two strategies - concentrating activation on specific subgroups, but consecutively moving from lower intensity activation to more intensive activation as unemployment persists. As we only find a negative relation between employment integration and ALMP participation after conditioning on the initial differences in activation intensity, this seems to imply that caseworkers focus on specific subgroups in the activation. The early timing of the negative effect in East Germany suggests that here, ALMP participation is considered a substitute for the availability of vacancies during the initial integration. In contrast, for men and in West Germany, the receipt of vacancy information does not change the initial probability to participate in ALMP, but negatively affects participation later on, which might naturally arise due to the reduced unemployment level.

2.6.3 Employment Quality

As outlined in Section 2.2, the receipt of vacancy information may affect the quality of accepted employment relationships. The sign of the quality difference is

theoretically ambiguous. A higher information level is expected to increase the reservation value of employment and hence the quality of accepted employment. In contrast, simultaneous monitoring or a low quality of vacancy information may result in a reduction of the employment quality. As we cannot distinguish between a monitored and pure information vacancy information in our analysis, the two countervailing effects may cancel out on average, which has to kept in mind when interpreting the results. Table 2.6 presents the effects of vacancy information on the first accepted employment spell, as before, the effects are presented by labor market subgroup and for the two different propensity sore specifications.

Table 2.6: Successful channel and employment characteristics of first employment spell.

	Women		Men		East Germany		West Germany	
	Spec 1	Spec 2	Spec 1	Spec 2	Spec 1	Spec 2	Spec 1	Spec 2
Found through PES	**0.104**	**0.104**	**0.072**	**0.069**	**0.091**	**0.092**	**0.082**	**0.078**
s.e.	0.020	0.021	0.017	0.018	0.027	0.029	0.015	0.016
t-stat	5.138	4.995	4.228	3.861	3.313	3.169	5.571	4.887
TWA	0.007	0.015	0.025	0.012	0.026	0.02	0.017	0.02
s.e.	0.022	0.023	0.022	0.024	0.027	0.029	0.018	0.019
t-stat	0.309	0.638	1.175	0.485	0.981	0.684	0.956	1.031
Short-term work	0.021	0.024	0.002	-0.004	-0.014	-0.02	0.005	0.009
s.e.	0.031	0.034	0.025	0.026	0.041	0.043	0.023	0.023
t-stat	0.673	0.712	0.079	-0.162	-0.354	-0.457	0.243	0.380
Hourly wage (log)	-0.009	0.002	-0.000	-0.001	-0.013	-0.007	-0.005	-0.001
s.e.	0.019	0.020	0.016	0.018	0.024	0.027	0.013	0.014
t-stat	-0.484	0.119	-0.006	-0.045	-0.514	-0.266	-0.354	-0.066
Hours worked (log)	0.002	0.002	**-0.017**	**-0.018**	-0.025	**-0.031**	-0.004	-0.005
s.e.	0.019	0.021	0.009	0.010	0.017	0.018	0.012	0.012
t-stat	0.102	0.112	-1.798	-1.780	-1.439	-1.687	-0.304	0.395
Common support treated[1]	2	1	0	3	1	0	5	2
Mean SB	1.526	1.886	1.073	1.298	2.374	2.301	1.004	1.136
N	1,321		1,857		999		2,157	

Note: The treatment effect estimates were estimated using kernel matching on the propensity score with an Epanechnikov kernel and optimal bandwidth that was selected to minimize the difference in characteristics in the matched sample. Standard errors are obtained by bootstrapping with 300 replications. Bold number indicate significance at the 10%-level. [1]Number of treated excluded from the estimation due to lacking or low overlap.

Overall, we find very little evidence of a heterogenous quality in the accepted employment relationships. While we find a substantial and significant increase in the probability to exit unemployment via the PES, we do not find a significant impact on any of the outlined quality indicators, except for the number of hours worked, which is reduced by about 2% for men, and about 3% for East Germans. Note, that in combination with a zero effect on the hourly wages earned, this may be interpreted as overall negative impact on the daily wages earned, which

was previously also found by van den Berg et al. (2013). Note also that the effect estimates for working at a PSA are economically significant for both East and West Germans, but lack statistical significance. Overall our findings suggest, that the receipt of early vacancies has a small negative effect on the employment quality.

2.6.4 Alternative Treatment Definition

To assess the sensitivity of our estimates to the definition of the treatment indicator, we redefine the treatment to also include the receipt of part-time vacancy information and jobs in temporary work agencies. This results in a small reallocation from the control to the treatment group and an increase of treatment group sizes between 5% and 13%. While the labor market characteristics of the treatment groups are largely unchanged by this redefinition, we can see a slight widening of the gap in the offered activation measures between treatment groups, that is strongest for East Germans.

The results on the channels-specific transition rates and the ALMP participation are depicted in Figure A2.1 in the Appendix. While the effects are fairly similar across the two specifications, the magnitude of the effects is somewhat increased for some labor market groups. Focussing on Specification 2, we find that the effect on PES exit rates increases to 95% for East Germans at the end of the observation period (relative to previously 85%). For all other subgroups, the exit rates from the PES channel remain the same. As part-time and TWA employment are likely to be considered to be of worse quality than fulltime employment, this suggests, that East Germans have a higher willingness to also accept employment of lower quality - which is in line with the notion that other channels of search are not very productive. With respect to the exit rates from the non-PES channels we find that the effect estimates are reduced for women and East Germans. While the positive peak in early non-PES exit rates completely disappears for women, the effect remains positive but becomes smaller for East Germans. The cumulative negative effect on the ALMP participation is increased for all subgroups except women, and now amounts to 22%-points for men, 16%-points for West Germans and 13%-points for East Germans.

The effect estimates for employment quality are depicted in Table A2.1 in the Appendix, showing that the magnitude of the effects is largely unchanged. In East Germany, the observed negative on hours worked loses statistical significance,

although the level remains at similar in terms of magnitude. We find a small increase in the probability to accept TWA employment for East Germans; however the effects are not significant.

2.6.5 Employment Stability

Our previous analysis on the quality effect of vacancy information suggested a slight reduction in the hours worked, but otherwise did not show a significant deterioration in the quality of accepted jobs. Note, however that the observed quality indicators provide only limited insight into whether the accepted jobs were well-matched with respect to abilities of the unemployed, and working conditions, all of which may influence the expected stability and the duration of the employment spell. A more reliable approach to measure the match quality is the actual observed duration of the employment relationship. As our observation window only covers a rather short period in time, it is difficult to make reliable statements about this, as most spells will be right-censored. Assuming that censoring is random this does not bias our effect estimates, but is likely to result in very noisy results.

Table A2.2 in the Appendix provides tentative evidence on the differential probability to survive in employment until the end of the observation period, for our preferred Specification 2 and for both treatment indicators. As expected, the standard errors are very large, so that none of the effect estimates is significant. Considering only receivers of full-time vacancy information, we find a negative effect in the probability to remain unemployed that ranges between 3%-points and 4%-points for women, East Germans, and West Germans, and is hence rather small in economic terms. For men, in contrast, the effect is close to zero. Considering the extended treatment we find that the effects become less negative, except for men - for whom they become more negative. Note, that van den Berg et al. (2013) also find a small but significant negative effect on employment stability for vacancies received early during the unemployment spell.

2.7 Conclusion

While a number of studies have previously addressed the role of vacancy information in the job search process (Fougère et al., 2009; van den Berg et al., 2013), little

evidence is available regarding their role amongst the overall set of unemployment activation programs. As most countries employ multiple types of high-intensity activation schemes (training programs, job creation schemes, etc.) that are quite costly to maintain, it is important to understand the use and potential of low intensity activation schemes, like job-broking services, in relation to these high intensity programs. Interestingly, the effectiveness of job broking services has not received much attention in the recent policy debate; the registration rates of vacancies are voluntary and are rather low in most countries, and vacancies have the reputation of being negatively selected.

Our study analyzes the effectiveness of early vacancy information from the public employment services (PES) in Germany, taking into account that caseworkers usually have multiple types of activation at their disposal. Exploiting very detailed survey data of unemployment entrants between 2007 and 2008, we can observe the simultaneous offer of vacancy information and offers of more intensive activation schemes, the participation in which would entail a temporary reduction in the employment probability due to locking-in. A descriptive analysis of the relation between the two types of measures shows a strong positive correlation, suggesting that caseworkers use these activation measures interchangeably, rather than focus on either a direct integration into the labor market, or the integration into auxiliary activation schemes.

In our empirical analysis we estimate the effects of vacancy information on the exit rate from unemployment, the probability to participate in ALMP, as well as the quality of subsequent employment. We take account of the positive relation between vacancy receipt and labor market conditions by controlling for an extensive set of individual-specific and local labor market characteristics that are likely to affect both the treatment probability and the probability to exit unemployment. We further account for the heterogenous activation strategies by controlling for the different types of activation offered, and employ a flexible kernel matching approach using the propensity score to estimate the treatment effects. The analysis is conducted separately by gender and for East and West German unemployed.

We show that receiving early vacancy information has a significantly positive effect on the early exit rates from unemployment, which is largely driven by a direct increase in the exit rate from the PES channel. The effect is stronger for women than for men, supporting earlier findings of Fougère et al. (2009), who show that

the effectiveness of vacancy information is higher when the relative productivity of alternative channels is low. We also find that the effect is stronger for West Germans than for East Germans, which can be explained by the relatively higher competition for jobs in the PES channel in East Germany, as the unemployment rate in East Germany is still about twice the size than in West Germany during the period of observation.

We further find that vacancy information from the PES may also increase the exit rate through non-PES channels (i.e., social networks, internet, newspapers, etc.), which could be explained by the simultaneous increase in the monitoring intensity. While we are unable to observe monitoring, we offer an alternative explanation by showing that the timing of the increase in non-PES exit rates coincides with the timing of a significantly reduced participation rate in ALMP, suggesting that non-receivers of vacancy information are locked in ALMP, preventing them from exiting unemployment. East Germans experience a large but temporary positive effect on non-PES exit rates during the early unemployment, the positive effect appears only later for West Germans and for men but persists until the end of the observation period. While this suggests that caseworkers in East Germany resort to early and temporary ALMP measures as substitutes for vacancy information, West German caseworkers use intensive measures as a consequence of a failed early integration into the labor market. The cumulative foregone participation in ALMP after one year is quite substantial and amounts to around 18%-points for men, 11%-points for East Germans and 15%-points for West Germans. For women, the reduction in ALMP participation is at 8%-points and hence not as strong as in the other subgroups. With respect to the quality of the accepted jobs, we observe a small but significant reduction in the weekly hours worked for men and East Germans. While it can be found that a large number of vacancies at the PES are posted by temporary work agencies, we observe an economically significant increase in TWA employment for East Germans that is not statistically significant, however.

Our analysis hence shows that early job broking activities have long-lasting effects on the unemployed activation process by increasing the early matching between workers and vacancies, and by reducing the propensity to enter more intensive and more expensive activation measures later on. Taken at face value, a direct policy conclusion emerging from these results would be to increase the number and quality of vacancy registration and to improve the counseling competence

of caseworkers. As previous studies show, badly matched vacancy referrals often deter companies to register their vacancies at the PES (Engström et al., 2012; Müller et al., 2011). Therefore, an increased connectedness of the PES to local firms and firms with high demand for labor would be an important prerequisite to achieve this (Behncke et al., 2008). Note, that a high registration rate of vacancies may also have direct spill-over effects on the monitoring and counseling quality as caseworkers would be better informed about the overall state of the labor market by observing a higher share of the overall vacancies available.

It is important to keep in mind, however, that our analysis is not fully conclusive about potentially negative effects of vacancy information on the stability of the accepted employment relationships, which could increase the risk of re-entering unemployment and hence the need for ALMP during subsequent spells of unemployment. Clearly, it may be optimal to select training measures over a direct integration if this increases the long-run stability of employment relationships. As descriptive analyses suggest that caseworkers use both measures interchangeably in the activation process, further research is needed to assess which of the two measures is the most efficient and cost-effective in the long-run.

Appendix

A2.1 Tables

Table A2.1: Successful channel and employment characteristics of first employment spell. Extended treatment indicator.

	Women	Men	East Germany	West Germany
Found through PES	**0.103**	**0.06**	**0.086**	**0.073**
s.e.	0.022	0.019	0.026	0.017
t-stat	4.729	3.153	3.263	4.357
TWA	0.007	0.013	0.029	0.015
s.e.	0.023	0.021	0.029	0.018
t-stat	0.311	0.615	0.993	0.841
Short-term work	0.008	0.006	-0.011	0.008
s.e.	0.032	0.025	0.039	0.024
t-stat	0.258	0.247	-0.285	0.329
Hourly wage (log)	-0.004	-0.001	-0.004	-0.007
s.e.	0.019	0.017	0.024	0.015
t-stat	-0.222	-0.036	-0.152	-0.446
Hours worked (log)	0.002	**-0.024**	-0.025	-0.01
s.e.	0.019	0.011	0.018	0.012
t-stat	0.108	-2.224	-1.367	-0.798
N	1,322	1,851	1,004	2,154
Common support treated[1]	13	9	5	6
Mean SB	1.678	1.486	2.019	1.349

Source: IZA Evaluation Dataset S, own calculations.
Note: Treatment is defined as receiving vacancy information of fulltime, part-time and TWA employment. The treatment effect was estimated using kernel matching on the propensity score with an Epanechnikov kernel. Standard errors are bootstrapped using 300 replications. Bold numbers indicate significance at the 10%-level. [1]Number of treated excluded from the estimation due to lacking or low overlap.

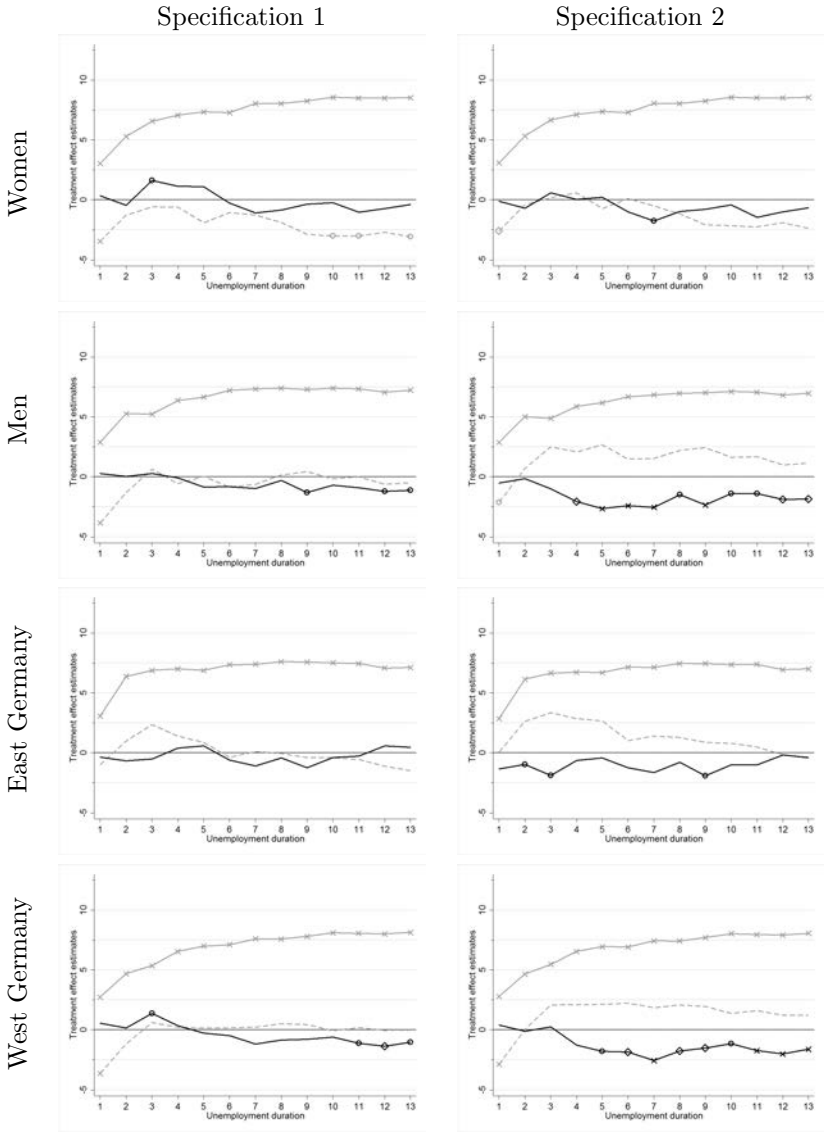
Table A2.2: Stability of first employment spell.

	Women	Men	East Germany	West Germany
Treatment: fulltime vacancies				
Survival probability	-3.69	0.09	-3.61	-3.20
s.e.	4.02	3.87	5.17	3.12
t-stat	-0.92	0.02	-0.70	-1.03
Treatment: fulltime, part-time, TWA vacancies				
Survival probability	-1.95	-0.62	-1.50	-2.57
s.e.	4.61	3.81	5.19	3.09
t-stat	-0.42	-0.16	-0.29	-0.83
N	1,322	1,851	1,004	2,154

Source: IZA Evaluation Dataset S, own calculations.
Note: The treatment effect was estimated using kernel matching on the propensity score with an Epanechnikov kernel. Standard errors are obtained by bootstrapping with 300 replications. Bold numbers indicate significance at the 10%-level.

A2.2 Figures

Figure A2.1: ALMP participation and channel-specific exit rates, extended treatment indicator.



Source: IZA Evaluation Dataset S, own calculations.
Note: The black solid lines depict the treatment effect on the participation probability in ALMP. The gray solid (dashed) lines depict the effect the cumulative incidence function considering only exits from the PES channels (all non-PES channels). Specification 1 does not include information on the alternative activation offers, Specification 2 includes this information. The treatment indicator is the receipt of fulltime vacancy information, part-time vacancy information and TWA information. Standard errors are obtained by bootstrapping with 300 replications. X's indicate significance at the 1%-level, diamonds indicate significance at the 5%-level and O's indicates significance at the 10%-level

Chapter 3

Fighting Youth Unemployment: The Effects of Active Labor Market Policies*

3.1 Introduction

Young individuals entering the labor market are generally considered a population at risk, exhibiting an above-average turnover rate between jobs and an increased probability of entering unemployment. The employment situation of youths[1] is also particularly sensitive to economic fluctuations (Verick, 2011), which was recently demonstrated in the aftermath of the financial crisis. Between 2008 and 2009, youths in the European Union experienced an increase in unemployment rates of about five percentage points to a 20% average, compared to a two percentage-point increase for adults to an average level of 11%.[2]

The prevalent youth-adult unemployment gap can be explained naturally by the initially low search skills and little work experience of labor market entrants, which results in increased levels of turn-over. Although this vulnerability is expected to be only transitory, some youths encounter difficulties during the school-to-work transition process caused by more structural problems. Recent

*This chapter is based on the paper *Fighting Youth Unemployment: The Effects of Active Labor Market Policies*, joint with Marco Caliendo and Steffen Künn (Caliendo et al., 2011).

[1]We follow the general definition of youth as being 25 years or younger.

[2]Based on unemployment rates for youths (aged 15 and 24) and adults (aged 25 and 54) in 2008 and 2009 in the EU-27, from *Eurostat*.

studies on the youth labor market situation in developed countries show that a persistent share of youths experience longer-term unemployment spells, with a strong imbalance towards youths with low educational attainment (Quintini et al., 2007). From an individual and a social perspective, this is a point of concern. Long unemployment spells are found to exhibit "scarring" effects on later labor market outcomes that seem to be more severe for young than for adult workers (compare, e.g., Ellwood, 1983). While the adverse effects on future employment probabilities are particularly persistent for low-educated youths (Burgess et al., 2003), the negative effects on wages seem to persist independently of individual characteristics (Gregg and Tominey, 2005). Potentially driven by foregone work experience or negative signalling, Korpi (1997) and Goldsmith et al. (1997) also show that the unemployment experience is associated with a decrease in subjective well-being and self-esteem, which might lead to a negative effect on current and future employment probabilities. In terms of social costs, there is evidence that rising levels of youths unemployment are not only related to an increase in spending on unemployment benefits and social assistance, but also to the depreciation of human capital, rising crime rates, drug abuse and vandalism (see Bell and Blanchflower, 2010, for an overview).

Against this backdrop, the majority of European countries spends significant resources each year to fight youth unemployment and improve the integration prospects of struggling youths. Active labor market programs (ALMP) are a common tool to achieve these goals. Between 1999 and 2002, countries in the EU-15 spent a yearly average of 1.3 billion euros on ALMP specifically targeted at unemployed youths (OECD, 2004). Although the primary objective of these programs lies in the fast integration in the first labor market, they may also target the continuation or take-up of vocational training for under-educated youths. The types of programs in use are manifold, ranging from targeted measures that account for the specific needs of labor market entrants, to the use of more "standard" ALMP, such as training, wages subsidies or job creation schemes. The prevalence of youth ALMP—introduced during the 1980s and 1990s—has continually increased during the past decade. In 2007 the number of young ALMP participants in the EU-15 amounted to approximately 14% of the youth labor force (between 15 to 24 years). The quantitative importance of ALMP thereby stands in stark contrast to the low level of knowledge regarding their effectiveness. Existing evaluation results of youth ALMP in Europe provide a rather heterogeneous picture of program

benefit[3], suggesting that some of the measures implemented significantly reduce the employment probabilities of youths in the short to medium run. More evidence on the effectiveness of ALMP for youths is hence urgently needed to draw lessons for future policy design. Extrapolating from evaluation results for the adult workforce is misleading, given the distinctive characteristics of young labor market entrants. Moreover, the assessment of long-term effects is particularly important, as ALMP may not affect employment outcomes directly, but through their impact on participation decisions in longer-term education.

Our analysis uses Germany as a case study to contribute to the evaluation literature of youth ALMP in Europe. Due to data restrictions, so far no comprehensive quantitative analysis of the effectiveness of ALMP for youths in Germany was conducted.[4] Our study aims to fill this gap. Even though Germany is considered a role-model of youth labor market integration, with its extensive dual apprenticeship system, a non-negligible share of youths faces structural difficulties of integrating into the labor market. After leaving general education, youths face two stylized barriers: the transition from general education to vocational schooling or training ("first barrier") and the subsequent transition from training to employment ("second barrier").[5] In the late 1990s specific ALMP targeted at unemployed youths were put into place, with measures more suited to accommodate the specific barriers faced by youths. Participation in ALMP has since substantially increased, calling for a thorough assessment of their effectiveness. We analyze the impact of participation in various ALMP in Germany, including job creation schemes, wage subsidies, short- and longer-term vocational training measures, as well as measure aimed at promoting participation in the vocational training system. We use administrative data on an inflow sample of youths into unemployment in 2002, in which we observe participants and non-participants of ALMP for a period of six years, until 2008. The main outcome of interest is the probability to be in regular employment, but we also investigate the effects on participation in further education as an intermediate policy objective. The long observation period allows a meaningful assessment of the short- and long-term program impacts in both cases.

[3]See, e.g., Centeno et al. (2009) for Portugal; Dorsett (2006) for the UK; Larrson (2003) for Sweden; and Brodaty et al. (2001) for France and Caliendo and Schmidl (2011) for a recent overview.

[4]Compare Ehlert et al. (2012) for a recent evaluation of an innovative pilot project that was conducted in three German cities.

[5]See Dietrich (2001) for an in-depth discussion of the barrier-concept.

Exploiting the detailed information on individual pretreatment characteristics we identify the program impact in a quasi-experimental evaluation framework. Based on a justifiable conditional independence assumption, we apply Inverse Probability Weighting (IPW). To account for dynamic treatment assignment and differences in program availability, we estimate the treatment effects separately by elapsed unemployment duration and calendar month of entry into unemployment. We further account for the differential labor market characteristics of East and West Germany, by conducting the analysis separately for the two regions.

The setup of this chapter is as follows. Chapter 3.2 briefly depicts the labor market situation of youths in Germany and the structure of the education system. Chapter 3.3 sets the stage for our evaluation by providing details on the estimation approach, the data used and the programs analyzed. Chapter 3.4 focuses on the implementation of the estimation strategy, and the results are presented in Chapter 3.5. Chapter 3.6 concludes.

3.2 Institutional Background

3.2.1 The German Education System

To set the stage for the following analysis it is helpful to briefly recall the structure of the German education and vocational training system (see Figure 3.1 for an overview).[6] The general secondary schooling system precedes the selection into the vocational training system and has three parallel types of schools: low (*Hauptschule*), medium (*Realschule*) and high (*Gymnasium*) secondary schooling. The vocational training system (*'upper secondary'* and *'tertiary'*) accommodates a variety of pathways that differ with respect to their degree of work–training interaction and their academic content; the higher the academic content, the higher is the required secondary schooling certificate needed to enter. For pupils finishing the lowest type of school the only immediately available vocational training option is the dual apprenticeship, unless they decide to acquire a higher general schooling degree. Pupils who obtain a medium schooling certificate, regularly spent one more year in general schooling and can choose between entering the dual apprenticeship system or full-time vocational schooling, where a state-approved professional de-

[6]Unless otherwise indicated, the following section relies heavily on the official description of the German education system provided by the Kultusministerkonferenz Germany and the EURIDYCE Unit (2009).

gree can be obtained outside the dual system, in a broader range of professions. Finally, pupils who finish the highest schooling type are allowed to participate in any type of vocational education (see shaded areas in Figure 3.1). The shares in Figure 3.1 indicate that medium secondary schooling is by far the most important one in Germany, with an average share of 38% (44%) of graduates in West (East) Germany.[7] It can also be seen that youths in the East have on average a higher level of schooling attainment than their Western counterparts. In both regions a persistent share of 10% leaves lower secondary school with no certificate.

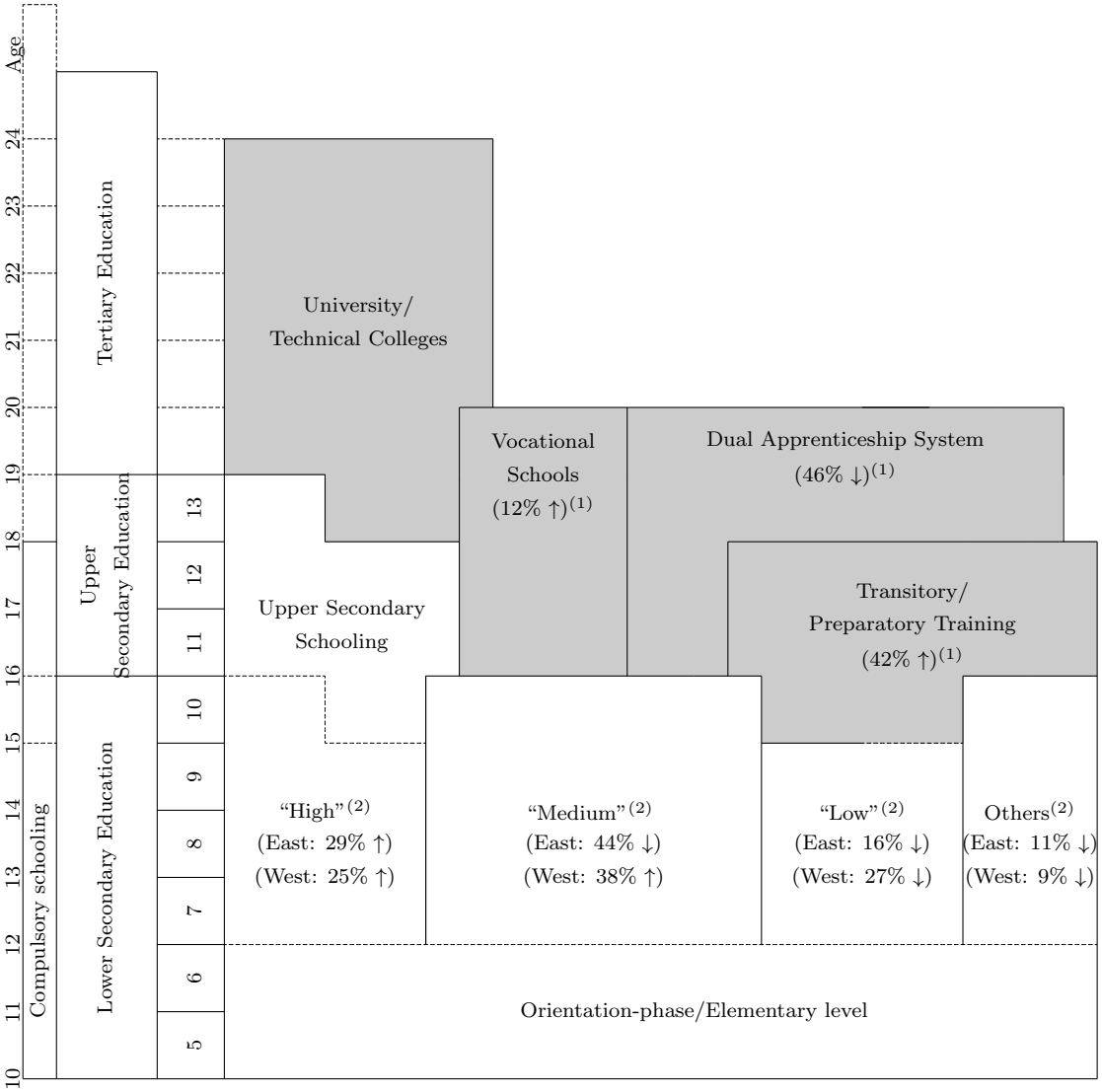
The dual apprenticeship system is the most important option of the vocational training system, accounting for roughly half of all entries each year. The majority (roughly 80% in 2004) of the applicants has a certificate from a low or medium level school (Autorengruppe Bildungsberichterstattung, 2006). Since the demand for apprenticeships mostly exceeded supply in the early 2000s, access to the dual apprenticeship system is competitive and particularly problematic for youths with low previous educational attainment. Given that it is particularly these youths who have only few further options for obtaining vocational education, they are likely to enter unemployment at this "first barrier". At risk of experiencing longer unemployment spells or exiting into inactivity, an extensive preparatory/transitory training system has been put into place aiming to prepare these youths towards a successful entry into the apprenticeship system or other options of the vocational education (see Neumann et al., 2010, for an overview). From 2000 to 2010, participation rates in the preparatory system have increased by about 50% —in years of low demand for apprentices, more youths enter the preparatory system than the apprenticeship system (Bundesministerium für Bildung und Forschung, 2009).

Due to the high labor market orientation of the vocational training system in Germany, the transition from vocational training into employment is generally characterized by relatively low levels of friction—although not all youths manage a smooth transition at this "second barrier". A lack of data that tracks youths after graduation from vocational education makes it difficult to assess the specific unemployment risks youths face after graduation. Reinberg and Hummel (2005) provide general figures for the unemployment risk of youths with different levels of vocational education. They show that individuals with no vocational qualification

[7]Statistics are taken from Bundesministerium für Bildung und Forschung (2009) and the Federal Statistical Office.

are up to three times more likely to be unemployed than youths with qualification—compared to youths with tertiary education they are eight times as likely.

Figure 3.1: The German education system



Source: BIBB 2009, Federal Statistical Office.
Note: Shaded areas denote the vocational part of the education system. (1)Average annual shares of yearly entries into vocational education between 1998 and 2006. (2) Average annual shares of yearly school leavers at the secondary level between 1998 and 2006. Arrows indicate trends in these years.

3.2.2 Youth Unemployment and ALMP in Germany

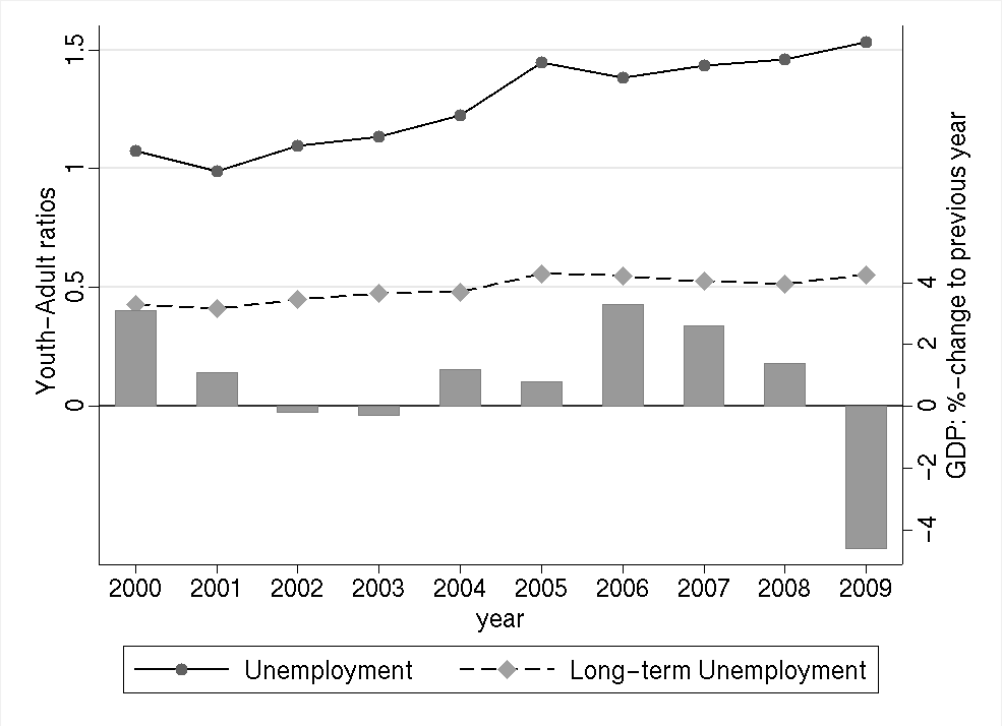
To assess the particularities of the employment situation of youths compared to the general population, it is helpful to relate youth labor market outcomes to the ones of more senior workers. A persistent pattern to be found across all European countries is that youths are usually more likely to enter unemployment than adults,

but that their unemployment spells are more transitory, i.e., they exit unemployment more often than older workers (compare, e.g., Caliendo and Schmidl, 2011, for a recent overview on the employment patterns of youths across the EU-15). Descriptive evidence on the overall economic conditions and the unemployment situation of youths in Germany during the period of our investigation exhibit a similar pattern, as can be seen from Figure 3.2. The youth-adult unemployment ratio gradually increased from almost identical levels in 2000 to about 1.5 in 2009, whereas the long-term unemployment ratio oscillates persistently at around 0.5. Compared to the EU-average, where the unemployment ratio is around 2 to 3, youths in Germany face a comparably low risk of entering unemployment, which is generally attributed to the strong labor market link of the apprenticeship system. In terms of the probability for young people to enter long-term unemployment, however, Germany is amongst the European countries with the highest risk and this is clearly cause for concern.

The rise in the youth-adult unemployment ratio during the observation period can be partially explained by the slowing German economy after 2000, but potentially also by an institutional reform in 2001, reducing the legal restrictions on part-time and fixed-term work. The extensive labor market reforms between 2002 and 2005 (the *Job AQTIV Act* and *Hartz*-reforms) further extended the realm of temporary work arrangements (see Kluve and Augurzky, 2007, for a more detailed description of the *Hartz*-reform changes), thereby leading to a strong increase in youths entering the labor market in "atypical" employment relationships with less stable long-term employment outcomes.

To fight unemployment Germany strongly relies on ALMP. The majority of ALMP schemes are financed by the federal government and the regulations regarding their implementation are contained in the German Social Code III *(SGB III)*. Unemployed youths who fulfill the eligibility criteria, are entitled to participate in the standard ALMP schemes available in the *SGB III*, e.g., training measures, wage subsidies, job creation schemes, etc. As part of the above-mentioned labor market reforms, significant adjustment of the implementation practice of ALMP were made after 2000, with the objective of reaching a faster activation of unemployed individuals. Besides an increase in monitoring efforts, this lead to the expansion of ALMP offering job search assistance and short-term training courses. Furthermore, the *Job-AQTIV Act* of 2002 integrated specialized youths measures within the *SGB III*, that became effective only in 2004. Before the integration

Figure 3.2: Unemployment and long-term unemployment youth-adult ratios, and GDP growth rates in Germany between 2000 and 2009



Source: Federal Statistical Office; Statistics of the Federal Employment Agency

of these measures into the *SGB III*, the only youth-specific ALMP on the federal level existed within the program of *Immediate Action Program for Lowering Youth Unemployment (JUMP)*. *JUMP* was introduced in 1999, following an increasing importance of ALMP in the European and German policy debate as means to deal with the increasing number of youths who were unemployed or unable to find an apprenticeship placement. By providing additional financial means of around one billion euros per year, reducing the eligibility criteria for ALMP participation of unemployed and disadvantaged youths, it was intended to enable a faster integration of youths into ALMP.[8] Furthermore, *JUMP* introduced some new measures specifically aligned to the requirements of unemployed youths, which have later on been partly integrated into the *SGB III*. Originally set up for only one year, *JUMP* was extended and finally expired in 2004 (between July 2003 and December 2004 the program was called *JUMP Plus* intending to support 100,000 long-term unemployed youth).

[8]For a detailed synopsis of the objectives and measures associated to the introduction of *JUMP*, see Bundesministerium für Arbeit und Soziales/Bundesministerium für Bildung und Forschung (1999)

3.2.3 Programs under Consideration

Statistics from the German federal employment agency on the overall numbers of entries into ALMP indicate a substantial increase in participation rates among youths between 2000 and 2010. In 1999 around 600,000 youths were registered in ALMP within *SGB III*—in 2009 the figure was 1.9 million. Between 1999 and 2003, there was on average an extra of 156,000 youths each year entering the programs of *JUMP* (see Dornette and Jacob, 2006, for a detailed participant structure of *JUMP*). Regarding the type of assistance offered, the ALMP in place can be grouped into three broad categories. The most important one in terms of entry numbers are counseling and placement help, with about 60% (50%) of all yearly entries in the *SGB III* in East (West) Germany.[9] Furthermore there are longer-term measures either aiming to promote the integration of youths into an apprenticeship or to help them integrate into the first labor market (training programs, wage and self-employment subsidies, and job creation schemes). Participation in ALMP (compared to the workforce) is generally higher in East Germany, where labor market conditions are less favorable.

In our analysis we assess the impact of seven types of programs, which constitute the most important ones in terms of participation numbers during the period under study (compare Section 3.3.3). Table 3.1 contains a list of the programs, a brief description of their content and their duration. Programs offered both within the regular activation schemes of the *SGB III* and within *JUMP* are grouped together if official implementation guidelines, participant structure and program duration suggested similar content.[10]

Job search measures (JS) include job search monitoring and the assessment of the career opportunities of individuals. Short-term training programs (STT) offer courses of a very short duration to improve auxiliary skills that are important in the application process, e.g. computer classes or language courses. The intended short duration of both programs aims to facilitate job search activities during participation, so that locking-in in these programs is expected to be small.

[9]Shares are provided by the statistical office of the federal labor agency; entries into ALMP between 1999 to 2009 without mobility aid, which technically only includes a cash-transfer to increase the mobility of youths.

[10]The administrative data used contains a very detailed listing of programs, differentiated by content and sources of funding, we aggregate programs with comparable content. In the case where *JUMP* contained a program similar to the regular activation measures, we compared the two measures with respect to their duration, participant structure, etc. and formed a common group only if they did not significantly diverge.

Table 3.1: Description of the programs under scrutiny and sample frequencies.

Abbreviation	Program content and regulatory framework	Participants		Observed duration (months)			
		East	West	East		West	
				50%-ile	90%-ile	50%-ile	90%-ile
JS	*Job Search and Assessement of Employability:* So-called "profiling" immediately after individuals enter unemployment, including professional counseling by the employment agency (EA), short-term measures to improve employability and mobility aid. Conclusion of an informal contract to systematize and monitor search effort, as well as measures to be taken by the EA for a quick and successful re-integration of the unemployed.	1,345 (25.1%)	1,915 (27.3%)	1	2	1	3
STT	*Short-Term Training:* Full- or part-time training measures aimed at improving the employability of youths, including coaching for the application process, and training of specific skills. In the *SGB III* the former should have a maximum duration of two weeks, the latter of eight weeks. *JUMP* measures are not considered.	979 (18.3%)	1,885 (26.8%)	2	4	2	6
JWS	*JUMP Wage Subsidies:* Wage subsidy to regular employment with minimum 15 hours per day at the maximum amount of 60% (40%) of the full wage, for a maximum duration of one (two) years. No minimum duration in unemployment necessary. Post-subsidy employment of half the subsidized period.	991 (18.5%)	628 (8.9%)	12	21	6	13
WS	*SGB III Wage Subsidies:* Wage subsidy to regular employment at the maximum amount of 50% of the full wage, for a maximum of one year. No minimum duration in unemployment. Post-subsidy employment of the same duration as the subsidized period, but a maximum of 12 months.	439 (8.2%)	502 (7.1%)	6	13	4	11
JCS	*Job Creation Schemes:* Working opportunity in areas of the public interest, e.g. infrastructure, social work. Low level of remuneration subsidized by the EA. In the *SGB III* the maximum duration of 12 months could be extended if it leads to regular employment. Very similar program within *JUMP*, here placement subordinate to placement in training or regular employment—parallel qualification measures should be implemented, but could be suppressed if they do not seem sensible.	680 (12.7%)	570 (8.1%)	7	12	7	12
FT	*Further Training Measures:* Long-term training measures for youths with or without professional degree, providing them with job-specific skills. Intensity of training was normalized to 25 to 35 hours per week. The total duration of the measures should not exceed one third of the regular vocational training, i.e. approximately one year, but could be extended if necessary. *JUMP* measures are not considered.	409 (7.6%)	515 (7.3%)	6	12	5	11
PT	*Preparatory Training:* Practical training/internship within a company that should help find and successfully participate in regular vocational training. Duration of training could vary within the *SGB III*. Within *JUMP* it was limited to one year, and potentially also included catching up on the lower secondary schooling degree.	510 (9.5%)	1,012 (14.4%)	7	12	6	12
Total		5,353	7,027				

Note: Program description based on policy guidelines for *JUMP* and the legal text of the *SGB III* in place in 2002. Calculations based on the estimation sample.

However, due to this short duration JS and STT measures are not suited to reduce structural deficits of labor market entrants. Often used as device to assess the employability of youths, it is particularly likely that youths participate in further ALMP subsequent to participation in JS or STT. As sequential program participation renders causal estimation of the impact of short-term programs more difficult, we address this issue in Section 3.5.3 specifically.

Job creation schemes (JCS) and further training (FT) are longer-term measures with a median duration of five to seven months, aimed at overcoming more structural problems of integration in the labor market. JCS are predominantly practically oriented, providing some type of work experience for youths with very little previous labor market experience and potentially low labor market attachment. As participants receive low levels of remuneration during program participation, locking-in in these programs is expected to be high for youths with few outside options. In contrast, FT measures are predominantly focused on youths with vocational qualification, who seem to require additional qualification to succeed in the labor market. The program usually comprises classroom training and may vary between part- or full-time courses.

In contrast to these supply-oriented measures, the wage subsidies offered within the *SGB III* (WS) and *JUMP* (JWS), are aimed to overcome demand side restrictions. The two programs differ with respect to the size of the subsidy and the time period for which it is granted. While the subsidy in WS was regularly limited to one year and provided 50% of the monthly wage, JWS could either be taken up for one year and 60% replacement, or two years and 40% of replacement; employers had to guarantee a period of post-subsidy employment which was equivalent to the subsidized period for WS and half the subsidized period for JWS.

Preparatory practical training measures (PT) aim to enhance the chances of youths struggling at the "first barrier", i.e., at entering the vocational training system. The program consists in a subsidized internship within a firm where predominantly basic practical skills and literacy are conveyed. Some employers might also use this as a probation period before offering a full apprenticeship position within the firm.

3.3 Estimation Strategy and Data

3.3.1 Identification of Causal Effects

We base our analysis on the potential outcome framework (Roy, 1951; Rubin, 1974) where D denotes the treatment indicator, Y^1 the potential outcome in the case of treatment ($D = 1$) and Y^0 the outcome without treatment ($D = 0$). The observed outcome for each individual i is given by $Y_i = Y_i^1 \cdot D_i + (1 - D_i) \cdot Y_i^0$. Our aim is to estimate the average treatment effect on the treated (ATT) that is formally given by $\tau = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1)$. As we are faced with the fundamental evaluation problem of not observing each individual simultaneously in the both the treatment and the non-treatment state, we need a meaningful substitute for the counterfactual (the second term on the right hand side). Approximation by the observed average non-treatment outcome of the non-treated, i.e., $E(Y^0 \mid D = 0)$ does generally not lead to a meaningful estimate, as participants and non-participants are likely to be (self-)selected groups with differential outcomes even in absence of the program. In the absence of random treatment assignment selection into the treatment is assumed to occur systematically based on observable or unobservable characteristics (or both).[11]

In the case where the participation decision depends on observable characteristics W only, we can estimate the ATT by conditioning on these variables, rendering the counterfactual outcome independent of treatment, i.e., $Y^0 \amalg D|W$ (*conditional independence assumption*, CIA). Rosenbaum and Rubin (1983b) show that instead of conditioning on a potentially extensive set of characteristics W directly, conditioning on the probability of treatment participation $P(D = 1|W)$ (the propensity score) suffices to achieve balance between treatment and control group. To ensure that we can find an adequate counterfactual for each treated individual it is furthermore required that the covariates influencing assignment and outcome do not deterministically predict treatment participation, i.e. that $Pr(D = 1 \mid W) < 1$ holds for all W (*weak overlap*). Furthermore, it is required that general equilibrium effects do not occur, e.g., the treatment participation of one individual can not have an impact on the outcomes of other individuals, independent of their treatment participation (*stable unit treatment value assumption*, SUTVA). The validity of this assumption is likely to depend on the scope of the

[11] See, e.g., Caliendo and Hujer (2006) for further discussion.

program as well as size of the resulting effects (Imbens and Wooldridge, 2009). As on average only 12% of the active youth population in Germany participated in ALMP from 2000 to 2007, the scope for general equilibrium effects seems rather limited in our case, so that we expect the SUTVA to hold.

The validity of the CIA is more difficult to justify, as it requires that all relevant variables that simultaneously influence participation and outcome can be controlled for (compare, e.g., Smith and Todd, 2005a or Sianesi, 2004). The availability of informative data is therefore crucial. Although there is no common rule on the particular set of information necessary, the ALMP evaluation literature provides helpful guidance on the question which variables to include. Lechner and Wunsch (2013) argue that more information lowers the bias, and highlight the importance of information on labor market history, caseworker assessments, job search effort, timing of unemployment and program start, health indicators, characteristics of last employer and regional characteristics. As our data is based on detailed administrative records, we are able to reproduce the set of variables suggested by Lechner and Wunsch (2013) to a very large extent (see Table 3.4). When dealing with youths, however, the importance of, e.g., observing past labor market histories to capture relevant but potentially unobservable selection variables (motivation, labor market skills, regional particularities, etc.) is likely to lose substantial power as labor market biographies do not yet exist, or are only limited. Hence, besides including labor market histories for those youth who have already labor market experience (employment and earnings, unemployment, inactivity and treatment participation during the three years prior to unemployment entry), we also include further productivity signals which are likely to justify the CIA. Specifically, we rely on information from the caseworkers (number of placements offers and last contact to labor agency before current unemployment spell) which show to be powerful predictors of treatment assignment. This is not surprising as the caseworkers perception on the labor market performance of unemployed is likely to be more important for the participation decision of low experience youths than for adults. Provided with this additional strong signal of unobserved ability of young unemployed, we argue that the CIA is a reasonable identification strategy in our context.

3.3.2 Definition of Treatment and Control Group

To estimate causal effects in the potential outcome framework, the definition of the treatment status requires clarification. Our question of interest is whether participation in an ALMP program has an impact on labor market outcomes of youths, in contrast to a situation where the program had not been available. In our setting, all unemployed youths are potentially eligible to participate in a program—and they may do so at different points in time—which complicates the straightforward definition of a group of participants and non-participants. As pointed out by Sianesi (2004), defining a treatment group by conditioning on ever observing individuals in treatment simultaneously restricts the control group to individuals who have successfully exited into employment before they could participate in a program, which would introduce bias in the effect estimates.

In the evaluation literature two streams exist to deal with this issue, a "static" and a "dynamic" approach. The dynamic approach makes no direct assumptions about the occurrence of the treatment but considers the timing of treatment as a stochastic process.[12] For the definition of the two groups this means that the distinction between treated and controls is made recurrently at each point in time, based on the observed state of all eligible individuals, and is therefore independent of any treatment status at a later point. Although this selection mechanism is realistic in our setting, the approach has the disadvantage of limited interpretability of the estimates. As the control group includes future program participants, the estimated effects have to be seen as a mixture of "participation vs. non-participation" and "participation now vs. participation later" (see Lechner et al., 2011). In the case of multiple available programs the estimated effects additionally include a relative effect compared to participation in a different program. The static approach on the other hand considers participation vs. non-participation in a particular program based on observing individuals up to a pre-determined point in time and thereby requires conditioning on future outcomes for the non-treatment group (Lechner et al., 2011). The interpretation of the estimated effects is more obvious as only "never-treated" (within a certain time period) non-participants contribute to the counterfactual outcome. As pointed out, the restriction on future outcomes is likely to create a control group consisting of a positively selected subgroup of

[12]See Abbring and van den Berg (2003, 2004) for a discussion in a duration model framework and Fredriksson and Johansson (2008); Sianesi (2004) for an application of semi-parametric matching.

all eligible unemployed and might therefore bias the results downwards.[13]

As we are interested in the effect of participation vs. non-participation, and given the variety of ALMP offered in Germany which render relative effects rather untransparent, we follow Lechner et al. (2011) and apply the static evaluation approach.[14] To do so we have to define a cut-off in unemployment duration at which individuals are assigned to the treatment group (if they participate before the cut-off) and control group (if not). The choice of the cut-off should balance two opposing influences. On the one hand, the estimation bias due to the restriction on future outcomes is increasing with the time window (Fredriksson and Johansson, 2003); on the other hand, a small entry window increases the variance of the estimates due to lower observation numbers, and might also reduce the external validity of the results due to potential seasonal effects. Therefore, we decide to specify the first 12 months of unemployment as our entry window. First, this is not too restrictive on control outcomes since 50% (40%) of non-treated individuals in East (West) Germany are still unemployed after 12 months. Second, it secures a sufficient number of treated observations and reduces the influence of seasonal effects as it captures the complete year.[15] Hence, we assign youths to the group of participants if they enter an ALMP program under consideration (see Table 3.1) within the first 12 months of their unemployment spell and to the group of controls if not. Note, that we discard individuals who participate in any other program within the first 12 months. When individuals participate in multiple programs during their unemployment spell, we focus on the first one in the main analysis.[16]

[13]Lechner et al. (2011) argue that this argument would even strengthen the effectiveness of programs in the case of positive results.

[14]We test the sensitivity of our results with respect to the choice of the evaluation approach and provide results using the dynamic approach in Section 3.5.3.

[15]The dynamic changes in the selection process due to the changes in the composition of unemployed, and potential changes in the types of programs offered during this time period are controlled for in the estimation process (see Section 3.4.2).

[16]About 50% (33%) of treated in the East (West) participated in multiple programs during their unemployment spell, with about 10% (5%) participating in further ALMP within the first 12 months. However, we focus on the first program as subsequent program participation has to be considered as the outcome of the first treatment.

3.3.3 Data and Descriptives

To assess the impact of program participation on labor market outcomes, we use data from the administrative part of the *IZA Evaluation Dataset S*.[17] It is based on the *Integrated Employment Biographies* (IEB) by the Institute for Employment Research (IAB) and consists of a random draw of unemployment entries between 2001 and 2008. It combines different administrative data sources, i.e., the Employment History, Benefit Recipient History, Training Participant History and Job Search History, and contains detailed daily information on spells in employment subject to social security contribution, unemployment, and participation in ALMP.[18] Linked to the information on the respective labor market status, the data include information on income from wages and benefits, on the previous labor market history and socio-economic characteristics of individuals.

We restrict our estimation sample to unemployment inflows in 2002.[19] This guarantees a sufficiently large observation window (at least 72 months after entry into unemployment) and allows us to obtain long-term impact estimates even for the longer running programs. Our choice of the year 2002 also takes account of the adoption of the *JobAqtiv Act* in the beginning of 2002, which entailed significant changes in the strategy of unemployment activation and implementation practice. Besides avoiding potential structural breaks in the implementation of programs between 2001 and 2002, the evaluation results for the programs under the new "regime" are also more relevant for current policy discussion, as their set-up resembles much more the set-up of programs in place today. Based on our initial inflow sample into unemployment in 2002, we only keep youths (aged 25 or younger) and apply several further sample selection criteria which are summarized in detail in Table A3.1 in the Appendix. We end up with an estimation sample of 51,019 unemployment entrants, corresponding to 17,515 youths from East and 33,504 youths from West Germany. Applying the definition of treatment status as discussed above, we identify 5,353 (7,027) youths in the East (West) participating in one of the programs under scrutiny within the first 12 months of unemployment. By restricting treatment to those ALMP entries in the first 12 months after unemployment entry, we capture about 62% (65%) of all individuals

[17] For a detailed description of the *IZA Evaluation Dataset S* see Caliendo et al. (2011).

[18] This does not include information about self-employment, civil servants or inactivity.

[19] Where we observe multiple entries into unemployment per individual, one spell is randomly drawn.

who enter one of the programs in our total observation period of 72 months in the East (West). Non-participants are defined as individuals who do not participate in any ALMP within the first 12 months of unemployment but who are potentially treated later in months 13-72, which is relevant for approximately 27% (14%) of non-participants in the East (West). Since the administrative data record only specific labor market states, we have missing observations for spells of schooling and education, military service, self-employment or inactivity. Some of these states are particularly likely to occur for young individuals. To overcome this problem we apply an imputation method that relies on information for the planned activity in the subsequent spell, and the last activity before unemployment entry recorded for each unemployment spell. By this procedure we are able to fill 92% of all missing monthly information, decreasing the share of monthly missings from initially 25.7% to 2.1%. Inspection of the type of information filled further reveals that non-randomly missing information does not pose a problem in our analysis (see Appendix A3.2 for details).

Table 3.1 provides the number of observations for each of the programs under investigation and moments of the distribution of program duration. As expected we find that the majority of our participants enter short-term measures, i.e., job search (JS) and short-term training measures (STT). Together, they account for almost half of participants in East and West Germany. This is naturally explained by our definition of treatment, as we focus on the first treatment after unemployment entry. Wage subsidies (WS) constitute the second most important types of measures. While WS are equally important in terms of participation shares in East and West, JWS are taken up twice as frequently in the East than in the West and have a longer duration. Furthermore JCS measures are used more extensively in East than in West Germany, potentially reflecting the lack of employment opportunities for low-educated youths in the East. Finally we find that PT are used in the West more often than in the East, with 14% of youths in the West and 10% of ALMP participants in the East.

Table 3.2 presents selected descriptive statistics of the program participants in East and West Germany (measured on entering unemployment). About two third of program participants are male and the majority of youths is older than 20 years. Migrant participation rates reflect the strong migrant populations differences between East and West Germany with 3% (12%) of participants having a migration background in the East (West). Further differences across East and

Table 3.2: Selected descriptive statistics of participants and non-participants

East Germany								
	JS	STT	JWS	WS	JSC	FT	PT	NP
Gender (Female)	0.37	0.42	0.42	0.40	0.30	0.31	0.41	0.39
Age (above 20 years)	0.72	0.73	0.71	0.72	0.66	0.77	0.27	0.56
Migration status	0.02	0.09	0.01	0.04	0.03	0.02	0.07	0.05
Having children	0.05	0.08	0.05	0.05	0.08	0.07	0.07	0.05
Health restrictions	0.10	0.08	0.04	0.06	0.17	0.05	0.09	0.07
School leaving certificate								
None	0.06	0.05	0.01	0.03	0.14	0.05	0.19	0.07
Lower secondary school	0.37	0.30	0.23	0.26	0.47	0.30	0.44	0.25
Middle secondary school	0.52	0.56	0.65	0.59	0.36	0.58	0.32	0.44
Upper/specialized secondary School	0.06	0.08	0.10	0.11	0.03	0.07	0.06	0.24
Professional training								
None	0.23	0.29	0.13	0.22	0.47	0.17	0.89	0.52
Apprenticeship/university	0.77	0.71	0.87	0.78	0.53	0.83	0.11	0.48
During the last three years before unemployment entry, months spent in ...								
regular employment	18.26	15.05	21.06	18.84	13.00	16.44	3.69	11.69
ALMP	4.42	4.41	3.24	3.64	5.17	5.10	3.47	2.71
inactivity	8.02	11.64	7.70	8.72	11.38	9.18	24.06	16.54
unemployment	5.76	5.24	4.12	4.98	6.85	6.18	3.99	3.99
Last activity before entry into unemployment								
Regular employment	0.63	0.59	0.62	0.56	0.58	0.64	0.40	0.54
Education, training, never employed	0.28	0.34	0.34	0.36	0.31	0.28	0.40	0.36
Other	0.08	0.07	0.05	0.08	0.12	0.07	0.20	0.10
Number of placement propositions	4.77	3.95	3.27	3.24	3.45	3.93	0.93	1.89
West Germany								
	JS	STT	JWS	WS	JSC	FT	PT	NP
Gender (Female)	0.36	0.38	0.34	0.36	0.30	0.33	0.38	0.39
Age (above 20 years)	0.72	0.70	0.73	0.77	0.48	0.81	0.29	0.61
Migration status	0.16	0.27	0.13	0.19	0.17	0.16	0.19	0.16
Having children	0.05	0.06	0.05	0.06	0.03	0.06	0.03	0.05
Health restrictions	0.08	0.07	0.06	0.07	0.08	0.08	0.06	0.07
School leaving certificate								
None	0.13	0.12	0.09	0.10	0.31	0.10	0.23	0.13
Lower secondary school	0.50	0.49	0.50	0.52	0.55	0.50	0.52	0.44
Middle secondary school	0.29	0.31	0.33	0.29	0.12	0.34	0.21	0.28
Upper/specialized secondary School	0.08	0.07	0.08	0.09	0.02	0.06	0.04	0.15
Professional training								
None	0.46	0.48	0.35	0.40	0.85	0.36	0.93	0.55
Apprenticeship/University	0.54	0.52	0.65	0.60	0.15	0.64	0.07	0.45
During the last three years before unemployment entry, months spent in ...								
regular employment	18.94	16.77	19.73	18.95	8.92	20.10	5.71	14.81
ALMP	2.72	2.15	2.74	2.34	3.10	2.31	3.14	1.77
unemployment	4.35	3.71	4.28	4.61	4.98	4.35	3.20	3.24
inactivity	9.50	12.89	8.70	9.67	17.78	9.08	21.46	14.68
Last activity before entry into unemployment								
Regular employment	0.70	0.65	0.74	0.69	0.59	0.74	0.48	0.63
Education, training, never employed	0.21	0.27	0.18	0.23	0.24	0.19	0.38	0.26
Other	0.09	0.08	0.07	0.08	0.17	0.07	0.15	0.11
Number of placement propositions	4.22	3.85	4.27	4.56	3.02	3.58	1.47	2.32

Note: Characteristics are measured at point of entry into unemployment. Numbers are shares unless indicated otherwise.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

West emerge in terms of the pretreatment educational attainment. While the average program participant in the East has acquired a middle secondary school certificate, their counterpart in the West has a lower secondary school certificate. Furthermore, about 75% of youths in the East have already received some type of apprenticeship training compared to only about 50% in the West. In line with the observed differences in program importance this underscores that youths in the West seem to require help at overcoming supply-sided restrictions caused by their insufficient level of educational attainment, while unemployed youths in the East are rather held back by the low labor demand. For example, the importance of measures to overcome the "first barrier" in the West can be explained by the low schooling levels of West German youths.

The comparison of participant characteristics across program types shows a clear divide in terms of labor market attachment. The labor market histories during the three years preceding unemployment entry show that youths in either type of wage subsidies (WS and JWS), longer-term training measures (FT) and job search assistance (JS) have spent more time in (full-time) employment and less time in inactivity (e.g. schooling) than participants in other programs and non-participants. Although they have spent a comparable amount of time in unemployment, they are also slightly older, have received a larger number of placement offers during their current unemployment spell, and in the East they are also better educated than the rest. The higher relative labor market attachment of program participants compared to to non-participants is somewhat suggestive of "cream-skimming" or at least a positive selection into these program based on these observed characteristics.

Individuals with adverse labor market prospects seem to be concentrated in JCS and PT programs. Given the differential objective of PT measures, the adverse characteristics (e.g., they are on average younger, did not obtain a school leaving certificate, and have received significantly fewer placement offers) of participants in PT are not surprising. The characteristics of JCS participants are similarly adverse, suggesting that it is also the low educational attainment that keeps them from integrating into the first labor market. Furthermore JCS participants are older and exhibit above average shares of youths with health restrictions in the East—suggesting that these youth face more structural difficulties of integrating in the labor market than the other program participants. Note, that the programs' objective (compare Section 3.2.3) is the provision of work experience

but not the increase in educational attainment. The first descriptive assessment of program characteristics hence suggests that placement in JCS is not primarily seen as stepping stone to further employment, but more as last resort for keeping these youths in the labor force.

3.4 Empirical Implementation

3.4.1 Inverse Probability Weighting

Based on the assumptions outlined in Section 3.3.1, the treatment and control group can be made comparable by conditioning on the propensity score (PS), i.e., $E(Y^0 \mid D = 1, P(W)) = E(Y^0 \mid D = 0, P(W))$, which then identifies the average treatment effect on the treated τ. Based on the PS, different approaches have been suggested to estimate an adequate counterfactual outcome, where the predominately used methods are semi-parametric matching or reweighting (see, e.g., Imbens, 2004). The most suitable method has to be chosen depending on the study context. Given our large set of covariates and the relatively homogenous groups of treated and controls we apply inverse probability weighting (IPW) (Imbens, 2000, 2004). The IPW estimator has preferable finite sample properties compared to different matching algorithms under the requirement that the propensity scores are estimated and the weights are normalized to one (shown by Busso et al., 2014b, in a Monte Carlo study). Huber et al. (2010) also show that IPW performs well under extensive variation of the data set-up, although it is outperformed by some advanced matching estimators. Given the major advantage of a lower computational burden during the bootstrapping procedure for the estimation of standard errors IPW seems to be an appropriate choice in our setting.

The idea of IPW is to adjust the outcomes of the non-treated by weighting them with the inverse of the estimated propensity scores $\hat{P}(W)$. An estimate of the parameter of interest τ^{IPW} is then obtained as the difference between the average outcome of the treated and the reweighted average outcome of the non-treated:

$$\tau^{IPW} = \left[\frac{1}{N^1}\sum_{i\in I^1} Y_i\right] - \left[\sum_{i\in I^0} \frac{Y_i \hat{P}(W_i)}{1-\hat{P}(W_i)} \Big/ \sum_{i\in I^0} \frac{\hat{P}(W_i)}{1-\hat{P}(W_i)}\right] \tag{3.1}$$

where $\hat{P}(W_i)$ is the estimated propensity score and the division of the counterfactual outcome by $\sum_{i\in I^0} \frac{\hat{P}(W_i)}{1-\hat{P}(W_i)}$ ensures that the weights add up to one (see Imbens,

2004). One concern associated with IPW is that it is particularly sensitive to large values of the propensity scores as they receive disproportionately large weights in the construction of the counterfactual (see Frölich, 2004). However, the relevance of this problem decreases with sample size as each observation has asymptotically less influence on the estimate (Huber et al., 2010). As we have a large number of non-treated observation of our disposal resulting in a average treated-control ratio of approximately 1 to 20, this issue is less of a concern in our application. To further reduce this problem, we apply a rather restrictive common support condition (see Section 3.4.3). In addition we test the sensitivity of our results with respect to this potential outliers in Chapter 3.5.3 by trimming the distribution of the propensity scores of the non-treated.

3.4.2 Perfect Alignment of Treatment and Control Groups

As pointed out by the previous literature, participant characteristics and the type of treatment received may vary with the timing of entry into a program (compare, e.g. Sianesi, 2004 and Fitzenberger and Speckesser, 2007). As we define treatment over a period of 12 months after entry into unemployment we need to take account of potential dynamics in the selection into treatment or out of unemployment during this period. To mimic the selection process up to a particular point in time only individuals with similar unemployment durations should be compared. Given the small number of monthly treatment entries in our sample, estimation of the propensity score within monthly cells is not feasible. Instead we adopt an approach suggested by Fitzenberger and Speckesser (2007), consisting of stratified estimation of the PS within larger time windows combined with a "perfect" (i.e. monthly) alignment of treated and controls for the estimation of the treatment effect.

For the estimation of the PS we stratify the sample of treated into three subgroups based on their elapsed unemployment duration until treatment entry: (1) one to three months of unemployment duration, (2) four to six months and (3) six to twelve months. The treatment group in the respective cells hence consists of all individuals receiving treatment within these months of their unemployment spell. The control group consists of youths who are still unemployed in the first months of the respective stratum and who are not treated in the first 12 months of their unemployment spell. Based on the estimated propensity score, weighting of the controls is done within the "alignment cells". Besides aligning individuals

perfectly on the month of entry into the program, we further take account of seasonal labor market conditions and program variability across calender time (see Sianesi, 2004), by aligning individuals perfectly by calender month of entry into unemployment.[20] The construction of counterfactual is hence done within *monthly* cells of both the unemployment entry and unemployment duration, whereby only controls receive weights that were unemployed at least until the month of program entry of the treated. The resulting estimator can be written as:

$$\tau^{IPW} = \frac{1}{N^1} \sum_{c=1}^{12} \sum_{p=1}^{12} \tau_{cp}^{IPW} \cdot N_{cp}^1 \tag{3.2}$$

where τ_{cp}^{IPW} is then estimated in each cell following Equation (3.1). N^1 denotes the total number of treated and N_{cp}^1 the number of treated in each cell defined by calendar month of unemployment entry c and the months in unemployment before treatment entry p. As the estimation of treatment effects within each cell yields 144 single effects τ_{cp}^{IPW}, with c denoting calendar month of entry into unemployment and p the month of entry into treatment, we aggregate the single effects to τ^{IPW}.[21] The aggregation is obtained by creating a weighted average of the monthly effects, with weights being determined by the distribution of monthly program starts and monthly unemployment entries among participants. See A3.3 in the Appendix for a more detailed description of perfect alignment.

3.4.3 Propensity Score Estimation and Implementation

Table 3.3 provides the number of observations for each of the three subgroups of treatment entry. It can be seen that treatment participation is strongly concentrated on the first quarter of unemployment duration—except for the case of JCS in the East, where youths are most likely to enter after six months in unemployment. It can also be seen that controls are highly likely to exit unemployment during the first quarter of their unemployment spell. In particular, we see a reduction of the control sample for about one quarter (one third) in the East (West) during the first three months in unemployment. Despite the reduction in sample sizes with increasing unemployment duration, each time window contains a sufficient number

[20]Note, that the propensity score specification includes indicators for the calendar month of unemployment entry

[21]Note that while treated are assigned to mutually exclusive cells defined by c_1 and p_1, they are opposed to non-treated with the same entry into unemployment $c_1 = c_0$ but $p_1 \leq p_0$.

of treated and controls to obtain a meaningful estimate of the propensity score.

Table 3.3: Timing of (potential) entry into treatment, for participants and non-participants

East Germany								
Entry	JS	STT	JWS	WS	JCS	FT	PT	NP
1 – 3 months N	758	516	609	299	202	181	257	12,119
%	56.36	52.71	61.45	68.11	29.71	44.25	50.39	100.00
4 – 6 months N	256	228	195	75	156	136	127	9,304
%	19.03	23.29	19.68	17.08	22.94	33.25	24.90	76.77
7 – 12 months N	331	235	187	65	322	92	126	8,444
%	24.61	24.00	18.87	14.81	47.35	22.49	24.71	69.68
Total	1,345	979	991	439	680	409	510	
West Germany								
1 – 3 months N	1,059	1,049	311	322	283	289	588	26,410
%	55.30	55.65	49.52	64.14	49.65	56.12	58.10	100
4 – 6 months N	438	429	177	115	121	123	230	17,561
%	22.87	22.76	28.18	22.91	21.23	23.88	22.73	66.49
7 – 12 months N	418	407	140	65	166	103	194	14,874
%	21.83	21.59	22.29	12.95	29.12	20.00	19.17	56.32
Total	1,915	1,885	628	502	570	515	1,012	

Note: Calculations are based on the estimation sample. Non-participants are considered controls in the respective time window if they are observed unemployed at least until the first month of the time window.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

For each program we estimate three binary *probit* models on participation in the program vs. not participating in any program within each of the respective time windows. The specification of the respective models was chosen as to include all covariates that potentially influence the selection into treatment and the success of the program. Table 3.4 contains a listing of the covariates used in our preferred specification. We include all variables that show up highly significant in at least one of the models. We only modify the estimation when there is a lack of variation between treated and controls in the respective time windows.[22] Given the differential characteristics of program participants, the sign and power of control variables in predicting treatment vary strongly across programs and entry time, in particular for the extensive set of information on past labor market history. Independent of program, the most important variables include schooling and vocational training information, calendar month of entry into treatment; potential entry in 2003; last

[22]We tested the sensitivity of our results by specifying more parsimonious models but found very little differences in the estimated effects.

contact to the employment agency; and the number of placement offers.[23] The latter two variables are of particular interest, as they proxy the closeness between youths and the employment agency and give potential signals for the labor market performance of youths as perceived by the caseworker. In particular, we observe a strong and significant inversely U-shaped relation between placement propositions and treatment participation for all programs except PT, which means that youths with extremely low or high number of employment options are less likely to participate in ALMP.

Based on the predicted values of the propensity scores, weights are constructed within each of the 144 cells. To ensure that we only compare individuals with similar values of the PS and reduce the incidence of extreme values in the PS distribution we exclude observations outside the region of common support by dropping treated and non-treated individuals who have PS values above (below) the maximum (minimum) value of the respective other group (Dehejia and Wahba, 1999). This predominantly yields to a deletion of non-treated individuals at the lower end, and very few treated individuals at the upper end of the PS distribution (see Table A3.3 in the Supplementary Appendix).[24] After imposing common support we perform weighting for all outcomes in each of the 60 months following program entry to obtain the short-, medium- and long-term treatment effects; standard errors are obtained by bootstrapping the entire matching procedure (including propensity score estimation) using 200 replications.

3.4.4 Balancing Tests

As the essential objective of IPW is to balance the distribution of observable characteristics between participants and non-participants, we test the success of the procedure by comparing the differences in the distributions of covariates of treated and weighted controls. Among the many approaches to do so, we choose a simple comparison of means t-test, and the mean standardized bias (MSB) in the weighted sample.[25] The MSB is defined as the differences in covariate means as a percentage of the square root of the average sample variances of the treatment

[23]The predictive power of the respective models ranges closely around 70% for all models, see Table A3.3 in the Supplementary Appendix. Full estimation results are available on request.

[24]We investigate the robustness of our results with respect to the choice of the common support and potential outliers in the sensitivity analysis in Section 3.5.3.

[25]See Caliendo and Kopeinig (2008) for a more detailed discussion of matching quality issues.

Table 3.4: Set of covariates included in the propensity score estimation

Information category	Specification details
Socio-demographic characteristics	Gender (dummy: Female) Age (dummy: below or above 20 years) Living situation: - living alone - living together married - living together not married Migration status (dummy) Having children (dummy)
Education level and health condition	School leaving certificate - none - lower secondary degree - middle secondary degree - upper/specialized secondary degree Having finished professional/vocational training (dummy) Health restrictions (dummy)
Information on last activity/employment	Last activity before entry into unemployment - regular employment - education, training, never employed - other Occupational group of previous job - agriculture - manufacturing, technical occupations - services - other Having professional experience (dummy) Daily income from last regular employment (log) Information available on working time at last employer (dummy)
Labor market history for past year and past three years	During the last year before unemployment entry (linear) - months spent in employment - months spent in unemployment - months spent in ALMP - months spent in inactivity - months spent in full-time employment(1) - months spent in part-time employment(1) During the last three years up to unemployment entry (linear) - months spent in employment - months spent in unemployment - months spent in ALMP - months spent in inactivity - months spent in full-time employment(1) - months spent in part-time employment(1) During the last three years up to unemployment entry (dummy) - never been in regular employment - never been in ALMP - never been in inactivity - never in full-time employment(1) - never in part-time employment(1)
Information on current unemployment and caseworker information	Months of remaining benefit entitlement (linear) Quarter of entry into unemployment (4 dummies) Unemployment spell lasts until 2003 (dummy) Months since last contact to employment agency - never contacted before - less than six months - more than six months - information missing Information available on preferred working time (dummy) Number of placement propositions by caseworker (linear and squared)
Regional Characteristics	Unemployment rate (linear) GDP growth during last year (log)

Note: This baseline specification was modified if observations where dropped from the analysis due to lack of variation. In particular we dropped the variable "information of working time wanted" for the case of JCS, WS, PT and FT measures; information on previous employment occupation for PT and FT; the square of the placement proposition for WS and PT; and the information on migration status for FT.
(1) The information of working time available can be divided into three categories, full-time, part-time and "not quite full-time". The latter was dropped from the analysis.

and control group, whereby it is generally assumed that a MSB below 5% reflects a well-balanced covariate distribution in the sample. We control for 53 variables in our PS specification and find that around half of the variables are rejected to have equal means in a one-sided 5% significance t-test before weighting is conducted. After weighting, however, the same test finds for all programs that none of the variables has unequal means. Similarly encouraging results are obtained using the MSB as a criterion. Before weighting the MSB is around 20%, but afterwards it is below 3% for all programs and time windows in East Germany and below 2% in the West. Overall, this indicates that reweighting yields a control group that is very similar to the treatment group with respect to their observable characteristics at point of entry into treatment.[26]

3.5 Main Results and Sensitivity

3.5.1 Key Results

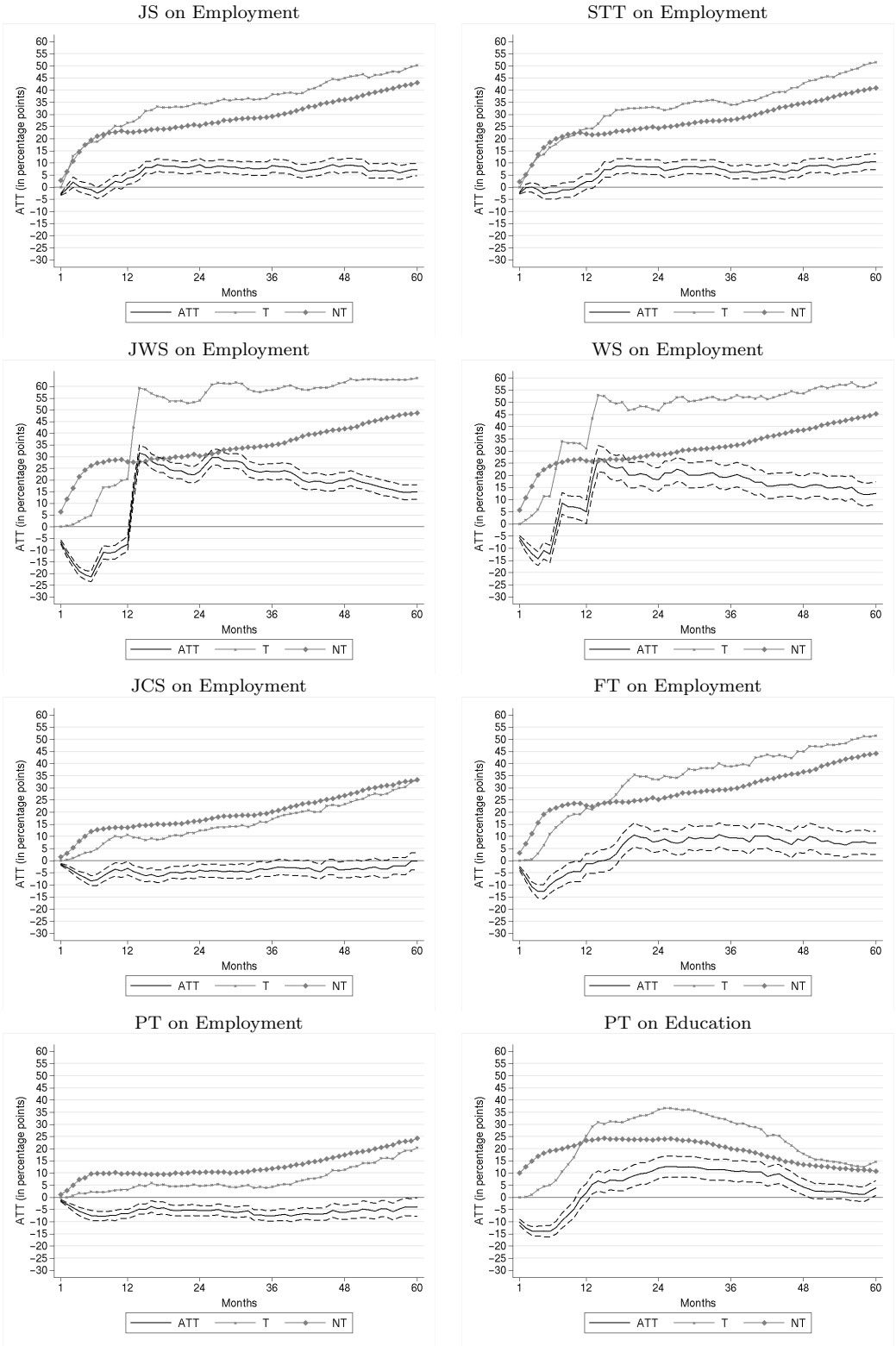
As our primary outcome of interest we consider the integration in unsubsidized regular employment.[27] Figures 3.3 (East Germany) and 3.4 (West Germany) plot the treatment effect estimates on the employment probabilities during the 60 months following program entry. Monthly effects are calculated as the difference between treated and (weighted) control outcomes, which we also plot to facilitate interpretation. Additionally, we provide the cumulative effects of program participation after 30 and 60 months in Table 3.5. We focus on overall effects irrespective of timing of entry and address differences only if they are of interest.

The monthly outcome plots reveal that except for JCS and PT measures, all programs significantly improve the labor market prospects of participants. Following initial locking-in and transition phases, the treatment impact stabilizes for all programs at around two years after program entry. The long-run impact of program participation—after the third year of program entry and onwards—amounts to a monthly employment boost between 5 to 20 percentage points, depending on program and region.

[26] See Tables A3.4 and A3.5 in the Supplementary Appendix for the detailed results of the t-test and the MSB.

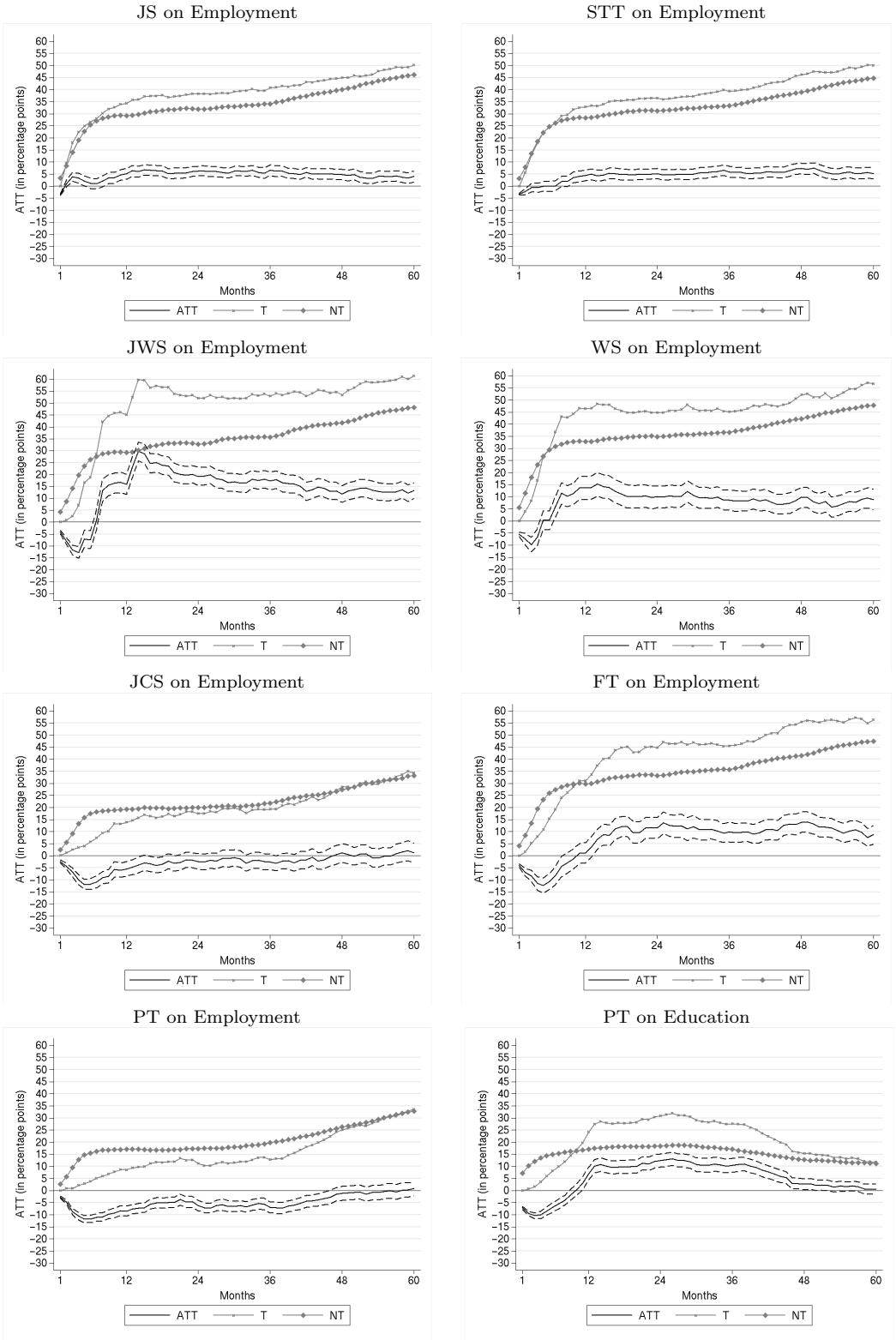
[27] We only consider employment subject to social security contributions as a success. This excludes "marginal employment", i.e. jobs that pay only up to 400 Euro and entail reduced social security contributions from the employer.

Figure 3.3: Causal effects of program participation in East Germany over time—aggregate results over all program entries.



Note: The black solid line depicts the average treatment effects on the treated (ATT) - the dashed black line provides the 95% confidence-band based on bootstrapping with 200 replications. The ATT is the difference between the average monthly outcomes of the treated (T) and the *weighted* average outcomes of the non-treated (NT) - the corresponding values are shown in gray.

Figure 3.4: Causal effects of program participation in West Germany over time—aggregate results over all program entries



Note: The black solid line depicts the average treatment effects on the treated (ATT) - the dashed black line provides the 95% confidence-band based on bootstrapping with 200 replications. The ATT is the difference between the average monthly outcomes of the treated (T) and the *weighted* average outcomes of the non-treated (NT) - the corresponding values are shown in gray.

We see that WS and JWS are the most successful programs in East Germany in the long-run (i.e. at the end of our observation period) with an average impact of 20 to 25 percentage points. Similarly, JWS is the most successful program in West Germany, with a 20 percentage point program impact, while here the effects of WS and FT are around 10 percentage points. The difference in relative impacts of wage subsidies and training measures in both regions seems to be in line with the notion that West German program participants are more constraint by their adverse labor market characteristics than demand side restrictions. Hence, programs that aim at gradually enhancing labor market skills, i.e. long-term classroom training or long-term practical experience are more apt to overcome the entry barriers faced by West German youths.

The labor market integration of participants in wage subsidies (JWS and WS) takes place in discontinuous jumps, suggesting an immediate integration into the labor market. As firms were required to offer a minimal period of unsubsidized employment following the subsidy, this is driven by the continuation of the employment relationship within the same firm. Even though we see a small decline in the employment probabilities when the employment guarantees expire, the overall employment levels of the treated remain remarkably high (between 45% to 60%), such that wage subsidies can be seen as stepping stone into stable unsubsidized employment. In contrast to the immediate integration of participants in wage subsidies, participants in training measures (JS, STT and FT) experience a period of high intensity transitions into employment after the program has ended. This period lasts for about six to twelve months and can be seen as causal for the persistent employment gap between treated and non-treated individuals during the rest of the observation period. Training measures in the East perform similar independent of their duration—with a long-term employment impact of about 10 percentage points; whereas in the West short-term training (JS and STT) increases the employment probabilities of participants less than long-term training (FT). The effects for JS and STT have to be interpreted with caution, since a significant share of youths in the East (40%) and West (27%) participate in further ALMP programs. We address this issue in our sensitivity analysis in Section 3.5.3. In contrast to the previous programs, JCS and PT do not exhibit any positive long-term employment impact on program participants. In particular we find that participation in these programs decreases the probability of entering employment in the medium-run, even though the negative effect phases out to zero over the

course of the observation period.

A further thing to note is that youths participating in longer-term measures experience severe locking-in effects during program participation—around 10 to 20 percentage points. If one interprets the level of locking-in during program participation as an initial investment, the cumulative benefit of program participation should be taken as measure for the net program effectiveness. The strength of locking-in depends on the opportunity costs of participation that are a function of, e.g., the program duration and the timing of entry into the program. Since non-participating youths experience particularly strong transitions out of unemployment during the first six months in their unemployment spell, this substantially aggravates the opportunity costs of entering the program during this phase. Table 3.5 presents the cumulative employment effects (30 and 60 months after program entry) overall and differentiated by entry strata.

Several issues emerge considering the employment outcomes: First, it can be seen that the relative cumulative long-run effectiveness of programs is largely consistent with the relative monthly long-run effectiveness. After 60 months, participants in wage subsidies yield the largest cumulative effects (up to nine months in East and five to nine months in West Germany). For the shorter programs JS and STT the cumulative effects are significantly positive between three and four months. For the longer FT measures, the effects are partly not significant after 30 months (due to long duration of the program), but turn positive after 60 months (3 and 4.5 months in East and West Germany). JCS and PT are the two programs with negative cumulative employment effects throughout. Second, we find that for almost all programs the cumulative effects are increasing with the timing of entry. In particular, we do not find significant differences in the monthly employment effects by entry time[28], so that the high opportunity costs of an early entry largely drive these results. Compared to individuals entering in the first three months of their unemployment spell, the locking-in costs are significantly reduced for later program entries. The largest differences across entry strata occur for JWS in the West, with a six-months cumulative gap for the earliest and the latest entries after 60 months.

Even though the integration into regular employment is the primary outcome of interest, we also test whether programs increase the participation in further un-

[28]Detailed monthly outcome plots by entry time into the program are available from the authors upon request.

Table 3.5: Cumulative treatment effects 30 and 60 months after program entry on regular employment probabilities

		East Germany				West Germany			
	Entry time/ Σ	All	1-3	4-6	7-12	All	1-3	4-6	7-12
JS	Σ 30	**1.49**	**0.94**	**2.28**	**2.15**	**1.37**	0.48	**2.09**	**2.86**
	(s.e.)	(0.25)	(0.35)	(0.56)	(0.54)	(0.22)	(0.28)	(0.43)	(0.51)
	Σ 60	**3.81**	**3.33**	**5.35**	**3.74**	**2.85**	**1.41**	**3.54**	**5.76**
	(s.e.)	(0.54)	(0.72)	(1.18)	(1.13)	(0.42)	(0.56)	(0.83)	(0.99)
STT	Σ 30	**1.27**	0.61	**1.75**	**2.28**	**0.98**	0.02	**2.18**	**2.17**
	(s.e.)	(0.31)	(0.43)	(0.58)	(0.70)	(0.23)	(0.32)	(0.48)	(0.48)
	Σ 60	**3.65**	**2.82**	**4.86**	**4.28**	**2.75**	**1.86**	**4.69**	**3.00**
	(s.e.)	(0.57)	(0.82)	(1.16)	(1.34)	(0.45)	(0.61)	(0.88)	(1.03)
JWS	Σ 30	**3.10**	**1.60**	**5.47**	**5.51**	**4.16**	**2.34**	**4.86**	**7.28**
	(s.e.)	(0.31)	(0.38)	(0.62)	(0.73)	(0.38)	(0.50)	(0.57)	(0.80)
	Σ 60	**9.09**	**7.37**	**12.36**	**11.27**	**8.53**	**6.16**	**9.20**	**12.92**
	(s.e.)	(0.62)	(0.78)	(1.39)	(1.55)	(0.71)	(0.99)	(1.23)	(1.63)
WS	Σ 30	**3.53**	**2.94**	**5.55**	**3.89**	**2.42**	**1.80**	**3.22**	**4.11**
	(s.e.)	(0.49)	(0.56)	(1.08)	(1.17)	(0.47)	(0.53)	(0.87)	(1.46)
	Σ 60	**8.49**	**8.12**	**10.40**	**7.96**	**4.92**	**3.60**	**6.70**	**8.32**
	(s.e.)	(1.02)	(1.14)	(2.36)	(2.57)	(0.86)	(1.00)	(1.62)	(2.60)
JCS	Σ 30	**-1.47**	**-2.86**	**-1.01**	-0.81	**-1.38**	**-2.47**	-0.02	-0.52
	(s.e.)	(0.25)	(0.46)	(0.49)	(0.42)	(0.30)	(0.40)	(0.70)	(0.58)
	Σ 60	**-2.38**	**-3.76**	-1.12	**-2.13**	**-1.63**	**-2.59**	-0.47	-0.84
	(s.e.)	(0.56)	(1.01)	(1.07)	(0.84)	(0.64)	(0.95)	(1.50)	(1.21)
FT	Σ 30	0.27	**-1.79**	**1.81**	**2.09**	**1.23**	0.48	**2.35**	**2.00**
	(s.e.)	(0.44)	(0.61)	(0.71)	(1.01)	(0.44)	(0.58)	(0.85)	(0.90)
	Σ 60	**2.86**	-0.07	**5.15**	**5.28**	**4.47**	**3.61**	**6.03**	**5.04**
	((s.e.)	(0.98)	(1.35)	(1.53)	(2.17)	(0.83)	(1.09)	(1.69)	(2.01)
PT	Σ 30	**-1.64**	**-2.09**	**-0.87**	**-1.50**	**-2.14**	**-2.65**	**-0.99**	**-1.96**
	(s.e.)	(0.20)	(0.29)	(0.31)	(0.44)	(0.20)	(0.24)	(0.38)	(0.49)
	Σ 60	**-3.43**	**-4.13**	**-2.45**	**-3.01**	**-3.09**	**-3.98**	**-1.15**	**-2.69**
	(s.e.)	(0.43)	(0.59)	(0.70)	(0.93)	(0.42)	(0.51)	(0.86)	(0.95)

Note: Cumulative effects are obtained by summing up the monthly treatment effects. Standard errors in parentheses are obtained by bootstrapping with 200 replications. Bold numbers indicate significance at the 5% level.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

subsized education or training, i.e., apprenticeships or higher secondary/tertiary schooling. As the administrative data only records apprenticeship participation we use the filling procedure described already in Section 3.3.3 (further details in Appendix A3.2) to impute information on alternative training spells. The treatment estimates on effects on the monthly probability to participate in unsubsidized education for participants in PT programs are depicted in the lower right panel of Figures 3.3 and 3.4. For the other measures—which are aimed at integration into employment—the cumulative impacts on education participation are depicted in Table 3.6. It can be seen that PT measures do indeed significantly improve par-

Table 3.6: Cumulative treatment effect 30 and 60 months after program entry on education participation

		East Germany				West Germany			
	Entry time/ Σ	All	1-3	4-6	7-12	All	1-3	4-6	7-12
JS	Σ 30	**-1.14**	**-1.02**	-0.62	**-1.82**	**-0.99**	**-0.55**	**-1.16**	**-1.93**
	(s.e.)	(0.14)	(0.16)	(0.39)	(0.31)	(0.14)	(0.19)	(0.27)	(0.28)
	Σ 60	**-1.64**	**-1.54**	-0.84	**-2.49**	**-1.4**	**-0.71**	**-1.61**	**-2.93**
	(s.e.)	(0.26)	(0.31)	(0.68)	(0.62)	(0.25)	(0.34)	(0.48)	(0.45)
STT	Σ 30	**-1.26**	**-0.76**	**-2.10**	**-1.56**	**-1.00**	**-0.73**	**-1.15**	**-1.55**
	(s.e.)	(0.19)	(0.28)	(0.30)	(0.40)	(0.15)	(0.21)	(0.30)	(0.33)
	Σ 60	**-1.54**	**-1.06**	**-2.64**	**-1.54**	**-1.31**	**-1.04**	**-1.65**	**-1.65**
	(s.e.)	(0.34)	(0.47)	(0.58)	(0.73)	(0.25)	(0.36)	(0.50)	(0.59)
JWS	Σ 30	**-2.49**	**-2.23**	**-2.77**	**-3.07**	**-2.20**	**-1.70**	**-2.43**	**-3.01**
	(s.e.)	(0.15)	(0.18)	(0.28)	(0.37)	(0.16)	(0.25)	(0.27)	(0.34)
	Σ 60	**-3.91**	**-3.48**	**-4.14**	**-5.08**	**-3.16**	**-2.14**	**-4.15**	**-4.17**
	(s.e.)	(0.27)	(0.35)	(0.64)	(0.54)	(0.32)	(0.52)	(0.48)	(0.77)
WS	Σ 30	**-2.32**	**-2.23**	**-3.18**	**-1.73**	**-1.34**	**-1.05**	**-2.01**	**-1.55**
	(s.e.)	(0.23)	(0.26)	(0.47)	(0.71)	(0.22)	(0.29)	(0.40)	(0.58)
	Σ 60	**-3.73**	**-3.98**	**-4.01**	-2.28	**-2.20**	**-1.84**	**-2.98**	**-2.57**
	(s.e.)	(0.40)	(0.45)	(0.94)	(1.18)	(0.40)	(0.52)	(0.81)	(0.89)
JCS	Σ 30	**-1.58**	**-1.30**	**-1.32**	**-1.88**	**-0.96**	-0.21	**-1.64**	**-1.75**
	(s.e.)	(0.22)	(0.37)	(0.38)	(0.36)	(0.28)	(0.42)	(0.52)	(0.43)
	Σ 60	**-1.82**	-1.26	**-1.77**	**-2.2**	-0.73	0.25	-1.19	**-2.06**
	(s.e.)	(0.43)	(0.71)	(0.79)	(0.66)	(0.54)	(0.80)	(1.13)	(0.80)
FT	Σ 30	**-1.85**	**-1.60**	**-2.12**	**-1.96**	**-1.79**	**-1.67**	**-1.94**	**-1.95**
	(s.e.)	(0.21)	(0.34)	(0.33)	(0.58)	(0.21)	(0.27)	(0.40)	(0.58)
	Σ 60	**-2.91**	**-2.87**	**-2.73**	**-3.25**	**-2.40**	**-2.19**	**-3.00**	**-2.26**
	(s.e.)	(0.43)	(0.65)	(0.69)	(0.98)	(0.43)	(0.51)	(0.75)	(1.07)
PT	Σ 30	0.65	0.82	-0.26	1.22	**1.47**	**2.17**	1.09	-0.23
	(s.e.)	(0.42)	(0.63)	(0.83)	(0.87)	(0.27)	(0.38)	(0.57)	(0.56)
	Σ 60	**2.67**	**3.01**	0.81	**3.88**	**3.14**	**4.40**	**2.42**	0.17
	(s.e)	(0.71)	(1.06)	(1.32)	(1.40)	(0.47)	(0.65)	(0.96)	(0.99)

Note: Cumulative effects are obtained by summing up the monthly treatment effects. Standard errors in parentheses are obtained by bootstrapping with 200 replications. Bold numbers indicate significance at the 5% level.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

ticipation in education. After about one year after entry into the program, participants experience a stable positive increase in education probabilities of around 10 percentage points between month 12 to 48. Coinciding with the approximate three-year duration of an apprenticeship in Germany this is indicative of successful completion of a professional training. Also with respect to education outcomes we find that the timing of program entry matters, as we observe an actual decline in effectiveness for later entries (see Table 3.6) in the West. Potentially driven by discouragement or rapid reduction in human capital for the rather young participants of PT, the fast integration into education seems to be crucial in order to

avoid negative long-term effects of unemployment. Table 3.6 also shows that none of the programs aimed at integrating youths into the first labor market have a positive impact on the education probabilities.

Further evidence for the education effect is given by a descriptive analysis of the share of youths having obtained a professional qualification until the end of our observation period (i.e. at most 72 months after initial unemployment entry), in Table 3.7. It can be seen that participants in PT have a significantly higher share of apprenticeship graduates at the end of the observation period, that is 20%-points higher in the East and 17%-points in the West. The increase is only at 8%-points and 6% for East and West German non-participants, respectively.

For youths who participated in employment programs. the average level of professional training does not increase strongly (about 3%-points on average) For East Germany this is not surprising as youths exhibit above average shares of professional training already at program start. In the West, however, about one third of participating youths still do not have any type of professional training at the end of our observation period. Again, youths participating in JCS fare much worse than the rest with about 40% (75%) of youths being without any professional degree after 72 months.

3.5.2 Effect Heterogeneity

In this section we inspect effect heterogeneity across gender and pretreatment schooling levels (below or equal vs. higher than lower secondary schooling certificate). To account for potential differences in the timing and nature of selection into treatment and to ensure that we only compare treated and non-treated within the region of common support we repeat the estimation procedure outlined in Section 3.3 for each of the respective subgroups. This leaves us with 14 distinct program-subgroup cells in East and West Germany (compare Table A3.6 in the Appendix for details of sample size).[29] What should be kept in mind is that the separation of the analysis for the respective subgroups entails that the results are not directly comparable. For example, a higher level in the estimated effects for women does not indicate that the program is more beneficial for women than it is for men,

[29]Due to the small number of observations within some cells, we modify the original PS specification on a case-by-case basis by successively excluding covariates with low explanatory value to obtain the optimal specification in terms of correct predictions rates. Full estimations results and further details are available upon request.

Table 3.7: Comparison of participant and non-participant highest vocational degree at point of entry into unemployment and 72 months later.

		East Germany			West Germany		
		$t = 0$	$t = 72$	Δ	$t = 0$	$t = 72$	p-value
	Professional training						
	none	0.23	0.18	**-0.05**	0.46	0.39	**-0.07**
JS	apprenticeship	0.76	0.80	**0.04**	0.53	0.59	**0.06**
	university	0.00	0.02	**0.02**	0.01	0.02	**0.01**
	Professional training						
	none	0.29	0.24	**-0.05**	0.48	0.41	**-0.07**
STT	apprenticeship	0.70	0.73	**0.03**	0.50	0.56	**0.06**
	university	0.02	0.03	**0.01**	0.02	0.03	**0.01**
	Professional training						
	none	0.13	0.10	**-0.03**	0.35	0.32	**-0.03**
JWS	apprenticeship	0.85	0.88	**0.03**	0.64	0.67	**0.03**
	university	0.01	0.02	**0.01**	0.01	0.02	**0.01**
	Professional training						
	none	0.22	0.18	**-0.04**	0.40	0.34	**-0.06**
WS	apprenticeship	0.76	0.79	**0.03**	0.59	0.63	**0.04**
	university	0.02	0.03	**0.01**	0.01	0.03	**0.02**
	Professional training						
	none	0.47	0.39	**-0.08**	0.85	0.74	**-0.11**
JCS	apprenticeship	0.52	0.58	**0.06**	0.14	0.22	**0.08**
	university	0.01	0.03	**0.02**	0.01	0.04	**0.03**
	Professional training						
	none	0.17	0.13	**-0.04**	0.36	0.32	**-0.04**
FT	apprenticeship	0.83	0.86	**0.03**	0.62	0.65	**0.03**
	university	0.01	0.01	0.00	0.02	0.03	**0.01**
	Professional training						
	none	0.89	0.68	**-0.21**	0.93	0.74	**-0.19**
PT	apprenticeship	0.09	0.29	**0.20**	0.06	0.23	**0.17**
	university	0.02	0.03	**0.01**	0.01	0.03	**0.02**
	Professional training						
	none	0.52	0.41	**-0.11**	0.55	0.48	**-0.07**
NP	apprenticeship	0.46	0.54	**0.08**	0.43	0.49	**0.06**
	university	0.02	0.04	**0.02**	0.02	0.03	**0.01**

Note: Δ depicts raw differences between the two values; bold numbers indicate significance at the 5%-level from a one-sided t-test.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

but that women have a higher benefit compared to non-participating women than men have compared to non-participating men. In the Appendix selected monthly treatment effects estimates on the employment probabilities in Tables A3.7 (gender) and A3.9 (schooling levels); cumulative effects on employment and education outcomes can be found in Tables A3.8 and A3.10.

Effects by Gender Our estimates reveal very minor differences in the monthly employment effects across gender. Only the long-run persistency of effects appears

to differ for some programs. In East Germany we find for all programs except PT, that two to three years after program entry the average monthly treatment impact of women declines substantially and then stabilizes again at a lower (but positive) level towards the end of the observation period. In the West we find a similar, but less pronounced long-term reduction in treatment effects for female participants in STT, JS and WS. This is potentially explained by an increased labor force attachment among women with a successful program participation, who delay the timing of family planning (compare Lechner and Wiehler, 2011, for similar results on ALMP in Austria). Examples on short-to medium-run differences between young men and women occur for participants in WS, and training measures in the West. For the case of WS we find that after an initially similar program impact, the employment probabilities of men in East and women in the West decline substantially during the 12 months following program participation, while they remain stable for the other groups. These differences are most likely driven by differences in take-over probabilities of the firm receiving the subsidy, the cause of which would however require a more in-depth analysis of firm and participant characteristics. In the case of STT and FT measures in the West we find that women seem to benefit much less from STT measures than men (the cumulated effect only amounts to 1.5 months), but benefit more from longer-term training in FT. The latter finding is in line with the observation that young women generally perform better in school-based training than young men—a validation would require a direct comparison of the subgroups however.

Effects by Schooling Levels Youths with different levels of pretreatment schooling have different returns to program participation. By and large these differences can be summarized into programs being more effective for high-skilled youths in terms of employment outcomes. In particular we find that participants in WS, JS, STT and FT with high levels of pretreatment schooling spend on average six months longer in employment than their non-treated counterparts over the whole observation period—compared to three months for youths with low schooling levels (see Table A3.9). We also observe that the periods of locking-in go beyond the median program duration for youths with a low schooling degree, which would correspond to further program enrollment. In the case of a successful further participation, the true gap in program success for youths with low and high pretreatment schooling in the first program is expected to be even larger. An exception from these differential effects is given by JWS and JCS measures, which seem to be

equally beneficial (detrimental) in terms of employment outcomes. The program effect of participation in JCS is either zero or slightly negative for both subgroups, while *all* youths participating in JWS have a cumulative employment gain of eight to ten months. As such the finding on JWS is an encouraging deviation from the our earlier findings as it is also driven by similar long-run effects, and not solely by the leveling of locking-in and program effects. In terms of education outcomes for participants in PT measures (last two rows of Table A3.9), we also observe that youths with higher schooling levels experience higher rates of education participation between month 12 to 36.

3.5.3 Sensitivity Analysis

We test the sensitivity of our results with respect to the crucial assumptions made in the main analysis. First, we consider the problem of further program participation and investigate to what extent our treatment estimates of the first participation in JS and STT measures are driven by participation in further measures. Second, we apply a dynamic evaluation approach that changes the composition of the control group. Finally, we check whether different variants of imposing common support alter our results. Table A3.11 in the Appendix presents the estimated cumulative employment effects from the sensitivity analysis together with the results obtained in the main analysis as a reference.[30]

Further Program Participation We have noted in Chapter 3.5.1 that the effects for JS and STT have to be interpreted with caution, since a significant share of youths participate in further ALMP programs. To be more specific, 44% (31%) of the JS participants in East (West) Germany participate in a further ALMP program within one year and the same is true for 38% (24%) of the participants in STT. As only individuals for whom the program did not lead to an entry into employment are assigned to further programs, the effectiveness of the initial measures would require the consideration of fully dynamic selection effects, which is beyond the scope of this analysis (see Lechner and Miquel, 2010, for an estimation approach). Instead we assess the sensitivity of our findings by restricting the sample of treated to individuals who participate in only one program during the first twelve months of their unemployment spell. This is insightful as it provides

[30]Results on education probabilities are not presented separately as their sensitivity is very similar to employment outcomes. But they are available upon request.

an indication whether any of the positive employment effects are attributable to participation in the initial program. As we exclude only youths for whom the program was unsuccessful, our sensitivity estimates are likely to be more positive than for the average participant. The results in Table A3.11 show that the new results are very similar to the results from the main analysis. We repeated this exercise not only for participants in JS and STT but also for the other programs (where the probabilities of subsequent participation is much lower). The medium- and long-run cumulative effects are very similar to the reference estimates for all programs; none of the cumulative effects after 60 months in the sensitivity analysis differs significantly from the main results.

Dynamic Evaluation Approach We assess the sensitivity of our results with respect to the choice of the evaluation approach and re-estimate our results using a dynamic approach, as outlined in Section 3.3.2. We hence redefine our control group to include youths who participate at any point in time later during their unemployment spell and who potentially participate in other programs. We find that the point estimates vary slightly using the dynamic approach, but none of these changes are significant at a conventional level. The observed increase in effects for the majority of programs is most likely due to controls entering other programs under investigation. As they experience periods of locking-in themselves, the opportunity cost of participating in the program of investigation is reduced. Given the large size of our never-treated control group, all of the observed changes are only minor and insignificant. We hence conclude that the choice of the evaluation approach has no significant implications for our results and using the dynamic approach does not change the overall evidence on program effectiveness.

Alternative Imposition of Common Support A necessary condition for the identification of treatment effects is the existence of corresponding non-participants over the whole support of the treated PS distribution, where limited overlap may be particularly distorting when using IPW (as pointed out by Frölich, 2004). We chose the "Min-Max"-condition in Section 3.4.3, but several alternatives have been suggested. Black and Smith (2004) argue that the imposition of a more restrictive trimming of the PS distribution might be beneficial if treated (controls) with very low (high) values of the PS are more likely to suffer from measurement error in the treatment variable, and remaining unobserved factors are more important here. To assess the sensitivity of our results with respect to this issue we conduct several robustness tests. First, we exclude control observations with very large

values of the PS (above the 99 percentile). Second, we exclude areas of the distribution where there is only low overlap between treated and controls and restrict the common support to an "optimal" area defined by $\alpha \leq P(W) \leq 1-\alpha$, whereby α is chosen to balance two opposing variance components (as suggested by Crump et al., 2009). While the variance of the estimate increases due to the lower number of observations, it decreases with an improved level of overlap between treated and non-treated.[31] Finally, we restrict the propensity score distribution even more, by dividing the distribution into twenty equidistant percentiles and estimate the effects only in regions where we have at least 5% of treated and non-treated observations. Clearly, restricting the estimation to areas of "thick support" reduces the validity of the results and might potentially lead to changes in estimated effects. This has the drawback that it is unclear whether changes are due to effect heterogeneity, large weights of outliers, or unobserved heterogeneity in characteristics. The results in Table A3.11 show that our effect estimates hardly change. This confirms our expectations discussed in Section 3.4.1, namely that due to a large sample of non-participants and a restrictive common support condition ("Min-Max" cut off rule) this issue is of minor relevance in our case.

3.6 Conclusion

Plagued with a persistent problem of long-term unemployment among youths, Germany is one of the European countries with the highest expenditures on youth ALMP—at 1.7 billion euros per year between 1999 and 2002. Between 2000 and 2010 about 1.4 million youths entries into ALMP were recorded each year—and the number is increasing. This evaluation study provides the first comprehensive assessment of the short-to-long-term employment impact of participation in various ALMP programs in place.

Based on a representative sample on young unemployment entries in 2002, we investigate the effectiveness of program participation vs. non-participation using an quasi-experimental estimation approach with IPW. Analyzing a broad range of instruments that belong to the common set of policy tools employed in European countries, we add to the previous European evaluation literature dealing with youth ALMP. We conduct the analysis separately for youths in East and West

[31] The implementation of this is done using the STATA tool *optselect.ado* provided by the Crump et al. (2009).

Germany, shedding some light on the effectiveness of the respective measures to improve the employment situation of youths under differential social, economic and labor market conditions.

In terms of improving the employment probabilities of unemployed youths, the overall picture of the different ALMP analyzed is rather positive, indicating a persistent and stable employment effect. In particular, we find a significant increase in employment probabilities for almost all measures examined. Focusing on the long-term employment impact, the strongest effects are observed for participants in wage subsidies (10 to 20 percentage points); job search assistance, short- and longer term training measures yield smaller but also persistently positive effects (5 to 10 percentage points). With respect to education outcomes we find that preparatory programs aimed at integrating youths into an apprenticeship are successful in doing so. In contrast to the aforementioned beneficial employment programs, public sector job creation schemes (JCS) are found to be harmful for the employment prospects of participants in the short- to medium-run and ineffective in the long-run. Put more drastically, if one considers the initial program participation as investment into future labor market outcome, the return of participating in JCS is negative throughout the whole observation period of five years. This is consistent with previous evaluation results for other countries that show the ineffectiveness of JCS for youths (compare, e.g., Dorsett, 2006, for the "environmental task force" implemented in the New Deal for Young People in the UK), and for the adult population (compare, e.g., Caliendo et al., 2008). Against these overwhelmingly negative findings for JCS it is surprising that during the current economic crisis policy makers still consider the temporary extension of these measure to counteract soaring levels of youth (long-term) unemployment rates (compare OECD, 2011).

In terms of a differential impact of the respective measures under different labor market conditions, our analysis provides evidence from the comparison of the employment impact for program participants in East and West Germany. For all measures we find similar qualitative results, suggesting that the programs can be sufficiently adapted to benefit in either type of economic environment. However, we also find that the relative benefit of longer-term training measures (FT) compared to wage subsidies (WS) seems to be higher in the West than in the East, which needs to be interpreted with the significantly lower pretreatment education levels of West German youths in mind. While youths in the East are characterized by high initial schooling levels, the provision of work experience by removing demand-

side barriers seems to be the most important hurdle to integrating into the labor market. In contrast, youths in the West have much less favorable labor market characteristics and hence seem to benefit more from an improvement in human capital endowment. Further evidence for this is given by the finding that only youths with high schooling levels in the West experience a positive long-term employment impact of participation in preparatory training. For youths in the East, the acquisition of a professional degree might not be sufficient to protect them from struggling at the "second barrier".

Recent statistics on youth unemployment levels in Germany (and similarly in other European countries) show that the probability to enter unemployment is significantly higher for low-educated than medium-educated youths, with a steadily increasing gap. Together with the expected shortage of labor in the medium-run the by far most vulnerable labor market group will be low-educated youths, making them the most important target of policy intervention. Our analysis provides evidence however, that the these youths are not sufficiently accommodated in the current policy set-up. In particular we find that all programs except JWS improve the labor market prospects of youths with high levels of pretreatment schooling to a greater extent than that of youths with low levels of pretreatment schooling. This suggests an insufficient adjustment of the respective measures for the requirements of unskilled youths. We further find that youths who are assigned to the most successful employment measures within the first twelve months in unemployment, compared to later- or never-participants, have much better characteristics in terms of their pre-treatment employment chances. As the program assignment process is likely to favor individuals for whom the measures are most beneficial, the observed strong positive selection of youths into ALMP—in particular in the East—supports our interpretation of a systematic lack of ALMP alternatives that could benefit low-educated youths.

Our analysis also indicates potential avenues for the improvement of ALMP for low educated youths. So far, none of the programs aimed at labor market integration increases the education participation of youths. By readjusting existing labor market programs to accommodate participation in further education or training as intermediate objective, the integration of low-educated youths into the labor market could be done in a more sustainable manner. Secondly, we find that wage subsidies of shorter duration work better for high-schooling youths, while wage subsidies with longer duration work equally well for low and high educated

youths. This suggests that low educated youths require more time to turn the subsidized work experience into a stepping stone to a stable employment entry. By extending the access to longer-term professional experience for these youths, an additional barrier of labor market integration for these could potentially be removed.

Appendix

A3.1 Sample Selection

Table A3.1: Documentation of sample reduction

	Loss of Individuals	Number of Individuals
Total inflows into unemployment		851,258
Implemented restrictions		
Entries in 2002 only	607,702	243,556
Youth only (≤25 years)	187,898	55,658
Data cleaning[(1)]	913	54,745
Other programs[(2)]	2,960	51,797
Missing in any variables of the PS specification	778	51,019
Estimation sample		51,019
East Germany		17,515
Participants		5,353
Non-participants		12,162
West Germany		33,504
Participants		7,027
Non-participants		26,477

[(1)] We exclude individuals with missing information only (except an unemployment spell of a maximum of one week) and also individuals who die during our observation period.
[(2)] Individuals participating in different programs of ALMP to those under scrutiny (see Table 3.1) are excluded.

A3.2 Imputation of Missing Information

To overcome the potential problem of non-randomly missing outcome information, we impute missing spells with information that is recorded with every registered spell of unemployment, employment or benefit receipt. For each of these spells the main planned activity subsequent to the spell is available. Furthermore, for each registered spell of unemployment, additional information on the previous activity is recorded by the caseworker.

For example, if an individual leaves unemployment because he has to serve in the army (which was compulsory for men within our observation period), he disappears from the registered data. Military service is recorded as the reason for leaving the unemployment status and we fill the missing period with this information. If he returns to unemployment after having served in the army, this can be verified, as

we again should observe the military service as the previous activity. However, we only observe the previous activity if the individual registers as unemployed. If he or she finds employment, we have to rely on the initial leaving information of the unemployment spell before military service. Table A3.2 summarizes the missing information that could be filled by this procedure.

Table A3.2: Documentation of filling procedure

	Individuals		Months	
	N	%	N	%
Total	51,019	100	3,673,368	100
Affected by missings	36,493	71.53	942,564	25.66
Filled			866,707	23.59
participants			113,278	13.07
non-participants			753,429	86.93
Remaining missings	6,576	12.88	75,857	2.07
Filling details				
Participants				
% positive employment			21,430	19.30
% positive education			20,179	17.81
Non-participants				
% positive employment			145,454	18.92
% positive education			161,270	21.40

Source: Own calculations, based on the *IZA Evaluation Dataset*.

From the distribution of missing information across program participants and non-participants we see that the data contain significantly more missings for non-participants. This can be explained by a lower attachment of these individuals to the FEA and the resulting lower contact frequency to the caseworker. However, we also find that the type of imputed information is similarly distributed across the two groups for both outcomes considered, so that non-randomly distributed missings should not pose a problem for our analysis.

A3.3 Details on Perfect Alignment

The participants and non-participants are matched directly conditional on the calendar month of entry into unemployment and elapsed unemployment duration. As a starting point we estimate the average treatment effect on the treated for each cell, i.e., participants who entered unemployment in month c of the year and have

a program start after months p in unemployment are compared to non-participants who also entered unemployment in the calendar month c and are still unemployed after in month p after unemployment registration. Hence, within each cell defined by calendar month of unemployment entry and months elapsed before program entry, the effects are defined as:

$$\begin{aligned} \tau_{cp}^{IPW} &= E\left(Y^1 \mid D=1, P(W), \text{UE-Entry}=c, \text{Prg-Entry}=\text{p}\right) - \\ &\quad E\left(Y^0 \mid D=1, P(W), \text{UE-Entry}=c, \text{UE-Duration} \geq p\right) \end{aligned}$$

In a second step the single cell-effects are aggregated to obtain the aggregate effect τ^{IPW}. For this, the 144 monthly effects τ_{cp}^{IPW} are weighted by the distribution of participants across cells:

$$\tau^{IPW} = \sum_{c=1}^{12} \left(\sum_{p=1}^{12} \tau_{cp}^{IPW} \cdot \frac{N_{cp}^1}{N_c^1} \right) \cdot \frac{N_c^1}{N^1},$$

with N_{cp}^1 denoting the number of treated observations within each cell defined by unemployment and treatment entry; N_c^1 denoting the number of treated by calendar month of unemployment entry, and N^1 denoting the total number of treated. After canceling N_c^1 out the total effect τ^{IPW} can be written as:

$$\tau^{IPW} = \frac{1}{N^1} \sum_{c=1}^{12} \sum_{p=1}^{12} \tau_{cp}^{IPW} \cdot N_{cp}^1.$$

A3.4 Additional Tables

Table A3.3: Hit rates of predicted propensity scores and number of observations deleted in the Min-Max common support (CS)

East Germany									
Entry into program	1-3 months			4-6 months			7-12 months		
	Hit Rate	CS NP	CS P	Hit Rate	CS NP	CS P	Hit Rate	CS NP	CS P
JS	68%	762	0	71%	968	2	72%	1,179	0
STT	67%	511	0	68%	428	0	70%	1,364	1
JWS	68%	31	0	70%	1,745	0	74%	1,014	1
WS	62%	896	0	71%	2,637	0	77%	2,802	0
JCS	72%	107	0	71%	2,747	1	71%	411	5
FT	67%	2,292	0	74%	3,137	1	76%	2,663	0
PT	77%	2,873	0	75%	3,276	0	79%	2,815	0
West Germany									
Entry into program	1-3 months			4-6 months			7-12 months		
	Hit Rate	CS NP	CS P	Hit Rate	CS NP	CS P	Hit Rate	CS NP	CS P
JS	65%	191	0	67%	2,296	0	69%	2,002	0
STT	64%	113	1	66%	44	0	68%	1,515	1
JWS	66%	1,701	0	71%	3,474	0	71%	199	0
WS	65%	692	0	70%	1,585	0	79%	8,057	0
JCS	74%	6,348	0	73%	3,159	0	73%	4,853	0
FT	64%	679	0	72%	4,260	0	73%	4,032	0
PT	73%	6,607	0	70%	299	0	74%	2,800	0

Note: The number of deleted observations for treated and controls are the sum of the respective upper and lower bound restrictions. *Hit rate:* Share of participants correctly predicted by the propensity score; *CS NP:* Number of non-participants deleted due to the imposition of the common support condition. *CS P:* Number of participants deleted due to the imposition of the common support condition.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.4: Matching quality: balancing quality of IPW in East Germany —different indicators

Program		JS	STT	JWS	WS	FT	JCS	PT
	Entries between 1 to 3 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	33	20	31	19	25	17	26
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	39	26	35	24	30	24	29
Matched		0	0	0	0	0	0	0
Unmatched	10%-level	42	30	41	26	34	27	35
Matched		0	0	0	0	0	1	0
	Mean standardized bias							
Unmatched		19.93	12.88	23.34	14.73	21.60	15.66	23.65
Matched		0.85	1.22	1.06	0.83	1.53	1.59	1.76
	Entries between 4 to 6 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	34	31	26	8	12	18	7
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	35	34	30	16	12	21	13
Matched		0	0	0	0	0	0	0
Unmatched	10%-level	37	34	33	24	19	25	18
Matched		0	0	0	0	0	1	0
	Mean standardized bias							
Unmatched		26.94	23.16	21.81	17.30	12.38	18.11	14.29
Matched		1.17	0.97	1.05	1.27	1.31	1.61	2.00
	Entries between 7 to 12 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	33	26	28	16	30	12	14
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	37	35	31	22	35	24	22
Matched		0	0	0	0	0	0	0
Unmatched	10%-level	40	39	35	26	43	31	25
Matched		0	0	0	0	0	1	0
	Mean standardized bias							
Unmatched		24.10	19.67	29.55	23.84	19.31	21.22	17.98
Matched		1.58	1.36	1.48	2.81	1.35	2.67	2.18

Note: For the t-test we conducted a simple t-test on the comparison of equal means. Depicted are the number of covariates with significant differences across the two groups, at the respective significance level. We included all variables in the analysis that were used in the respective PS-specifications - the baseline specification contains a total number of 53 covariates.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.5: Matching quality: balancing quality of IPW in West Germany —different indicators

Program		JS	STT	JWS	WS	FT	JCS	PT
	Entries between 1 to 3 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	28	31	22	17	28	24	41
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	31	35	29	25	33	31	43
matched		0	0	0	0	0	0	0
Unmatched	10%-level	33	36	36	27	35	37	45
Matched		0	0	0	0	0	0	0
	Mean standardized bias							
Unmatched		11.68	10.28	13.87	11.23	18.69	15.71	22.73
Matched		0.52	1.01	0.64	0.60	1.60	1.08	1.13
	Entries between 4 to 6 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	33	30	20	20	11	19	27
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	37	37	27	26	15	24	36
matched		0	0	0	0	0	0	0
Unmatched	10%-level	39	38	30	29	21	32	38
Matched		0	0	0	0	0	1	1
	Mean standardized bias							
Unmatched		18.31	17.93	17.17	20.54	14.68	21.52	25.43
Matched		0.71	0.60	1.05	1.23	1.81	1.78	1.15
	Entries between 7 to 12 months in unemployment							
	t-test on equal means							
Unmatched	1%-level	34	33	27	5	15	19	20
Matched		0	0	0	0	0	0	0
Unmatched	5%-level	38	36	32	7	22	22	25
Matched		0	0	0	0	0	0	0
Unmatched	10%-level	42	38	36	15	28	23	29
Matched		0	0	0	0	0	1	0
	Mean standardized bias							
Unmatched		19.58	20.01	26.60	15.43	13.81	19.22	19.19
Matched		0.93	0.61	1.75	2.63	1.00	1.79	1.36

Note: For the t-test we conducted a simple t-test on the comparison of equal means. Depicted are the number of covariates with significant differences across the two groups, at the respective significance level. We included all variables in the analysis that were used in the respective PS-specifications - the baseline specification contains a total number of 53 covariates.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.6: Number of observations by gender and pre-treatment schooling levels for program participants and non-participants

		By gender				By pre-treatment schooling level			
		East Germany		West Germany		East Germany		West Germany	
		M	W	M	W	Low	High	Low	High
JS	N	854	491	1,230	685	590	813	1,202	713
	%	63.49	36.51	64.23	35.77	42.05	57.95	62.77	37.23
STT	N	564	415	1,165	720	354	654	1187	749
	%	57.61	42.39	61.80	38.20	35.12	64.88	61.31	38.69
JWS	N	574	417	412	221	248	757	380	260
	%	57.92	42.08	65.09	34.91	24.68	75.32	59.38	40.63
WS	N	262	177	320	182	134	313	324	190
	%	59.68	40.32	63.75	36.25	29.98	70.02	63.04	36.96
JCS	N	473	207	400	170	416	268	500	79
	%	69.56	30.44	70.18	29.82	60.82	39.18	86.36	13.64
FT	N	282	127	343	172	146	266	317	212
	%	68.95	31.05	66.60	33.40	35.44	64.56	59.92	40.08
PT	N	301	209	627	385	319	194	766	253
	%	59.02	40.98	61.96	38.04	62.18	37.82	75.17	24.83
NP	N	7,367	4,752	15,926	8,690	3,767	8,157	14,890	11,871
	%	60.79	39.21	64.70	35.30	31.59	68.41	55.64	44.36

Source: Calculations are based on the estimation sample.
Note: Low levels of schooling indicate a lower secondary schooling degree levels or none; high levels of schooling indicate a medium or higher secondary schooling degree.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.7: Treatment effect heterogeneity by gender - selected monthly employment effects

		East Germany							West Germany						
	Month.../ Gender	1	6	12	24	36	48	60	1	6	12	24	36	48	60
JS	Men	**-2.67**	-0.96	3.07	**9.05**	**10.52**	**11.44**	**9.12**	**-3.03**	1.94	**6.44**	**5.89**	**7.06**	**6.41**	**3.93**
	(s.e.)	(0.25)	(1.50)	(1.65)	(1.88)	(1.80)	(1.70)	(1.85)	(0.19)	(1.24)	(1.33)	(1.35)	(1.38)	(1.37)	(1.41)
	Women	**-3.30**	-1.34	3.69	**8.81**	**5.33**	4.23	3.41	**-4.24**	-0.55	1.90	**7.14**	**5.75**	1.94	**3.89**
	(s.e.)	(0.38)	(2.06)	(2.13)	(2.28)	(2.26)	(2.32)	(2.39)	(0.36)	(1.95)	(1.78)	(2.03)	(1.96)	(1.89)	(1.80)
STT	Men	**-2.21**	-1.55	2.00	**7.93**	**6.40**	**10.80**	**11.41**	**-2.93**	0.42	**4.38**	**6.23**	**7.86**	**9.66**	**7.30**
	(s.e.)	(0.21)	(1.79)	(2.00)	(2.13)	(2.01)	(2.09)	(2.11)	(0.18)	(1.34)	(1.52)	(1.43)	(1.35)	(1.43)	(1.32)
	Women	**-2.72**	**-4.33**	1.38	**7.84**	**5.10**	4.46	**9.03**	**-3.86**	-1.26	**4.25**	**3.55**	3.04	**3.82**	2.35
	(s.e.)	(0.37)	(2.13)	(2.27)	(2.36)	(2.51)	(2.35)	(2.41)	(0.30)	(1.52)	(1.79)	(1.68)	(1.84)	(1.74)	(1.89)
JWS	Men	**-6.31**	**-20.24**	**-7.11**	**21.27**	**20.98**	**20.81**	**15.58**	**-3.70**	**-5.89**	**14.46**	**19.57**	**19.48**	**13.10**	**14.43**
	(s.e.)	(0.55)	(1.24)	(1.99)	(2.09)	(2.43)	(2.27)	(2.28)	(0.44)	(1.97)	(2.46)	(2.60)	(2.29)	(2.54)	(2.23)
	Women	**-6.84**	**-24.13**	**-9.52**	**25.69**	**26.63**	**18.85**	**13.91**	**-4.98**	**-6.60**	**22.09**	**21.28**	**15.23**	**11.81**	**13.50**
	(s.e.)	(0.72)	(1.81)	(2.36)	(2.70)	(2.77)	(2.67)	(2.73)	(0.64)	(3.00)	(3.52)	(3.35)	(3.14)	(3.33)	(3.14)
WS	Men	**-4.87**	**-9.34**	**6.13**	**16.33**	**18.90**	**14.84**	**13.15**	**-4.90**	2.00	**16.04**	**11.83**	**11.62**	**11.71**	**11.82**
	(s.e.)	(0.56)	(2.08)	(2.96)	(3.11)	(3.25)	(3.33)	(3.42)	(0.53)	(2.62)	(2.62)	(2.89)	(2.71)	(2.64)	(2.65)
	Women	**-6.55**	**-13.25**	4.59	**23.17**	**22.59**	**17.24**	**12.66**	**-6.29**	-3.00	**8.66**	6.11	1.70	5.22	1.64
	(s.e.)	(0.92)	(2.95)	(3.54)	(3.72)	(4.43)	(3.94)	(3.75)	(0.91)	(3.54)	(3.52)	(3.57)	(3.85)	(3.57)	(3.63)
JCS	Men	**-1.67**	**-8.69**	-2.95	-2.15	**-4.48**	-2.69	-0.13	**-2.33**	**-9.86**	**-4.64**	-1.16	-1.21	4.46	2.53
	(s.e.)	(0.28)	(1.32)	(1.63)	(1.67)	(1.94)	(2.30)	(2.36)	(0.36)	(1.60)	(1.89)	(1.98)	(2.13)	(2.46)	(2.54)
	Women	**-1.26**	**-7.07**	-2.26	**-6.48**	0.55	-5.27	0.56	**-2.51**	**-16.21**	**-6.79**	**-4.97**	-5.29	-6.18	-0.83
	(s.e.)	(0.44)	(2.06)	(2.59)	(2.31)	(3.33)	(3.08)	(3.26)	(0.56)	(1.95)	(2.79)	(2.50)	(2.84)	(3.25)	(3.41)
FT	Men	**-2.94**	**-10.04**	-0.08	**9.26**	**9.81**	**9.34**	**7.69**	**-4.29**	**-12.67**	-0.44	**11.64**	**8.08**	**13.44**	**9.18**
	(s.e.)	(0.37)	(2.27)	(2.71)	(3.18)	(3.27)	(3.14)	(3.17)	(0.46)	(2.06)	(2.72)	(2.70)	(2.66)	(2.42)	(2.63)
	Women	**-3.73**	**-9.89**	-3.45	6.48	**10.50**	5.32	6.56	**-3.82**	**-7.88**	3.61	**11.43**	**13.28**	**14.44**	6.19
	(s.e.)	(0.84)	(3.28)	(4.36)	(4.49)	(4.61)	(4.61)	(4.78)	(0.64)	(3.44)	(3.83)	(4.08)	(3.67)	(3.74)	(3.91)
PT	Men	**-1.22**	**-7.63**	**-7.22**	**-5.06**	**-7.65**	**-6.11**	**-5.28**	**-2.93**	**-11.95**	**-8.63**	**-7.13**	**-7.88**	-2.03	-0.56
	(s.e.)	(0.30)	(1.38)	(1.44)	(1.57)	(1.64)	(1.91)	(2.68)	(0.31)	(1.06)	(1.36)	(1.44)	(1.60)	(2.03)	(1.96)
	Women	**-1.26**	**-7.36**	**-6.69**	**-6.75**	**-8.10**	**-5.95**	-3.29	**-2.10**	**-11.75**	**-9.14**	**-4.97**	**-6.34**	-0.50	1.98
	(s.e.)	(0.50)	(2.26)	(1.87)	(1.97)	(2.01)	(2.56)	(2.93)	(0.29)	(1.78)	(1.86)	(1.86)	(1.71)	(2.15)	(2.41)
PT$^{(1)}$	Men	**-9.70**	**-14.40**	0.23	**11.47**	**10.06**	**7.93**	**4.99**	**-7.01**	**-6.19**	**6.88**	**13.61**	**11.88**	**3.98**	1.82
	(s.e.)	(0.95)	(1.66)	(2.80)	(3.06)	(2.99)	(2.62)	(2.39)	(0.50)	(1.17)	(1.75)	(1.83)	(1.81)	(1.66)	(1.36)
	Women	**-10.69**	**-12.95**	3.60	**12.36**	**11.48**	-2.62	1.42	**-7.56**	**-7.94**	**7.46**	**11.11**	**8.77**	-0.07	-1.72
	(s.e.)	(1.33)	(2.39)	(2.99)	(3.33)	(3.46)	(2.68)	(2.36)	(0.70)	(1.70)	(2.41)	(2.29)	(1.95)	(1.76)	(1.66)

Note: Depicted are monthly ATT estimates on employment probabilities. $^{(1)}$ refers to the ATT estimates on the education probabilities. The ATT are written in bold when they are significant at the 5%-level. Standard errors are obtained by bootstrapping with 200 replications and are depicted in parentheses.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.8: Treatment effect heterogeneity by gender - cumulated effects after 30 and 60 months

		East Germany				West Germany			
		Employment		Education		Employment		Education	
	$\sum$/ Gender	30	60	30	60	30	60	30	60
JS	Men	**1.39**	**4.19**	**-0.91**	**-1.51**	**1.39**	**3.06**	**-1.01**	**-1.52**
	(s.e)	(0.35)	(0.68)	(0.19)	(0.36)	(0.26)	(0.54)	(0.17)	(0.30)
	Women	**1.44**	**2.74**	**-1.53**	**-1.85**	**1.16**	**2.15**	**-0.98**	**-1.21**
	(s.e)	(0.44)	(0.86)	(0.26)	(0.49)	(0.41)	(0.79)	(0.26)	(0.44)
STT	Men	**1.23**	**4.02**	**-1.29**	**-1.82**	**1.14**	**3.56**	**-1.14**	**-1.56**
	(s.e)	(0.40)	(0.81)	(0.23)	(0.41)	(0.29)	(0.55)	(0.18)	(0.31)
	Women	**1.00**	**2.71**	**-1.42**	**-1.53**	**0.67**	**1.42**	**-0.78**	-0.83
	(s.e)	(0.50)	(0.90)	(0.31)	(0.58)	(0.34)	(0.68)	(0.26)	(0.47)
JWS	Men	**2.70**	**8.49**	**-2.14**	**-3.63**	**4.13**	**8.96**	**-2.21**	**-3.42**
	(s.e)	(0.39)	(0.84)	(0.18)	(0.38)	(0.46)	(0.91)	(0.18)	(0.38)
	Women	**3.24**	**9.44**	**-2.65**	**-3.73**	**4.31**	**9.42**	**-2.32**	**-3.30**
	(s.e)	(0.48)	(0.96)	(0.25)	(0.44)	(0.62)	(1.30)	(0.34)	(0.66)
WS	Men	**3.42**	**8.28**	**-1.66**	**-3.11**	**2.89**	**6.03**	**-1.27**	**-2.32**
	(s.e)	(0.57)	(1.25)	(0.28)	(0.49)	(0.52)	(1.03)	(0.29)	(0.48)
	Women	**4.23**	**9.77**	**-3.09**	**-4.23**	**1.43**	2.28	**-1.42**	**-1.81**
	(s.e)	(0.66)	(1.46)	(0.40)	(0.70)	(0.76)	(1.46)	(0.45)	(0.88)
JCS	Men	**-1.36**	**-2.26**	**-1.36**	**-1.38**	**-0.99**	-0.57	**-0.81**	-0.66
	(s.e)	(0.30)	(0.73)	(0.28)	(0.52)	(0.34)	(0.78)	(0.29)	(0.56)
	Women	**-1.46**	**-2.16**	**-2.05**	**-2.87**	**-2.28**	**-3.86**	**-1.30**	-0.99
	(s.e)	(0.47)	(1.05)	(0.49)	(0.81)	(0.49)	(0.99)	(0.47)	(0.99)
FT	Men	0.64	**3.57**	**-1.91**	**-3.18**	0.84	**3.85**	**-1.85**	**-2.48**
	(s.e)	(0.56)	(1.18)	(0.30)	(0.55)	(0.50)	(0.94)	(0.27)	(0.52)
	Women	-0.26	1.66	**-2.29**	**-2.98**	**1.79**	**5.37**	**-2.28**	**-3.08**
	(s.e)	(0.85)	(1.78)	(0.41)	(0.72)	(0.80)	(1.57)	(0.39)	(0.74)
PT	Men	**-1.70**	**-3.42**	0.60	**2.85**	**-2.23**	**-3.55**	**1.65**	**3.74**
	(s.e)	(0.30)	(0.64)	(0.63)	(1.03)	(0.26)	(0.59)	(0.36)	(0.62)
	Women	**-1.66**	**-3.54**	0.62	**1.89**	**-2.06**	**-2.69**	**1.26**	**2.33**
	(s.e)	(0.41)	(0.79)	(0.64)	(1.06)	(0.38)	(0.75)	(0.50)	(0.76)

Note: Depicted are the cumulated treatment effects, summing up the monthly ATT between for 30 or 60 months following treatment entry. The effects are written in bold when they are significant at the 5%-level. Standard errors are obtained by bootstrapping with 200 replications and are depicted in parentheses.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.9: Treatment effect heterogeneity by pretreatment schooling - selected monthly employment effects.

		East Germany							West Germany						
	Month.../ Education	1	6	12	24	36	48	60	1	6	12	24	36	48	60
JS	Low	**-2.31**	0.11	**4.47**	**7.01**	**9.26**	**6.91**	**5.34**	**-2.80**	0.75	**4.17**	**4.96**	**4.76**	**3.04**	**3.31**
	(s.e.)	(0.29)	(1.69)	(1.90)	(1.96)	(1.97)	(1.98)	(2.05)	(0.21)	(1.35)	(1.33)	(1.29)	(1.48)	(1.36)	(1.51)
	High	**-3.37**	2.00	**6.78**	**14.38**	**11.99**	**13.02**	**9.58**	**-4.68**	1.62	**6.79**	**8.74**	**9.83**	**8.07**	**6.04**
	(s.e.)	(0.27)	(1.42)	(1.78)	(1.98)	(1.96)	(1.94)	(1.93)	(0.40)	(1.78)	(1.81)	(1.90)	(1.91)	(2.04)	(2.08)
STT	Low	**-1.75**	-1.37	1.46	**7.38**	4.75	**6.40**	**11.84**	**-2.70**	2.18	**6.21**	**6.15**	**6.70**	**9.07**	**7.13**
	(s.e.)	(0.23)	(1.61)	(1.92)	(2.43)	(2.55)	(2.65)	(2.87)	(0.16)	(1.15)	(1.36)	(1.31)	(1.46)	(1.47)	(1.48)
	High	**-2.66**	0.53	**5.90**	**12.37**	**10.17**	**12.20**	**10.68**	**-3.75**	**4.17**	**8.50**	**10.04**	**9.19**	**8.34**	**5.54**
	(s.e.)	(0.24)	(1.65)	(1.90)	(2.05)	(2.12)	(2.11)	(2.03)	(0.28)	(1.71)	(1.70)	(1.72)	(1.70)	(1.81)	(1.65)
JWS	Low	**-5.22**	**-15.26**	0.46	**25.02**	**24.21**	**22.63**	**17.74**	**-3.37**	-2.32	**16.23**	**15.95**	**13.13**	**12.34**	**12.33**
	(s.e.)	(0.67)	(1.86)	(2.93)	(3.40)	(3.59)	(3.51)	(3.46)	(0.43)	(2.30)	(2.61)	(2.35)	(2.36)	(2.42)	(2.21)
	High	**-6.80**	**-19.95**	**-7.76**	**25.27**	**24.93**	**20.05**	**14.89**	**-4.76**	**-9.10**	**20.40**	**27.27**	**25.83**	**12.89**	**15.39**
	(s.e.)	(0.45)	(0.98)	(1.63)	(2.09)	(1.90)	(1.86)	(1.72)	(0.71)	(2.59)	(3.25)	(3.31)	(3.02)	(3.13)	(3.07)
WS	Low	**-3.18**	**-6.41**	5.53	**17.45**	**14.86**	6.66	7.96	**-4.09**	2.33	**12.96**	**8.66**	**7.95**	**8.87**	**9.88**
	(s.e.)	(0.56)	(2.70)	(3.55)	(3.72)	(3.37)	(3.66)	(4.24)	(0.42)	(2.21)	(2.62)	(2.61)	(2.37)	(2.43)	(2.61)
	High	**-6.81**	**-11.26**	**7.98**	**20.09**	**22.46**	**18.94**	**15.37**	**-6.98**	5.47	**23.01**	**19.58**	**15.20**	**15.64**	**9.06**
	(s.e.)	(0.61)	(2.03)	(2.81)	(2.70)	(2.79)	(2.88)	(2.77)	(0.69)	(3.43)	(3.54)	(3.77)	(3.74)	(3.74)	(3.85)
JCS	Low	**-1.23**	**-6.63**	**-3.67**	**-3.25**	**-3.97**	-2.97	-2.77	**-2.09**	**-9.17**	**-3.73**	-2.61	-1.90	2.53	1.99
	(s.e.)	(0.23)	(1.00)	(1.44)	(1.51)	(1.80)	(2.22)	(2.24)	(0.31)	(1.08)	(1.45)	(1.46)	(1.56)	(1.87)	(2.09)
	High	**-1.55**	**-5.49**	1.42	-2.07	-0.38	-4.34	2.87	**-2.48**	**-7.57**	-3.05	5.30	-3.25	-5.14	-5.75
	(s.e.)	(0.25)	(1.54)	(2.39)	(2.23)	(2.66)	(2.69)	(2.98)	(0.58)	(2.89)	(4.23)	(5.39)	(4.80)	(6.02)	(6.35)
FT	Low	**-2.88**	**-7.08**	1.78	2.91	2.76	2.95	3.02	**-3.18**	**-7.14**	-0.19	**11.64**	**8.33**	**14.50**	**8.86**
	(s.e.)	(0.56)	(2.28)	(3.88)	(4.08)	(4.17)	(3.85)	(4.03)	(0.36)	(2.07)	(2.32)	(2.98)	(3.05)	(2.80)	(2.60)
	High	**-2.98**	**-5.94**	0.74	**14.26**	**16.25**	**14.63**	**12.23**	**-4.90**	**-11.55**	**8.42**	**16.64**	**14.48**	**15.27**	**8.80**
	(s.e.)	(0.41)	(2.46)	(2.73)	(3.11)	(3.00	(3.14)	(3.10)	(0.65)	(2.52)	(3.23)	(3.50)	(3.47)	(3.46)	(3.60)
PT	Low	**-0.81**	**-4.19**	**-5.02**	**-4.17**	**-5.27**	**-4.71**	-3.34	**-2.16**	**-8.94**	**-6.78**	**-4.49**	**-5.56**	-1.76	-0.37
	(s.e.)	(0.20)	(1.02)	(1.07)	(1.39)	(0.99)	(1.51)	(2.25)	(0.19)	(0.81)	(1.08)	(1.26)	(1.30)	(1.42)	(1.70)
	High	**-1.40**	**-6.53**	**-6.34**	**-6.20**	**-10.84**	**-7.28**	-4.37	**-2.77**	**-11.32**	**-7.90**	**-7.86**	**-8.76**	0.28	3.66
	(s.e.)	(0.27)	(1.37)	(1.43)	(1.79)	(1.99)	(2.68)	(3.44)	(0.36)	(1.22)	(2.01)	(1.98)	(2.44)	(3.34)	(3.25)
PT$^{(1)}$	Low	**-7.83**	**-12.80**	**2.15**	**10.21**	**10.65**	**5.65**	3.07	**-6.04**	**-5.77**	**5.07**	**10.99**	9.98	**2.97**	1.99
	(s.e.)	(0.79)	(1.32)	(2.24)	(2.68)	(2.79)	(2.21)	(2.08)	(0.34)	(0.98)	(1.46)	(1.63)	(1.59)	(1.27)	(1.25)
	High	**-12.79**	**-14.72**	2.18	**15.17**	**12.13**	1.44	3.47	**-12.01**	**-11.44**	**12.02**	**16.57**	13.26	2.13	-2.94
	(s.e.)	(1.21)	(2.67)	(3.82)	(3.52)	(3.18)	(2.94)	(2.88)	(0.95)	(2.26)	(2.97)	(3.25)	(3.19)	(2.59)	(2.05)

Note: Depicted are monthly ATT estimates on employment probabilities. $^{(1)}$ refers to the ATT estimates on the education probabilities. Low levels of schooling indicate a lower secondary schooling qualification or none; high levels of schooling indicate a medium or higher secondary schooling qualification. Depicted are the average treatment effects (ATT) on the employment probabilities in the months following treatment entry. The ATT are written in bold when they are significant at the 5%-level. Standard errors are obtained by bootstrapping with 200 replications and are depicted in parentheses.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.10: Treatment effect heterogeneity by pretreatment schooling - cumulated effects after 30 and 60 months

		East Germany				West Germany			
		Employment		Education		Employment		Education	
	Σ/ Education	30	60	30	60	30	60	30	60
JS	Low	**1.37**	**3.24**	**-0.57**	-0.71	**1.04**	**2.07**	**-0.73**	**-0.85**
	(s.e.)	(0.40)	(0.81)	(0.19)	(0.37)	(0.28)	(0.54)	(0.15)	(0.28)
	High	**2.58**	**5.90**	**-2.52**	**-3.61**	**1.92**	**4.22**	**-1.38**	**-2.27**
	(s.e.)	(0.36)	(0.73)	(0.21)	(0.37)	(0.38)	(0.79)	(0.26)	(0.41)
STT	Low	**1.24**	**3.24**	**-0.87**	**-1.26**	**1.36**	**3.59**	**-0.94**	**-1.07**
	(s.e.)	(0.41)	(0.89)	(0.30)	(0.50)	(0.25)	(0.52)	(0.16)	(0.29)
	High	**2.16**	**5.45**	**-2.61**	**-3.36**	**2.19**	**4.45**	**-2.18**	**-3.27**
	(s.e.)	(0.41)	(0.87)	(0.26)	(0.47)	(0.34)	(0.68)	(0.29)	(0.47)
JWS	Low	**4.00**	**10.49**	**-1.17**	**-1.68**	**3.83**	**8.01**	**-1.53**	**-2.53**
	(s.e.)	(0.56)	(1.26)	(0.23)	(0.40)	(0.44)	(0.87)	(0.19)	(0.40)
	High	**3.44**	**9.63**	**-3.54**	**-5.50**	**5.73**	**10.94**	**-3.84**	**-5.16**
	(s.e.)	(0.34)	(0.70)	(0.16)	(0.31)	(0.57)	(1.21)	(0.30)	(0.63)
WS	Low	**3.39**	**6.66**	**-1.38**	**-2.03**	**2.29**	**4.65**	**-1.14**	**-2.05**
	(s.e.)	(0.65)	(1.37)	(0.32)	(0.65)	(0.45)	(0.90)	(0.24)	(0.44)
	High	**4.22**	**10.09**	**-3.20**	**-5.15**	**4.85**	**8.83**	**-2.62**	**-3.99**
	(s.e.)	(0.52)	(1.06)	(0.28)	(0.45)	(0.79)	(1.52)	(0.48)	(0.82)
JCS	Low	**-1.35**	**-2.46**	**-1.40**	**-1.85**	**-0.99**	-1.00	**-0.98**	-0.90
	(s.e.)	(0.23)	(0.59)	(0.26)	(0.50)	(0.26)	(0.55)	(0.25)	(0.49)
	High	-0.74	-1.22	**-3.04**	**-3.62**	-0.47	-1.20	**-3.57**	**-3.59**
	(s.e.)	(0.43)	(0.96)	(0.39)	(0.75)	(0.83)	(1.99)	(0.95)	(1.65)
FT	Low	0.02	0.89	**-1.21**	**-1.98**	**1.11**	**4.16**	**-1.52**	**-2.01**
	(s.e.)	(0.65)	(1.39)	(0.30)	(0.51)	(0.51)	(1.07)	(0.22)	(0.50)
	High	**1.47**	**5.89**	**-3.15**	**-4.76**	**2.64**	**6.69**	**-2.75**	**-3.87**
	(s.e.)	(0.55)	(1.16)	(0.30)	(0.57)	(0.66)	(1.39)	(0.42)	(0.72)
PT	Low	**-1.10**	**-2.30**	0.72	**2.73**	**-1.61**	**-2.50**	**1.28**	**2.98**
	(s.e.)	(0.21)	(0.44)	(0.48)	(0.84)	(0.20)	(0.43)	(0.31)	(0.52)
	High	**-1.64**	**-4.20**	0.66	**2.60**	**-2.17**	**-3.07**	**1.86**	**3.69**
	(s.e.)	(0.31)	(0.72)	(0.77)	(1.11)	(0.38)	(0.89)	(0.65)	(1.06)

Note: Low levels of schooling indicate a lower secondary schooling qualification or none; high levels of schooling indicate a medium or higher secondary schooling qualification. Depicted are the cumulated treatment effects, summing up the monthly ATT between for 30 or 60 months following treatment entry. The effects are written in bold when they are significant at the 5%-level. Standard errors are obtained by bootstrapping with 200 replications and are depicted in parentheses.
Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Table A3.11: Sensitivity of the employment effect estimates

East Germany														
	JS		STT		JWS		WS		JCS		FT		PT	
Σ	30	60	30	60	30	60	30	60	30	60	30	60	30	60
Results from the main analysis														
ATT	**1.49**	**3.81**	**1.27**	**3.65**	**3.10**	**9.09**	**3.53**	**8.49**	**-1.47**	**-2.38**	0.27	**2.86**	**-1.64**	**-3.43**
(s.e)	(0.25)	(0.54)	(0.31)	(0.57)	(0.31)	(0.62)	(0.49)	(1.02)	(0.25)	(0.56)	(0.44)	(0.98)	(0.20)	(0.43)
A) Further program participation														
ATT	**2.16**	**3.92**	**1.53**	**3.61**	**3.33**	**9.37**	**4.09**	**9.55**	**-1.37**	**-2.32**	0.43	**2.75**	**-1.57**	**-3.30**
(s.e)	(0.32)	(0.61)	(0.36)	(0.70)	(0.32)	(0.64)	(0.47)	(0.95)	(0.26)	(0.60)	(0.48)	(1.04)	(0.24)	(0.53)
B) Dynamic evaluation approach														
ATT	**1.70**	**3.95**	**1.48**	**3.81**	**3.31**	**9.09**	**3.78**	**8.57**	**-1.31**	**-2.22**	0.44	**2.91**	**-1.44**	**-3.19**
(s.e)	(0.23)	(0.52)	(0.27)	(0.56)	(0.27)	(0.57)	(0.44)	(0.88)	(0.24)	(0.54)	(0.41)	(0.90)	(0.20)	(0.46)
C) Alternative imposition of common support														
C1) ATT	**1.62**	**4.15**	**1.28**	**3.58**	**3.32**	**9.53**	**3.61**	**8.65**	**-1.61**	**-2.71**	0.32	**2.99**	**-1.63**	**-3.56**
(s.e)	(0.28)	(0.56)	(0.30)	(0.56)	(0.31)	(0.60)	(0.45)	(0.92)	(0.26)	(0.58)	(0.44)	(0.88)	(0.20)	(0.46)
C2) ATT	**1.19**	**3.42**	**1.03**	**3.54**	**2.66**	**8.52**	**3.46**	**8.75**	**-2.07**	**-3.40**	0.03	**2.56**	**-1.78**	**-3.63**
(s.e)	(0.28)	(0.59)	(0.37)	(0.73)	(0.37)	(0.69)	(0.55)	(1.17)	(0.30)	(0.64)	(0.53)	(1.19)	(0.30)	(0.63)
C3) ATT	**1.73**	**4.39**	**1.52**	**3.99**	**3.32**	**9.22**	**3.59**	**8.82**	**-1.49**	**-2.44**	0.26	**3.08**	**-1.63**	**-3.18**
(s.e)	(0.33)	(0.64)	(0.32)	(0.62)	(0.31)	(0.70)	(0.47)	(0.98)	(0.30)	(0.71)	(0.62)	(1.22)	(0.29)	(0.63)
West Germany														
	JS		STT		JWS		WS		JCS		FT		PT	
Σ	30	60	30	60	30	60	30	60	30	60	30	60	30	60
Results from the main analysis														
ATT	**1.37**	**2.85**	**0.98**	**2.75**	**4.16**	**8.53**	**2.42**	**4.92**	**-1.38**	**-1.63**	**1.23**	**4.47**	**-2.14**	**-3.09**
(s.e.)	(0.22)	(0.42)	(0.23)	(0.45)	(0.38)	(0.71)	(0.47)	(0.86)	(0.30)	(0.64)	(0.44)	(0.83)	(0.20)	(0.42)
A) Further program participation														
ATT	**2.43**	**4.29**	**1.57**	**3.22**	**4.49**	**9.09**	**2.97**	**5.33**	**-1.15**	-1.13	**1.32**	**4.65**	**-2.07**	**-2.90**
(s.e.)	(0.25)	(0.49)	(0.26)	(0.48)	(0.36)	(0.72)	(0.50)	(0.98)	(0.31)	(0.66)	(0.41)	(0.84)	(0.21)	(0.47)
B) Dynamic evaluation approach														
ATT	**1.52**	**2.95**	**1.13**	**2.91**	**4.16**	**8.44**	**2.50**	**4.93**	**-1.20**	**-1.34**	**1.28**	**4.46**	**-1.92**	**-2.85**
(s.e.)	(0.21)	(0.42)	(0.22)	(0.42)	(0.32)	(0.61)	(0.43)	(0.86)	(0.30)	(0.60)	(0.41)	(0.86)	(0.18)	(0.41)
C) Alternative imposition of common support														
C1) ATT	**1.44**	**2.95**	**1.02**	**2.78**	**4.24**	**8.68**	**2.49**	**5.00**	**-1.50**	**-1.93**	**1.29**	**4.61**	**-2.17**	**-3.17**
(s.e.)	(0.21)	(0.42)	(0.21)	(0.43)	(0.36)	(0.69)	(0.44)	(0.84)	(0.28)	(0.62)	(0.44)	(0.85)	(0.20)	(0.46)
C2) ATT	**1.09**	**2.43**	**0.70**	**2.39**	**3.76**	**8.25**	**1.83**	**3.87**	**-1.79**	**-2.23**	0.83	**4.04**	**-2.13**	**-3.06**
(s.e.)	(0.28)	(0.54)	(0.27)	(0.53)	(0.42)	(0.82)	(0.52)	(0.98)	(0.36)	(0.78)	(0.48)	(0.98)	(0.25)	(0.51)
C3) ATT	**1.78**	**3.49**	**1.27**	**3.43**	**4.20**	**8.60**	**2.72**	**5.27**	**-1.15**	**-1.40**	**1.37**	**4.68**	**-1.96**	**-2.82**
(s.e.)	(0.30)	(0.55)	(0.25)	(0.48)	(0.35)	(0.73)	(0.44)	(0.86)	(0.35)	(0.81)	(0.45)	(0.89)	(0.23)	(0.53)

Note: The cumulative effects are obtained by summing up the monthly program effects over a period of 30 or 60 months after program entry. Standard errors in parentheses are obtained by bootstrapping the estimation procedure with 200 replications. Bold numbers indicate significance at the 5% level. The results from the main analysis are the aggregate cumulative effects from Table 3.5.

Sensitivity A) refers to the exclusion of further program participants within one year of unemployment duration.

Sensitivity B) refers to the extension of the control group to all future program participants and other program participants.

Sensitivity C) refers to modifications in the PS distribution that is used to weigh the nonparticipant outcomes. We estimate the effects in C1) by excluding non-participants with PS-values above the 99th percentile. In C2) we only include participants and non-participants in the analysis within the optimal region of common support: $\alpha < P(W) < (1-\alpha)$ as suggested by Crump et al. (2009). For C3) we divide the PS-distribution in 20 equidistant percentiles, and only estimate the ATT in regions where the density is above 5% ($F(P(W) > 5\%)$ in both groups.

Abbreviation index: JS: job search assistance; STT: short-term training; JCS: Job creation schemes; JWS: *JUMP* wage subsidies; WS: *SGB III* wage subsidies; FT: further training (medium to long-term); PT: preparatory training; NP: non-participants.

Chapter 4

More Practical Guidance for the Implementation of Matching and Weighting Estimators*

4.1 Introduction

Balancing the characteristics between two population subgroups is an empirical exercise of high practical relevance. To identify the causal effect of a treatment in the absence of random treatment assignment, conditional independence in outcomes between treatment groups can be established by balancing the relevant pre-treatment characteristics (e.g., Imbens, 2004). Even without the claim of causality, balancing the observable determinants of an outcome across two population groups may help answer the question to what extent outcome differences between these groups are related to group-specific unobservables, as is done in decomposition analysis (Fortin et al., 2011). Beyond the objective of removing the influence of observable characteristics on an outcome, balancing the characteristics across two subgroups might be beneficial in the design of empirical studies (Rubin, 2007), may increase the robustness of parametric outcome analysis (Ho et al., 2007), or may improve the power of instrumental variable approaches (Frölich, 2007a; Baiocchi et al., 2010).

In these and similar settings, semi-parametric matching and weighting on the propensity score are frequently applied. Matching on the propensity score

*This chapter is based on a joint paper with Marco Caliendo.

(PSM) was pioneered by Rosenbaum and Rubin (1983b), who showed that matching individuals on a one-dimensional summary score of characteristics instead of the potentially high-dimensional matrix of characteristic combinations will also lead to the balancing of characteristics across two subgroups. Their summary score is derived as the conditional probability of receiving treatment, with the unbalanced characteristics as explanatory variables. Matching on the propensity score is identical to reweighing observations based on their propensity score value, thus exhibiting great similarity with the concept of inverse probability weighting (IPW) encountered in the non-random sampling literature (Horvitz and Thompson, 1952). Here, the reweighing of control group members with an increasing function of the treatment probability redistributes weights from individuals who are less similar to the group of treated towards individuals who are more similar to them, thereby creating balance in characteristics.

Using balancing via PSM and IPW for the construction of counterfactual outcomes in the estimation of treatment effects has the advantage that missing overlap in the distributions of characteristics becomes evident and manageable, and avoids implicit extrapolations that often go unnoticed in parametric regression analyses (Cochran, 1957). Strong non-linearities in the relation between characteristics and outcomes that are a problem to parametric regression can be more flexibly accommodated (Basu et al., 2008). Furthermore in the case of binary outcomes, balancing on the propensity score may be more reliable than logistic regression when the number of non-zero outcomes is low (Cepeda et al., 2003). By their versatility, PSM and IPW are amply applied in diverse fields of empirical research, e.g., labor economics (Frölich, 2007b), health economics (Schreyögg et al., 2011), political sciences (Eggers and Hainmueller, 2009; Boyd et al., 2010), neurology (Saposnik et al., 2012), business administration (Armstrong et al., 2010), medical research (Austin and Mamdani, 2006), etc.

In the practical implementation, PSM and IPW unfortunately do not come as one-fits-all methods, but have to be adapted to the data at hand to achieve the required balance in characteristics. Implementation choices, such as the correct specification of the treatment model, the detection and elimination of overlap and the choice of an appropriate matching or weighting scheme have to be done manually by the researcher. Unlike in parametric regression analysis where the optimal fit is achieved via automated optimization of a least squares problem or a likelihood function, implementation of weighting or matching requires that balance

is assessed manually after each implementation step. For each implementation step the literature offers a large and oftentime competing array of guidelines. A large methodological literature has emerged addressing questions regarding the optimal choices implementation strategy when using these methods.

Against this background, the objective of this paper is twofold. The first aim is to provide a comprehensive summary of the practical guidelines currently available for the implementation of propensity score balancing methods. We keep the overview mainly non-technical to bridge the gap between theoretical and applied knowledge and to reduce uncertainty regarding the correct choice and application of these methods for applied researchers. In this respect we pick up the line of thought of Caliendo and Kopeinig (2008). By incorporating the findings of diverse fields of applied research we aim to facilitate the exchange between these strands of the literature, as they often work independently on similar problems. The second aim of this paper is to provide guidance on the estimation of conditional outcome differences using the balancing weights. We outline the prerequisites for a causal interpretation of the conditional outcomes, and present alternative estimation methods that may increase the robustness of estimates compared to the most commonly applied conditional differences in means estimator. While most applications address the calculation of balancing weights and the outcome analyses jointly, this paper deals with these two issues separately. With this it is underscored that the balancing questions need to be addressed independently of the identification questions.

The chapter is set up as follows. In Chapter 4.2 we start with a theoretical motivation for balancing observed characteristics in the context of the construction of counterfactual outcomes, with and without conditional independence, and outline the assumptions required for balancing on the propensity score. Based on a schematic overview of the practical steps to be taken when calculating balancing weights with PSM and IPW (see Table 4.1), we proceed by discussing them successively in Chapters 4.3 to 4.7. In general, these steps are rather similar for both methods, so they will be discussed jointly and differences are highlighted if they exist. For researchers who intend to use the balancing weights to estimate and compare conditional outcomes, Chapter 4.8 addresses the estimation of outcome differences and addresses issues as variance estimation, and sensitivity analyses to corroborate the conditional independence hypothesis. Here, we also outline avenues for the combination of PSM and IPW with DID and IV methods, as this

might further strengthen the assumption of conditional exogeneity, and review possible combination of the semi-parametric balancing with parametric outcome analyses that might increase the robustness of the conditional outcome differences. In the final Chapter 4.9, we discuss recent developments in the matching literature and outline extension in the literature addressing multiple treatments and dynamic treatment assignment. Chapter 4.10 concludes. The summary is complemented by Table A4.1 outlining the currently available software in `stata` or `R` implementing the estimation steps/methods mentioned in the text.

4.2 Theoretical Framework and Implementation Steps

In the following we outline the estimation set-up when aiming to balance characteristics across two population or treatment groups, and discuss the assumptions required to justify matching and weighting in the estimation of conditional outcome differences, with or without a causal interpretation. For simplicity we assume that the population of interest is divided into two mutually exclusive subgroups, with $D = (0, 1)$ representing a binary treatment or group identifier.[1] For each individual i we observe the group (treatment) status D_i and a set of characteristics X_i. If the group-status is not assigned randomly, the distribution of X_i generally differs across groups, so that $E(D_i|X_i) \neq E(D_i)$. The conditional probability of treatment is given by $p(X_i) = Pr(D_i = 1|X_i)$. We further observe an outcome of interest Y_i that is a function of the observed characteristics X_i, as well as of unobserved characteristics U_i, and assume that m_D represents the function linking the observed and unobserved characteristics to the outcome, e.g., the return function (also see Fortin et al., 2011),

$$Y_{0i} = m_0(X_i, U_i) \text{ and } \; Y_{1i} = m_1(X_i, U_i). \tag{4.1}$$

By this representation it can be seen that the difference in outcomes of the two subgroups, $\Delta_i = Y_{1i} - Y_{0i}$ is given by differences in either m_D, X or U. The idea of

[1] In the following we refer to the two subgroups as treated ($D = 1$) and controls ($D = 0$) throughout the paper, but they could also refer to binary population characteristics such as, e.g., gender or migration background. Note also, that the framework can be easily extended to more than two subgroups. We discuss the application of PSM and IPW for multiple-valued treatments in Section 4.9.2.

balancing is that the differences arising from X can be eliminated by establishing balance in the observable characteristics across groups.

Propensity score balancing methods provide a common tool to do so. By either matching on the conditional treatment probability $p(X_i)$, or weighting by the inverse of $p(X_i)$, balance in characteristics can be established. In the following we focus on the alignment of the distribution of X_{0i} in the group of controls to that of the characteristics in the group of treated X_{1i}. This allows to identify common parameters of interest, both in the evaluation literature – the effect of treatment on the treated – and the decomposition literature – the average difference attributable to differences in the return function[2]. Based on aligning the characteristic distribution in the two treatment groups, the outcomes of the treated can be compared with the outcomes of the balanced controls. A common parameter of interest is the *average* outcome difference, for which the average reweighed control outcome is given by $\bar{Y}^C_{PSM} = E_{X|D=1}[Y_{0i} \mid p(X_i)]$. Alternatively, the density, as well as the cumulative distribution function of Y can be estimated (Frölich, 2007b). We focus the subsequent discussion on the average effect.

Before outlining the two approaches to balance characteristics, we need to emphasize the importance of the overlap condition when aligning characteristics distribution. In particular, it is required that all characteristic values appearing in the treatment group also appear in the control group. Stated differently, this condition fails when certain characteristic values deterministically imply participation in the treatment. Hence, the overlap condition can be expressed as follows,

$$S_X = \{X \mid Pr(D = 1 \mid X) \; < \; 1\}. \qquad (Overlap)$$

In particular, when some characteristic combinations are only observable in the treatment group one is unable to infer how they will actually relate to outcomes in the control group, as $m_0(\cdot)$ and U_0 are unknown. The lack of overlap might risk identification of causal effects as incomparable individuals in terms of their observable characteristics are likely to exhibit unusual or extreme characteristic combinations, and are hence also more likely to systematically differ in terms of their unobservable characteristics — the additional regularity assumptions re-

[2]Note, that we focus on the so-called aggregate decomposition analysis that looks at the joint compositional effect of observables. DiNardo et al. (1996) and Fortin et al. (2011) discuss the use of IPW for recovering the "contribution" of single binary observable characteristics. A similar logic could be applied to PSM.

quired for identification are difficult to justify (see, e.g., Heckman et al., 1997; Ñopo, 2008).

4.2.1 Propensity Score Matching

In a seminal paper Rosenbaum and Rubin (1983b) show that the conditional treatment probability $p(X_i)$ is a "balancing score" in the sense that instead of conditioning on individual characteristics X_i for achieving balance, conditioning on this "propensity score", may yield balance in characteristics X_i across treatment status,

$$X_i \perp\!\!\!\perp D_i \mid p(X_i). \tag{4.2}$$

While this result may be used when aiming to create balance in X_i across treatment groups, Rosenbaum and Rubin (1983b) further show that this balancing property can be used to justify matching on the propensity score for removing the influence of X_i on Δ_i. Note, that this result is based on the law iterated expectations (Frölich, 2007b), and hence does not require conditional independence (see below in Section 4.2.3). Let ω_{ij}^{PSM} denote the balancing weights that are derived as a function of the distance between the propensity score values of treated and controls $\|p(X_i) - p(X_j)\|, i : D = 1; j : D = 0$, and $\sum_{j:D=0}^{N_0} \omega_{ij} = 1$. Matching estimators may differ in the maximally allowed distance or number of propensity scores to be matched, i.e., the definition of the "neighborhood", and how the scores within this neighborhood are aggregated. Note that in matching, $p(X_i)$ acts merely as a summary measure for combinations of X_i, so that consistent estimate of the treatment probability is not needed.

Based on the calculation of ω_{ij}^{PSM}, the conditional average control outcome $\bar{Y}_{0i}^{C}$, is calculated on the reweighed control sample. It can be shown that this amounts to a reweighting of control outcomes (Smith and Todd, 2005a; Busso et al., 2014a),

$$\bar{Y}_{PSM}^{C} = E_{X|D=1}[Y_{0i} \mid p(X_i)] = \sum_{i \in S_X, D_i = 0} \omega_i^{PSM} Y_{0i}, \tag{4.3}$$

using the weights $\omega_i^{PSM} = \frac{N_1}{N_0} \sum_i^{N_1} \omega_{ij}$, with $N_D, D = 0, 1$ representing the size of the respective treatment gorups.

4.2.2 Inverse Probability Weighting

The use of inverse probability weighting for balancing characteristics goes back to the literature on attrition and sample selection (see, e.g., Horvitz and Thompson, 1952; Wooldridge, 2002), as well as missing data problems (Robins et al., 1994), and has recently found its way to applied problems in the treatment evaluation literature (Hirano et al., 2003) and decomposition analysis (DiNardo et al., 1996). The idea of missing data imputation is to reweigh the observed sampling distribution (here, the controls) by the sampling probability (the treatment probability $p(X_i)$) to retrieve the correct distribution parameters for the target population (here, the treated). The balancing effect of IPW is also based on the law of iterated probabilities (see, e.g., Fortin et al., 2011). THe balancing weights are given as a direct function of the propensity score, i.e., with $\omega_i^{IPW} = \frac{p(X_i)}{1-p(X_i)}\frac{\pi}{1-\pi}$,whereby π represent the unconditional treatment probability in the sample. The conditional control outcome with IPW the weighting scheme are hence given by

$$\bar{Y}_{IPW}^{C} = E_{X|D=1}[Y_{0i} \mid p(X_i)] = \sum_{i \in S_X, D_i=0}^{N_0} \omega_i^{IPW} Y_{0i}. \tag{4.4}$$

In practice, the inverse probability weights are commonly normalized and scaled in empirical applications. Note that the prerequisites for IPW are slightly different than for PSM. As ω_i^{IPW} depends directly on the *value* of the estimated propensity score, a consistent estimate of the treatment probability is required for the estimation of Y_{IPW}^{C} (Wooldridge, 2007; Waernbaum, 2012).

4.2.3 Interpreting Conditional Differences

After reweighing the control outcomes with the balancing weights, the conditional outcome differences between treated and controls can be estimated is given by the sample analogue of

$$\Delta_{p(X)} = \bar{Y}_1 - \bar{Y}^C. \tag{4.5}$$

The question of what has been identified with parameter $\Delta_{p(X_i)}$ depends on the additional assumptions made with respect to the unobservable factors in equation (4.1). As balancing only accounts for outcome imbalance arising due to the observable characteristics X, differences arising due to differences in U_i or $m_i(\cdot)$ are not accounted for. When aiming to interpret $\Delta_{p(X_i)}$ causally, the additional

assumption of "unconfoundedness" or "conditional independence" (CIA) has to be made. The CIA states that conditional on X, the outcomes are mean independent of the treatment status. Under additive separability of X_i and U_i this formally amounts to ruling out differences due to unobserved characteristics in the outcome equation U_0,

$$Y_{0i} \perp\!\!\!\perp D_i \mid p(X_i) \;\Rightarrow\; U_{0i} \perp\!\!\!\perp D_i \mid p(X_i). \qquad (CIA)$$

A further assumption to be made is the "stable unit treatment value assumption" (SUTVA) or "simple counterfactual treatment" assumption. This alludes to general equilibrium or spill-over effects that emerge when, e.g., the existence of the treatment group, or the choices made by members of one group, alters the returns to characteristics of the other group (see, e.g., Miguel and Kremer, 2004). Arguments defending the absence of such effects are usually based on detailed knowledge of the institutional setup and the underlying decision making process. In small scale treatments, the SUTVA assumption pertains to differences to potential effects of treatment on $m(\cdot)$, as the availability of the intervention for the treatment group may alter the return function of the control group. In large scale treatments, or in case where D describes a population parameter, e.g., gender, the effects of D on $m(\cdot)$ and hence also the validity of SUTVA are difficult to assess, so that the remaining differences are to be interpreted as the result of both differences in the return function and differences due to treatment D, or as the "unexplained part in conditional differences", as in the decomposition literature. Only when both the CIA and SUTVA assumption hold, $\Delta_{p(X_i)}$ can be interpreted as the average treatment effect on the treated (ATT), i.e., the "causal effect" of D on Y.

4.2.4 Empirical Implementation Steps

The empirical implementation of PSM and IPW can be divided into six chronological steps, which we will take as the structure for this chapter. The individual steps and their correspondence in the section of this chapter are summarized in Table 4.1. Depending on what parameter is to be identified, the first issue one needs to address is the choice of an appropriate data that contains the characteristics to be balanced across the population groups. In general, all balancing endeavours should start with a thorough assessment of the relevant variables to be included in the propensity score. As outlined before, when balancing is done

for the purpose of estimating causal treatment effects, the required rigorousness of variable selection is particularly high, as the conditional independence assumption needs to be satisfied. The first step of implementation is to select identify and select the relevant variables that need to be balanced and to assess distributions of characteristics in the treatment and control group to uncover characteristics with extreme imbalance and potentially problematic overlap. We address these issues in Section 4.3.2. The second step of propensity score estimation commonly relies on parametric logistic or probit models, but recent attempts have been made to increase the robustness of the estimation by introducing more flexible parametric models, data-mining techniques, or automated balancing propensity scores. We outline the basic ideas of these approaches in Section 4.4.

The third step involves the choice of the matching or weighting estimator. As outlined in Caliendo and Kopeinig (2008) PSM estimators are subject to a bias-variance trade-off that is affected by the choice of tuning parameters. We outline the effects of the individuals tuning parameters, and provide some suggestions to reduce the trade-off. We then outline the different IPW estimators and summarize the experience with their balancing power in practice. While large sample theory is often uninformative on which estimators to choose in which data setting, we outline the findings of a growing simulation literature pointing to systematic differences across estimators in *finite* samples that might help the decision finding process, in Section 4.5. As balance can only established in an area of common support, the fourth step involves the definition of an area of support, which is most often done based on the overlap in propensity score distributions. As some matching and weighting estimators have been found very sensitive to low overlap, the definition of an area of *thick* support may be necessary. We present the different methods suggested in the literature in Section 4.6. In the fifth and final step of balancing, the balancing success needs to be assessed via tests of balance of the reweighed characteristics. This is very important, as balance is not automatically warranted for all variables. The recent literature points to a poor performance of conventional parametric tests on detecting remaining imbalance, so that multiple tests of balance should be conducted to account for different distributional features of characteristics. Parametric and non-parametric tests for the equality of means and other distribution moments are outlined in Section 4.7.

Table 4.1: Implementation steps and estimation options for balancing and effect estimation with PSM and IPW

Step 1	**Variable Choice and Data Inspection**	Section 4.3
	Variable selection	4.3.1
	◇ Based on theory and empirical evidence	
	◇ Data-driven selection of variables	
	◇ How and when to measure	
	Data checks: sample size, treatment control ratio, imbalance checks	4.3.2
Step 2	**Estimation of the propensity score**	Section 4.4
	◇ Parametric Estimation	
	◇ Model Specification and Balance Checking	
	◇ Data Mining Techniques	
Step 3	**Selecting Matching or Weighting Methods**	Section 4.5
	NN Matching and the choice of tuning parameters	4.5.1
	◇ Selecting a caliper	
	◇ Matching without replacement	
	Kernel Matching/local linear/local polynomial matching	4.5.2
	◇ Kernel Choice	
	◇ Bandwidth Choice	
	◇ Local averaging vs. local regression	
	Subclassifiation/Stratification	4.5.3
	◇ Strata Choice	
	◇ Optimal Full Matching	
	Inverse Probability Weighting	4.5.4
	◇ Outliers, thin support and trimming	
	Finite Sample Comparison of estimators	4.5.5
	Exact Matching and Fine Balancing	4.5.6
Step 4	**Common Support Condition**	Section 4.6
	◇ Min-Max-Rule	
	◇ Trimming	
	◇ Optimal Support	
	◇ Convex Hull	
	◇ Prematching	
Step 5	**Checking Balance (revisiting Steps 1 through 4)**	Section 4.7
	◇ Testing for differences in means	
	◇ Testing for similar distributions	
	◇ Multidimensional balance measure	

Note: Own summary, the last columns refers to the sections in the chapter.

4.3 Preparing the empirical analysis

4.3.1 Variable Choice

In some applications the set of relevant characteristics that need to be balanced are not or only impartially known. Hence, the definition of relevant characteristics constitutes an important implementation step preceding the statistical analysis. When aiming to balance characteristics to construct conditional outcome differences, a meaningful interpretation requires a detailed understanding of both the outcome generating process and the covarying differences in group characteristics to credibly assert that all remaining differences are attributable only to the treatment or group status. When treatment participation is based on individual choice, factors influencing the (expected) benefits and costs of participation should be considered. Selection models offer a well-founded economic theory to support the variable selection (see, e.g., Roy, 1951; Heckman and Navarro-Lozano, 2004) and the instutional determinants of the selection processes are commonly better understood than the outcome process. When limited data availability does not allow to observe all relevant characteristics, meaningful approximations need to be considered. For example, in context of labor market program evaluation, the pre-treatment history of labor market status and wages is found a very good predictor of expected program success by approximating unobserved labor market attachment and aspirations that also influence the decision to participate in these programs (Heckman et al., 1998; Sianesi, 2002, 2004; Lechner and Wunsch, 2013).

Statistical Association In some applications theoretical and empirical knowledge of the subject matter is not sufficiently evolved to specify the set of relevant variables in detail. In this case statistical goodness-of-fit indicators of the treatment model are commonly used to decide about the relevant variables. Indicators of predictive power, goodness-of-fit analyses (likelihood ratio-tests, χ^2-test, "hit-or-miss" ratios), or (mis-)specification tests (Shaikh et al., 2009) may help to discriminate between different models, with respect to the questions whether and how (i.e., including higher order and interaction terms) to include a given set of characteristics. It should be kept in mind that variable selection based on the treatment model alone tends to overemphasize the importance of variables that are strong predictors of the selection equation, and one may risk accidentally including characteristics that are affected by the treatment (Wooldridge, 2005).

A number of studies point to the perils of including characteristics that are not or only weakly related to the outcome (e.g., instrumental variables). Clarke et al. (2011) point to a particular correlation structure between observable determinants of the treatment equation and unobservables in the outcome equation that leads to an increase in the estimation bias when the former is included in the propensity score. On a similar note, the inclusion of instrumental variables can lead to significant deterioration of the consistency of PSM, when the CIA does not hold; the stronger the instrument, the higher the risk of inconsistency (Heckman and Navarro-Lozano, 2004; Bhattacharya and Vogt, 2012; Myers et al., 2011).

An alternative data-driven approach is to model the outcome equation. The relevance of specific characteristics may be assessed by the t-statistic of the coefficients (e.g., Patrick et al., 2011). Although less often used in practice when implementing IPW and PSM this may help prioritize variable selection and detect non-linear relationships between characteristics and the outcome of interest. Using outcome analysis to assess the functional form goes against the idea of a clear separation between "design" and " analysis" as promoted by Rubin (2007). As one might intuitively select the model with the strongest treatment effect estimate, this risks to reduce the objectivity of the balancing exercise. To avoid that the choice of characteristics is in any way linked to the success of the treatment, the regression could be done using control outcomes only, or pre-treatment values of the outcome that are not influenced by treatment.

Over-fitting A practical problem related to the inclusion of many characteristics strongly related to treatment is the exacerbation of the overlap problem. When treated and controls can be discriminated "too well", the common support condition might be violated so that for some treated the counterfactual cannot be estimated. At the same time, it has been found that the reliablity of PSM and IPW may be substantially reduced in areas of low support, due to the large weight given to only few control variables (Augurzky and Schmidt, 2001; Brookhart et al., 2006). Hence, while a number of studies defend over-fitting for the purpose of proxying potentially omitted relevant characteristics[3] (Rubin and Thomas, 1996; Zhao, 2008; Millimet and Tchernis, 2009, 2012) on the basis that the benefits of

[3]Note, the distinction between "over-fitting" and mis-specification, which are often used interchangeably in the literature. While the former refers to the inclusion of variables unrelated to the outcome, the latter refers to the inclusion of unnecessary higher order or interaction terms of relevant variables to increase the flexibility of the model. A similar distinction is used in Millimet and Tchernis (2009).

bias reduction outweighs the costs of the loss in efficiency, it is advised to apply "the more the better"- selection rule with caution. When the number of relevant characteristics that *must* be included is already quite high, the efficiency cost of including characteristics of only little relevance can be quite strong.

How and when to measure More general advice on the inclusion of variables concerns the timing of measurement of characteristics, as none of the characteristics in the selection equation should be influenced by the treatment participation (Holland, 1986). This not only precludes information measured *after* the participation of treatment, but also any measurement prior to participation that might be influenced by behavioral responses in expectation of the treatment. Examples include shocks to previous outcome measures, as Ashenfelter's earnings dip (Ashenfelter, 1978), or changes in behavior due to threat effects of treatments (see, e.g., Bergemann et al., 2011). In a more general analysis Chabé-Ferret (2014) shows that conditioning on pre-treatment outcomes only yields a consistent estimator, when the previous outcomes were not subject to unobserved and persistent shocks. Depending on the assumptions about the persistency of previous shocks to outcome, and their effect on the participation in treatment, it may be better to apply a difference-in-difference strategies as this may deal with shocks that are symmetric around the entry date, or not to condition on pre-treatment outcome values altogether.

Furthermore, the data selection process should be harmonized across treatment groups, as differently administered data sets might lead to heterogenous definitions of characteristics, differences in response rates and missing variables for the two groups to be compared (Heckman et al., 1998; Rosenbaum and Rubin, 1984). Rosenbaum and Rubin (1984) further notes to inspect the distribution of missing values across treatment status as this might also be indicative of pre-treatment differences. Rather than excluding individuals with missing values it may also be more beneficial to assign missing values a category of their own, or to use imputation methods to avoid a selective estimation sample (Mattei, 2009).

4.3.2 Data Inspection

Based on the selection of variables and the corresponding data source, the characteristics should be assessed descriptively. A first point of interest is the relative

size of the treatment and the control group in data or the sample of interest. The ratio of "treatment-to-control-observations", $r_{TC} = N_t/N_c$ is a good indicator for the expected power of the balancing exercise. The smaller the ratio, the more reliable the estimate of the control group. Treatment-control ratios of around 1/2 to 1/4 are to be considered sufficient, but a higher absolute number of controls may be required when the selection is strong as then required overlap is expected to reduce the relevant control sample. The treatment-control-ratio is decisive in the choice of the appropriate balancing method, as PSM and IPW estimators differ systematically with respect to their efficiency (see Section 4.5).

The second point to be assessed is the degree of divergence in characteristic distributions, i.e., the strength of selection into treatment. The stronger the difference in covariate distributions, the lower the overlap, and the more difficult the balancing challenge. An initial descriptive comparison of the characteristic distributions give of sample means (e.g., the *standardized bias*, see Section 4.7), variability and range of each characteristics, helps to identify "problematic" characteristics with extreme imbalance. Large differences in pre-balancing means and a lower range of the characteristic values in the control group compared to the treatment group suggest a rather "unfavorable" matching setting, that should be accounted for (Rubin, 1973). Options are to restrict the estimation to a particular subset of individuals to create a more harmonious initial sample of treated and controls, find an alternative data source, where the differences are less pronounced, or apply an exact matching schemes for these variables to ensure balance of problematic characteristics. In the causal estimation of treatment effects, balancing priority is given to variables that are both very influential in the outcome and the selection equation so one should put specific attention on the pre-balancing distribution of these characteristics (Rubin, 2004).

4.4 Propensity Score Estimation

After preparing the empirical analysis, the propensity score needs to be estimated. While variable selection is focussed on collecting all relevant confounders and thinking about their appropriate functional form, the specification of the score also serves the objective of balancing. As pointed out in Section 4.2.2 an important difference in the mechanics of PSM and IPW is that the consistency of the treatment model is not a necessary condition to achieve conditionally independent

outcomes for PSM. As long as the CIA condition holds, unbiased effect estimates may be obtained also from a misspecified selection equation. In IPW, where individuals receive a weight depending on the *value* of the propensity scores, the estimation of a consistent selection probability is of higher importance. Measures of overall predictive power of the model might be helpful in deciding whether or not the selection equation can be improved. However, as in PSM, the ultimate objective is to achieve a balanced sample, so that the value of any specification should be judged by their balancing performance after weighting and matching (Section 4.7). Most commonly, the score is estimated parametrically — in the following we discuss recent attempts to improve the performance of standard logit and probit models, and further outline the alternative of using data-mining techniques. We refrain from outlining non-parametric methods, as similar flexibility and greater precision can be achieved by using parametric approaches (see, for semi- or nonparametric approaches Hirano et al., 2003; Lehrer and Kordas, 2004; Frölich, 2006, 2007b).

Parametric Estimation of the Propensity Score In the majority of applications, logit or probit models are used to estimate the propensity score. The shape of these distributions is rather similar, with the logistic distribution exhibiting slightly larger tails than the normal distribution. Provided that the model is correctly specified, the use of the logistic or a probit link function does not matter much in practice (Zhao, 2008; Busso et al., 2014a). The compression of score values in the [0,1]-interval seems to be important, however, for ensuring the consistency of the estimator, so that linear probability models should be avoided by their feature of allowing extreme dispersions of the score values (Zhao, 2008; Kang and Schafer, 2007).

As the maximum likelihood estimates of logit and probit models are sensitive to outliers (see, e.g., Pregibon, 1982, on this issue), a recent strand of the literature suggests to employ link functions that are more robust to outliers. In particular, it is proposed to use estimators that flexibly adapt to specific distribution features, e.g., the kurtosis and the tails of the distribution, or to use link functions exhibiting even longer tails than the logit model to increase robustness against outliers. Liu (2005) suggests the so-called robit link function that is based on the Student t-distribution and allows flexible choice of a distribution parameter to accommodate specific features of the data at hand. For a particular choice of the parameter, the

t-distribution resembles the logistic distribution, but with slightly longer tails, and is hence more robust to outliers. Kang and Schafer (2007) apply the robit link in an application of IPW, and find that the consistency of the estimator can be slightly improved if conditional independence is given, and substantially improved if not all relevant confounders have been captured. Similarly, Koenker and Yoon (2009) propose two alternative flexible parametric families that allow the adaptation of the tail and kurtosis behavior of the distributions. A downside of these approaches is that they are not as easily implementable as the standard models, as the distribution parameters have to be estimated along with the other model parameters, and standardized computation tools do not exist. It is hence suggested to assess the relevance of outliers in a given application by testing the influence of individual variables in determining the shape of the propensity score, using e.g., standardized residuals or Cook's distance.

Model Specification and Balance Checking So-called "after-balancing" tests (Lee, 2013) are the only way to validate a particular specification of the propensity score. Hence, once a model is specified, the balancing method of choice is conducted and the level of balance assessed — if balance is not achieved, the model has to be re-specified, or a different balancing approach has to be applied[4]. It is often suggested to start the iterative specification process with a parsimonious specification of all relevant characteristics, check its balancing power and then add higher order or interaction terms step-wise when balance could not be achieved (e.g., Dehejia and Wahba, 1999; Lee, 2013).

To circumvent the arduous and adhoc-procedure of iterative specification checks, Imai and Ratkovic (2014) propose an estimation method that incorporates the balancing condition as an additional restriction in the estimation of the selection model. They show that this "covariate balance propensity score" (CBPS) may significantly improve the balancing performance of IPW and PSM, compared to the standard logit model, also exhibiting a higher robustness to misspecification of the treatment model. Note that this estimation technique still requires to think about the correct specification of the score in terms of interaction and higher order terms as input characteristics. So far, this approach has not been implemented in

[4]Ho et al. (2007) coin the "propensity score tautology" referring to the circular statements: "The estimated propensity score is a balancing score when we have a consistent estimate of the true propensity sore. We know we have a consistent estimate of the propensity score when matching on the propensity score balances the raw covariate"

many empirical studies, although it seems promising.

Data Mining Techniques Compared to the theory-based model specifications, statistical classification approaches are purely data-driven categorization algorithms that select the classification ("specification") of input characteristics with the highest predictive power of the propensity score model. The algorithms choose categorizations of variables, interaction and higher order terms automatically, and derive their power from choosing the terms that capture non-linearities in the selection process most appropriately in terms of model fit. In particular when the relationship between confounders and treatment selection is non-linear, neural networks, and methods involving classification and regression trees (CART), may significantly improve the performance of PSM and IPW compared to standard logistic or probit models (Setoguchi et al., 2008; Lee et al., 2010).

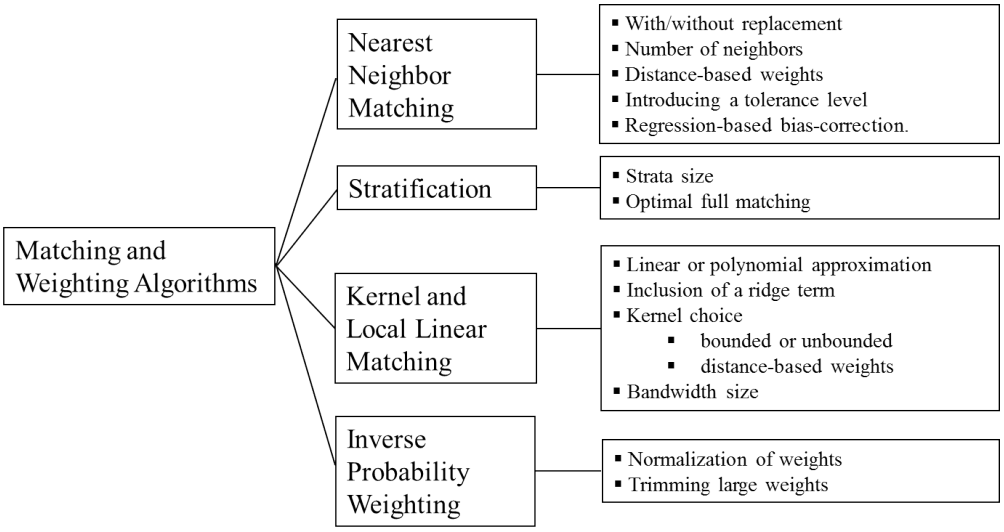
In particular "boosted CART" algorithms are often applied (Westreich et al., 2010). Here, regression trees with a given complexity are fit iteratively to the data, improving the model-fit in each iteration by reclassifying individuals that were misspecified by the model of the previous iteration. The power of this approach is derived from combining several individual regression trees via "boosting", thereby reducing the risk of classification error of one singular tree. The parameters to be chosen by the researcher are the complexity of the regression tree, i.e., the number of splits or end-nodes of the trees, the total number of iterations (trees), and the influence of each individual tree, i.e., the so-called learning, or shrinkage rate, all of which affect the fit of the model. A trade-off arises between achieving the best fit and efficiency of the model, so that the model should not be too complicated (compare McCaffrey et al., 2004; Schonlau, 2005, for some advise regarding the parameter choice).

Several boosting algorithms exist in the literature, the most versatile one being Friedman's gradient boosting (GBM) which has the nice feature of adapting its loss-function to the most common error distributions, *inter alia* the logistic one (see, e.g., Friedman, 2001; Ridgeway, 1999). McCaffrey et al. (2004) provide an interesting adaptation of boosting to the variable balancing context, where the GBM algorithm incorporates a "stopping-rule" to select the number of iterations at which maximal balance between characteristics across treatment groups is reached. Lee et al. (2010) show that this algorithm performs better than other CART approaches, resulting in IPW estimates with low bias and variance over a number of complex estimation scenarios.

4.5 Matching and Weighting Methods

Large sample analyses of matching and weighting algorithms show that most estimators lead to asymptotically unbiased results, although the efficiency of weighting estimators tends to be systematically higher than that of matching estimators (Hirano et al., 2003; Abadie and Imbens, 2006, 2011). While this suggests that estimator choice should be focussed on efficiency consideration, these analyses neglect that the small sample performance of estimator may be substantially different. As outlined in Caliendo and Kopeinig (2008), matching estimators are subject to a bias-variance trade-off that may become critical in finite, i.e., small samples. Further challenging data issues as low overlap, and non-linearities in the relation between treatment participation and the outcome of interest tend to distort the performance of the algorithms and should hence be accounted for (Frölich, 2004; Busso et al., 2014b,a; Huber et al., 2013). In particular, there seem to exist systematic differences of the robustness of estimators that were not picked up by the large sample theory. Figure (4.1) outlines the four large groups of methods used for balancing and provides an overview of the respective "tuning" parameters that have to be selected during integration.

Figure 4.1: Matching and weighting estimators and their "tuning parameters"



In the implementation of PSM, the bias and variance of the estimates is a direct function of the definition of the neighborhood, i.e., the set of closest control observations. The closer the matched neighbors to the treated in terms of their characteristics — and hence their propensity score value — the higher the matching quality. But, as outlined in Caliendo and Kopeinig (2008), choosing

parameters that maximize matching quality usually comes at the expense of an increased variability, as the number of controls that are matched is reduced. In IPW in contrast, balancing is achieved via the correct ordering and the value of the propensity scores, so that the implementation is predominantly concerned with the elimination of extreme values, and the maintenance of a smooth ordering of score values. The bias-variance trade-off is less of an issue. In the following we provide a brief description of the respective balancing methods and provided guidance on the choice of the tuning parameters. Subsequently, we compare the performance of these methods under challenging data settings. For this we draw heavily on the findings of Frölich (2004), Huber et al. (2013) and Busso et al. (2014a,b).

4.5.1 Nearest Neighbor Matching

In nearest neighbor matching, each treated individual is assigned a small neighborhood of controls that are closest in terms of their propensity score. Table 4.2 summarizes the formal definition of the neighborhood $\mathcal{A}(i)$ for treated i when using either pair matching, multiple neighbor matching or caliper/radius matching, and points to the bias-variance trade-off associated with the choice of neighbors K and caliper size ε. A similar trade-off exists with respect to the choice of matching with or without replacement, i.e., using individual control observations more than once, or removing them from the pool of controls after they have been used as a match. While the former reduces bias by ensuring that each treated is matched to the closest available control, this may reduce efficiency as the total sample of matched controls contains less *distinct* controls than the full sample.

The parameters to balance bias and variance have to be selected by the researcher. As there do not exist clear-cut formulas for this, one has to rely on guidance from the applied literature. A first piece of advice emerging from this literature is to use pair matching with replacement (with $K = 1$) as a reference for other matching estimator, as this estimator is purely focussed on bias minimization. Using the balance achieved here as a reference point, the sensitivity of balance to a systematic increase in the number of neighbors can be assessed. A second point of advice is to mitigate the trade-off between bias and variance by down-weighing distant observations in a neighborhood, i.e., to use distance-based weights, $1/(|p_i - p_j|)$ rather than uniform weights to aggregate the observations in the neihgborhood. This reduces the influence of distant matches while increasing the number of observations that can be used in a match. A third piece of advice

is to avoid the use of multiple-neighbor matching that choosing only a number of neighbors K. This method is generally outperformed by most other methods, as it offers no protection against outlier values, and the optimal choice of neighbors is very ad-hoc. Rather than selecting a maximal number of neighbors, a maximal distance, i.e., a caliper, should be selected. The use of a caliper has been found be particularly beneficial when X contains many continuous variables as this enforces the matching of controls with similar values in these variables (Austin, 2011).

Selecting a caliper The optimal caliper size is a function of the dispersion and relative location of propensity score values of treated and controls and should hence be selected as a function of these propensity score parameters. Rosenbaum and Rubin (1985a) and Austin (2011) propose to make the caliper a factor of the standard deviation of the propensity scores. Huber et al. (2013) suggest a factor of the relative propensity score distributions, e.g., to take 1/3 of the maximal observed distance in propensity score matches as the caliper. As smaller calipers reduce the number of control matches used, the size of the neighborhood should be monitored. Big gains in efficiency are already made with very small numbers of additional neighbors: with two (four) neighbors rather than one, efficiency may increase already by 50% (75%) towards optimal efficiency (Haviland et al., 2007). By imposing a caliper some treated may not find adequate controls and are hence eliminated from the estimation sample. While this is a natural imposition of the common support condition, one may wish to work around this redefinition of the population of interest by assigning these treated their closest neighbor outside the caliper (Lechner et al., 2011).

Matching without replacement When the pool of controls is large, efficiency considerations are usually secondary, so that nearest neighbor matching should be done with replacement. However, when the number of controls is too small (or the treatment ratio is close to one), a more efficient use of all controls may be important. Stepwise matching without replacement — so-called "greedy matching" — has the substantial disadvantage that the matching results depend on the ordering of matches, and that the quality of matching becomes worse for later matches. An alternative is given by "optimal matching", where matching is done without replacement, but the distance-minimization problem is done jointly for the whole sample to minimize the *overall* distance between treatment-control matches (Rosenbaum, 1989, 2002; Hansen, 2007). Haviland et al. (2007) show that this approach effectively reduces imbalance in the matched sample, while maintaining

the additional efficiency constraint. Similar to matching with replacement, optimal matching can be implemented with one or multiple neighbors, and incorporates caliper and radius matching. It further allows the use of variable neighbors for each match. Ming and Rosenbaum (2000) show that if the number of distinct controls is kept constant, a better balance can be achieved for the same level of efficiency. Note that this approach has not been assessed comparatively so that its ability in achieve a comparable level of balance as matching with replacement is not clear. It may hence be advisable to start the analysis with pair matching with replacement to assess the maximal balance that can be achieved, and compare it with the results of optimal matching, to see whether efficiency can be increased while maintaining a comparable level of balance.

4.5.2 Kernel, local linear and local polynomial matching

Kernel (KM), local linear (LLM) or local polynomial (LPM) matching exploit the sophisticated techniques developed in nonparametric regression analysis (see, Heckman et al., 1997 Heckman et al., 1998 Heckman et al., 1998 and Frölich, 2004). Here, the neighborhood for each treated i is defined by a symmetrically shaped kernel that assigns weights to control observations as a function of the scaled distance of propensity score values. The exact weighting scheme and the scaling are determined by the researcher, via the choice of the kernel function $G(\cdot)$ and the bandwidth parameter a_n, respectively. Table (4.2) depicts the formal representation of two most commonly used kernel functions, the Normal and the Epanenchnikov kernel, and outlines the bias-variance trade-off associated with the choice of a_n. Using these weights, the aggregation of controls in the neighborhood is done either by weighted averaging (*kernel* matching), weighted regression on an intercept and a linear term of the propensity score (*local linear* matching, and it's extension "ridge matching"), or a higher order terms of the propensity score (*local polynomial* matching) (see Smith and Todd, 2005a; Seifert and Gasser, 1996; Frölich, 2004, for details). The following section discusses advice from the literature on the choice these parameters.

Kernel choice The most commonly used kernels are the Gaussian (Normal) and the Epanechnikov kernel, both of which assign weights that are decreasing in the absolute distance of propensity score values. In the matching context, the distinctive characteristic of the two kernels is their (non-)boundedness. Unlike the

Table 4.2: Formal depiction of matching estimators and the bias-variance trade-off

Estimator	Neighborhood for treated i	Bias	Variance
Pair Matching	$\mathcal{A}_1(i) = \left\{ j : D_j = 0, \mathbb{1}\{\min_{j:D_j=0} \|\hat{p}(X_i) - \hat{p}(X_j)\|\} \right\}$		
Multiple Neighbor	$\mathcal{A}_K(i) = \left\{ j : D_j = 0, \sum_j \mathbb{1}\{\|\hat{p}(X_i) - \hat{p}(X_j)\| \leq \|\hat{p}(X_i) - \hat{p}(X_k)\|\} = K \right\}$	$K(+)$	$K(-)$
Caliper/Radius	$\mathcal{A}_\varepsilon(i) = \left\{ j : D_j = 0, \sum_j \mathbb{1}\{\|\hat{p}_j(X) - \hat{p}_i(X)\| < \varepsilon\} \right\}$	$\varepsilon(-)$	$\varepsilon(+)$
Kernel, LL, LP Matching			
Kernel argument	$s = (\|\hat{p}_0(X) - \hat{p}_1(X)\|)/a_n$	$a_n(+)$	$a_n(-)$
Gaussian kernel	$G_G(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2}$	(+)	(-)
Epanechnikov kernel	$G_E(s) = \frac{3}{4}(1 - s^2)\mathbb{1}\{\|s\| \leq 1\}$	(-)	(+)

Source: Own summary.

Gaussian kernel, the Epanechnikov kernel is bounded, assigning non-zero weights when the range of it's argument lies within $[-1, 1]$, see Table (4.2). In PSM, a_n is similar to a caliper, eliminating all controls j in a match i with the score distance $|p_i - p_j| > a_n$. The use of a bounded kernel may achieve better balancing quality and hence lower bias of estimates. The unbounded kernel has a higher efficiency by using a weighted average of *all* control observations. It is hence recommended to use a Gaussian kernel for efficiency reasons, but use an Epanechnikov kernel when there is a risk of including too distant matches, as, e.g., in regions of low overlap or when there are long tails in the distribution of control propensity scores.

Bandwidth choice In order to systematize the choice of a_n, many matching applications borrow from the literature on nonparametric kernel density estimation, using Silverman's (1986) "rule-of-thumb" approach , or "leave-one-out-cross-validation" (LOOCV). The latter selects the bandwidth that minimizes the mean integrated squared error (MISE) of a nonparametric regression estimator (see, e.g., Frölich, 2004; Busso et al., 2014b). These approaches are not optimal in the matching context as they not account for the relative location of treated and controls and are hence not selected to minimize the risk of bad matches where the treated is located and bad matches are likely to occur. Galdo et al. (2008) propose modifications of the LOOCV that account for the varying importance of control observations in the overall matching process[5], and uses locally varying bandwidth. These approaches provide modest but significant improvement of the match quality that is largest when selection into treatment is very strong, and when there are areas of low support. In general, a grid-search should be conducted to assess the sensitivity of the estimators to the bandwidth choice. When the propensity score

[5]A similar approach is proposed by Bergemann et al. (2009).

distributions are highly dispersed, the more sophisticate approach of Galdo et al. (2008) should be used. Alternatively, similar selection strategies as in radius or caliper matching could be used.

Local averaging or local regression The non-parametric regression literature suggests that local regression should be preferred over local averaging as it is more robust to outliers and performs superior in case of endpoints, by automatically adjusting to boundary observations, when a symmetrical kernel cannot be constructed (see Fan, 1992, 1993; Fan and Gijbels, 1995; Fan et al., 1997). Hence, when boundary observations arise due to low or lacking overlap, local linear matching (LLM) should be preferred over local constant kernel matching (KM) (Frölich, 2004). A more flexible approximation via the inclusion of local polynomials does not seem to improve results significantly. Also, a local logit approximation as in (Frölich, 2006) does not yield considerable gains over linear regression (Huber et al., 2013)

The ridge matching estimator (LRM) proposed by Frölich (2004) is based on the ridged local linear regression estimator proposed by Seifert and Gasser (1996); for a more in-depth discussion see Seifert and Gasser (2000). In LRM matching an increased stability of the estimates is achieved by two modifications of the LLM weights. First, the linear slope term is built around a smoothed average of control observations around $\bar{p}_i$ rather than p_i itself. This results in a higher numerical stability of the linear term and prevents the occurrence of negative weights. Second, in the denominator of the linear term a "ridge" parameter r is included aiming to improve the stability of the weights in areas where only few and distant controls are available. Comparative analyses show that LRM tends to outperform conventional LLM in terms of robustness, and bears the additional advantage of being less sensitive to the choice of the bandwidth parameter (Ham et al., 2011; Frölich, 2004). As to the choice of r, Seifert and Gasser (2000) provide a rule-of-thumb for that defines the optimal r with respect to kernel chosen.[6].

4.5.3 Subclassification/Stratification

Based on the ordering of individuals by their propensity score values, Rosenbaum and Rubin (1983b) show that by simply classifying treated and controls into sub-

[6]They suggest that $r \approx .35$ for the Gaussian kernel, and $r = 0.325$ for the Epanechnikov kernel.

groups of similar scores, the imbalance in characteristics of treated and controls is substantially reduced. The neighborhood is hence defined by the number and size of the subclasses or "strata" chosen. As the assignment to strata is mutually exclusive, subclassification uses all control observations without replacement.

Strata choice Clearly, the finer the strata, the higher the balance of characteristics within strata. However, if the strata are chosen too thin, some treated and controls will have to be eliminated. Rosenbaum and Rubin (1984) find that five subclasses of the propensity score are sufficient to reduce the imbalance within classes about 90%, so that the stratification by quintiles of the ordered scores is often used as a starting point. Tests of imbalance in propensity score values and characteristic distributions within the strata can help assess whether the size of stratum should be further reduced (Dehejia and Wahba, 1999, 2002). Based on a rather broad stratification, the size of a stratum is reduced, e.g., by dividing it further at the median propensity score value of the stratum, if some variables are unbalanced (see Section 4.7 for details on balancing tests to be used).

Optimal Full Matching Similar to optimal matching, optimal *full* matching stratifies the sample based on the minimization of the overall distance in propensity score values, and including *all* individuals in the matching process to improve efficiency (Rosenbaum, 1991; Hansen, 2004). Optimal full matching creates subclasses of the type one-one, one-many, many-one, whereby the size of the strata is defined by the number of treated and controls in each strata, which is allowed to vary across strata. It hence shares similar beneficial features with optimal matching, by allowing the treatment control ratio to vary across the distribution of propensity scores. The vector of optimal strata sizes of a given matching problem is defined by the minimal distance in propensity scores over all strata. Hansen (2004) proposes an extension that avoids that some strata absorb a very large number of treated or controls, while others are only pair-matched strata, as this may increase bias as well as variability of estimates. Stuart and Green (2008) compare the performance of both version of optimal full matching relative to conventional subclassification and find that constrained optimal full matching yields the highest degree of balance of characteristics across treatment groups.

4.5.4 Inverse Probability Weighting

In inverse probability weighting the predicted propensity score values are used to manually calculate the balancing weights, which are then used to reweigh control observations. The literature proposes different weighting functions that differ with respect to their normalization, and their consideration of unconditional sample probabilities. Let $\hat{p}_j(X)$ denote the propensity score estimate of individual j in the control population, and $\hat{p} = N_1/N_0 + N_1$ the observed frequency of treated in the sample. The different weights for the controls are given by[7]

$$w^{UN} = \frac{\hat{p}_j(X)}{(1-\hat{p}_j(X))} / \frac{\hat{p}}{1-\hat{p}} \tag{IPW1}$$

$$w^{N} = \frac{\hat{p_j}(X)}{(1-\hat{p_j}(X))} / \frac{1}{n_0} \sum_{i \in D_0} \frac{\hat{p_i}(X)}{(1-\hat{p_i}(X))} \tag{IPW2}$$

$$w^{LD} = \frac{\hat{p_j}(X)}{(1-\hat{p_j}(X))}(1-C_j) / \frac{1}{n_0} \sum_{i \in D_0} \left(\frac{\hat{p_i}(X)}{(1-\hat{p_i}(X))}(1-C_k) \right) \tag{IPW3}$$

with the correction term

$$C_j = \frac{(1 - \frac{\hat{p}_j(X)}{\hat{p}} A_j) \frac{1}{n} \sum_{i \in D_0} (1 - \frac{\hat{p}_j(X)}{\hat{p}} A_j)}{\frac{1}{n} \sum_{i \in D_0} (1 - \frac{\hat{p}_j(X)}{\hat{p}} A_j)} \quad \text{and} \quad A_i = \frac{1 - D_{0i}}{1 - \hat{p}_i(X)}.$$

The sum of weights in IPW1 only add up to one in expectation, which might result in problematic behavior in the practical application. In particular, IPW1 weights might take on extreme values above one, which may then exert substantial influence in the overall reweighting scheme, resulting in unreliable balancing outcomes. At the same time, this may have the unattractive consequence of resulting in unrealistic values of the treatment counterfactual (Busso et al., 2014a,b). The normalized weights in IPW2 as proposed in Imbens (2004) overcomes these problems, and show a higher robustness to outliers. As this comes at no cost, IPW2 should always be the preferred in empirical implementations. The additional modification of weights in IPW3 aims to further reduce the large sample variance of

[7] We follow the notation by Busso et al. (2014a).

the weighting estimator (Lunceford and Davidian, 2004; Busso et al., 2014a). In IPW3, a correction term C_j scales the weights of controls towards the unconditional sample probability. By this, observations with scores lower (larger) than the average sample probabilities are scaled up (down), so that the range of the propensity score distribution is compressed, thus reducing the variance of the estimates and decreasing (increasing) the influence of very large (small) weights. Simulation studies suggest, that the benefits of the correction term arises predominantly in large samples, in small samples, IPW2 and IPW3 estimators are expected to lead to similar results (Busso et al., 2014a).

Outliers, thin support and trimming As mentioned before, the success of reweighing crucially depends on the *value* of propensity scores and hence a correct specification of the treatment model (Huber, 2011). Several approaches might help in assessing the sensitivity of estimates. First, the stability of the ordering of individuals could be assessed by plotting the score values of individuals in different specification of the score. If singular outliers amongst controls are detected, the analysis could be redone without these outliers. Note, that when selection is strong, the propensity score distribution of controls is often characterized by a long thin tail at the upper end of the score distribution, who are then given a disproportionately large role in the weighting. Khan and Tamer (2010) and Busso et al. (2014a) show, that this type of "thin-support" may result in very high variance of treatment effect estimates and slower convergence rates, (even convergence at infinity), thus leading to biased results in finite samples. One way to deal with this is the use of robust link functions (see Section 4.4) or boosted CART rather than parametric regression as this might reduces the spread of estimated scores (Lee et al., 2010). Alternatively, trimming of large propensity score values is advised (see also Section 4.6). Huber et al. (2013) suggest to exclude 4% to 6% of individuals with the highest weights, Lee et al. (2011) on the other hand cap the weight of controls above some cut-off percentile of the weight distribution by setting them exactly equal to this cut-off percentile. However, by the ad-hoc nature of these cut-offs, it is in general advised to assess the sensitivity of the results to varying levels of trimming.

4.5.5 Finite sample performance of balancing methods

While all outlined methods are expected to perform similar in large samples with ample overlap, their small sample performance might differ substantially as pointed out by the simulations studies of Frölich (2004), Huber et al. (2013) and Busso et al. (2014a,b). A general point to emerge from these studies is that most methods are highly sensitive to low or lacking overlap in propensity score distributions. Here, very few number of control observations may either produce locally unstable matches, or decrease the consistency of IPW due to extreme weights. When conditional outcomes are to be estimated, non-linear relationships between the treatment probability and the outcome of interest located in the area of low support pose a substantial threat to unbiased estimation of treatment effects. Hence, in all applications the issue of low support should be addressed with care. In the following, we outline in more detail the performance difference amongst the outlined methods as found in of the above-mentioned studies. Over all studies, caliper matching with distance-based weights, ridged LLM matching, and IPW2 and IPW3 appear to be most robust estimators.

Small sample size In small samples (100 to 500 observations) or in settings with a high treatment-control ratio, only very few controls are available. This increases the variability of the control group and may decrease balancing quality, due to the high relative importance of each control observations. To decrease variability, matching estimators using a larger number of neighbors should be used in combination with down-weighting and calipers to reduce the influence of distant observations. Hence, caliper matching with distance-based weights, ridged LLM matching, or optimal matching are found to perform fairly well in small settings. Kernel and local linear matching tend to perform not so well, which may be explained by their higher sensitivity to the bandwidth size and boundary observations. Provided that the score can be correctly estimated, IPW2 and IPW3 tend to outperform matching estimators due to their efficient use of all control observations - the two IPW methods tends to perform similar in small to medium sized samples. As the sample size increases local linear and ridge matching quickly become competitive to weighting estimators, both in terms of bias and variance. Kernel matching remains somewhat erratic, however and should hence be accompanied by extensive sensitivity checks with respect to the choice of the bandwidth.

Non-linearities When conducting balancing for the purpose of conditional outcome estimation, non-linearities in the relationship between the treatment probability and the outcome of interest are detrimental to identification when the nonlinearity is located in areas of low overlap and small sample size. Due to the strong impact of bad matches, the bias of estimates may be increased substantially. While none of the estimators – except for IPW1 — is particularly sensitive to non-linearities (even extreme ones) if they are located in the thick support of the propensity score, *all* estimators are found to produce highly biased effect estimates even in the presence of only small deviations from a constant effect. The impact of the nonlinearity can be reduced by using matching estimators with a small caliper bandwidth size. Pair matching, ridge matching and IPW2 exhibit the highest robustness - although the relatively good performance of ridge matching tends to manifests itself only in larger samples, and IPW2 is particularly sensitive to problematic overlap and non-linearities in the upper tail of the distribution, also compare (Basu et al., 2008). Slight non-linearities can be overcome by combining PSM and IPW with parametric regression analysis (Section 4.15). Extreme non-linearities in the region of thin support are difficult to deal with, however, and might require the redefinition of the estimation sample altogether.

4.5.6 Exact Matching and Fine Balancing

The methods discussed so far balance all characteristics with the same priority. In some applications it might be beneficial to prioritize balancing of some important characteristics. This can be done by *exact* matching on singular characteristics. One way to do this is to stratify the sample into cells defined by these characteristics values and conduct balancing within these subgroups (Heckman et al., 1998). An alternative approach is to combine the previous estimators with multivariate exact matching or Mahalanobis metric matching (Rubin, 1980) that matches individuals through minimizing the Mahalanobis distance (MD) (see Imbens (2004) and Zhao (2004) for a discussion on how to adjust the MD weights in the matching context). Again, this can be implemented by either first balancing all characteristics via PSM or IPW and then matching exactly on a selected subset of important confounders. For example, Rosenbaum and Rubin (1985a) suggest to conduct caliper matching, and then conduct Mahalanobis matching within these calipers. Alternatively, the propensity score and characteristics can be matched in Mahalanobis matching — note however, this approach my result in a higher imbalance

of the characteristics include in the propensity score.

With more than one exact matching variable, the strata size defined by the combination of exact matching variables may become rather small, not allowing for reliable balancing within these cells. To reduce this problem, Mahalanobis matching can also be implemented using an optimal matching algorithm, i.e., minimizing the distance across the propensity score and additional covariates over all matches, instead of optimizing the distance at each match (Hansen, 2007). In a similar spirit, Rosenbaum et al. (2007) propose the method of "fine balancing", which aims to achieve exact balance in the *distribution* of characteristics in the matched sample rather than forcing exact matches in each matching step. They show that fine balancing can achieve exact balance, even when exact matching on a high-dimensional characteristic is not possible due to insufficient sample size.

4.6 Common Support

To avoid balancing outside the region of overlap, the area of common support between treated and controls should be routinely assessed, which can be conveniently done using the estimated propensity score values. Density plots of the estimated score values for each treatment group provide a good initial assessment of problematic overlap. A visual comparison of the support and relative location of the propensity scores of treated and controls, helps identify extreme values, long tails and regions of low densities arising from characteristic combinations with only little support in a given treatment group. When overlap is missing, the value space of balancing characteristics should be restricted to areas of joint overlap by removing treated or control observations that do not have a corresponding value in other group.

For the estimation of treatment counterfactuals, the imposition of the common support condition is done in terms of the treatment observations to avoid extrapolation for propensity score values that do not exist in the control sample. Hence, all treated observations with $\hat{p}_{1i}(X) > \max(\hat{p}_{0i}(X))$ and $\hat{p}_{1i}(X)) < \min(\hat{p}_{0i}(X))$ should be eliminated. Clearly, the elimination of treated individuals implies a redefinition of the estimated counterfactual, and hence the estimated treatment effect as estimation population now differs from the sample population (Rosenbaum and Rubin, 1985b). With treatment effect heterogeneity, the effect estimated over the common support effect may differ from the effect estimated on

the whole sample. For a meaningful interpretation of the estimator, differences in characteristic distributions between individuals within and outside of support need be documented (Rosenbaum, 2002). Lechner (2008) further proposes to calculate nonparametric bounds to assess the sensitivity of the final estimator to the deletion of treated observations for the imposition of the common support.

Besides ensuring the minimal overlap for reasons of identification a causal treatment effect as stated in the *Overlap* condition (Section 4.2), it may be optimal to restrict the estimation to an area of *thick* or *strict* overlap to reduce the risk of estimation bias. As outlined in the previous section, thick support is expected to increase the stability of both PSM and IPW estimators near thin-support boundary points. Furthermore, the definition of an area of thick support may reduce the problem of model-dependency. Parametric models used to estimate the score tend to fit well to observations in the "center" of covariate distribution and may hence produce rather different and unreliable predictions in the tails of the distribution (King and Zeng, 2006; Ho et al., 2007). Slight mis-specification of the propensity score model may therefore risk the balancing success predominantly through different tail values, so that the sensitivity of balancing to eliminating controls in the tails of the propensity scores should be assessed. The definition of the area of thick support is done in terms of the control observations only, or both treated and controls as the elimination of controls may also modifies the region of common support. In the following we outline different approaches to detect and establish areas of thick overlap.

Minima-Maxima Rule In most applications the problem of low or lacking overlap arises with extreme values in the tails of propensity score distributions. As the treatment model discriminates well between treated and controls, the distribution of predicted propensity score is skewed to the right for the treated, and skewed to the left for controls. As a consequence it is usually that $\max(\hat{p}_{1i}(X)) > \max(\hat{p}_{0i}(X))$ and $\min(\hat{p}_{1i}(X))) > \min(\hat{p}_{0i}(X))$. To avoid the influence of controls with extreme values of propensity scores, Dehejia and Wahba (1999) propose to eliminate controls that exhibit values outside of the range of propensity score values. The area of overlap is hence defined by extreme values of the treatment distribution.

$$\hat{S}_{MM} = \{\hat{p}_{0i}(X) : \ \hat{p}_{0i}(X) \in [\min(\hat{p}_{1i}(X)), \max(\hat{p}_{1i}(X))]\}.$$

The convex hull of variable values Instead of ensuring overlap based on values of the propensity score, this can also be done based on covariate values. This is more transparent in terms of sample definition and has the additional advantage that the observations are not discarded wrongfully in case of a misspecified selection model. King and Zeng (2006) and Ho et al. (2007) propose a multidimensional area of common support, based on the convex hull of variable values. The convex hull is defined as the subset of all observations whose characteristics lie within a polygon formed by connecting the minimum and maximum values of all variable combinations (see King and Zeng, 2006, for a more detailed explanation and a visualization). The common support is established by eliminating controls with variables values outside of the convex hull defined by variable values. The convex hull is quite demanding to calculate when there are more than three characteristics to be balanced. Statistical software is available (see Table A4.1). Note, that all variables should be included in the set of variables, i.e. including interaction and non-linear terms.

Trimming Instead of focussing on extreme characteristic values, areas of low overlap can be defined as a function of the propensity score densities. Heckman et al. (1997, 1998) and Smith and Todd (2005a) propose to eliminate controls with propensity score values whose density is lower than a threshold level q. Based on a kernel density estimate of treated and control propensity scores, $\hat{f}(\hat{p}_{di}(X)), d = 0, 1$, propensity score values are with zero density are eliminated, a joint ranking of the density estimates is conducted. Based on this a distribution function of density values is obtained, with q denoting the q-th quantile, and c_q the corresponding density values. The region of common support $\hat{S}_{P_q}$, is then defined by

$$\hat{S}_{P_q} = \{\hat{p}_{di}(X) : \hat{f}(\hat{p}_{di}(X)) > c_q\}.$$

Smith and Todd (2005a) and Frölich (2004) propose values of q between 0.02 and 0.1, but depending on the context, higher values might be appropriate. An important parameter in this approach are the estimated kernel densities and the related bandwidth choice, as this determines the sensitivity of the kernel estimator to non-smooth areas in the distribution. Smith and Todd (2005a) propose to use Silverman's (1986) rule-of-thumb to define the bandwidth. As this approach might result in the elimination of control observations also in the middle range of the

distribution, the sensitivity of matching estimators around the boundary values should be closely monitored.

Huber et al. (2013) propose to eliminate individuals based on the balancing weights ω_{ji}^{PSM} and $\omega^{IPW}i$ rather than the propensity score. By deleting control observations who receive extreme weights it is avoided that they influence the results excessively (note that in many application this may coincide with propensity score values with a low density). Based on a ranking of the assigned weights, the highest $t\%$ are deleted. Hence the common support is defined by

$$\hat{S}_\omega = \{\hat{p}_{0i}(X) : \omega(\hat{p}_{0i}(X))\mathbb{I}[\omega(\hat{p}_{0i}(X))/\sum_{j}^{N}\omega(\hat{p}_{0i}(X)) \leq t\%].$$

Huber et al. (2013) use 4% to 5% as cut-off. When implementing IPW they further propose to limit the influence of large weights instead of discarding the information altogether. This "truncation" procedure is implemented by setting propensity scores above a certain threshold value $\bar{\omega}$ equal to $\bar{\omega}$, with the threshold being fixed at some percentile of the propensity score distribution. While IPW requires that the true propensity score values are used, only some few extreme observations should be eliminated this way to avoid that the identification of the treatment counterfactual is lost.

Optimal support, or "10-90 rule of thumb" The elimination of controls in the definition of a common support has two opposing effects on the efficiency of estimates: While the reduction in sample size increases the variance of estimates, the elimination of problematic scores might improve the precision of estimates due the reduction of distant matches and extreme weights. Crump et al. (2009) hence suggest to use this trade-off to systematically choose a subset of the estimation sample for whom the treatment effect can be estimated most efficiently. In particular, they propose to conduct the estimation on a subset of values $\mathbb{A}^* = \{x \in X|\hat{p}_i(X) < 1-\alpha^*\}$, with α^* being chosen "optimally" in the sense that it minimizes the asymptotic variance of the estimator. The cut-off level α can be calculated numerically (see Table A4.1 for software). Crump et al. (2009) also show that the optimally chosen α is oftentimes similarly effective as a simple exclusion of individuals above the 90-th percentile of the propensity score distribution.

4.7 Assessing the Balancing Quality

Having calculated a set of weights ω^{IPW} or ω^{PSM}, statistical tests on the equality of distributions between treated and reweighed controls need to be applied to assess their balancing power. An inadequate choice of the neighborhood in PSM or specifying the wrong selection model in IPW might result in the calculation of weights that do not or only insufficiently balance the characteristic distributions, or might even increase imbalance for some characteristics. Tests on the equality distribution moments hence provide the only diagnostic check on the appropriateness of a set of estimated weights. The following Section outlines tests that are frequently used to measure and test for balance across treatment groups. Beforehand we address some general points to be considered when implementing them.

Whether and which balancing test should be used as reliable balancing indicators is a subject of an ongoing debate. A first issue to consider is that most conventional hypothesis tests are highly sensitive to sample size, the treatment-control ratio and other data features unrelated to the balance in characteristics (Imai et al., 2008; Ho et al., 2007), so that changes in these parameters may distort the test results. To account for the differential sensitivity with respect to distributional features of characteristics, it is advised to run multiple balance diagnostics simultaneously (Smith and Todd, 2005b; Sekhon, 2007; Lee, 2013), and use on tests on the equality of means as well as higher order moments. Sekhon (2007) proposes to use the largest minimal p-value over all tests to assess the overall degree of balance for a given specification. A second issue to keep in mind is that conventional significance values may not be valid in the balancing context. In the balanced sample, the underlying assumption that the sample is drawn from a normally distributed super-population is not expected to hold, so that conventional significance levels lose their meaning. Resampling methods, i.e., permutation and bootstrapping,[8] may be used to obtain more meaningful confidence levels. While they may improve performance of conventional significance values, permutation-based tests may also be overly sensitive to slight imbalance or not sensitive enough to reject imbalance (see Lee, 2013; Huber, 2011, for an application to tests for equality in means or distribution quantiles, respecively.). We hence advise to maximize (minimize) test-statistics and p-values without limit, irrespective of conventional significance cut-offs (see Imai et al., 2008; Ho et al.,

[8]See, e.g, Good (2005) for an introduction.

2007), to avoid the false acceptance of balance.

The variables to be tested include all characteristics in the propensity score, including interactions and higher order terms. In the context of estimating conditional outcomes, different characteristics may have different degrees of "balancing priority", by their relevance in the outcome equation, as outlined in Section 4.3.1. Whereas a small imbalance of strong predictors of the outcome might result in hugely biased results, rather large difference of not very relevant confounders will not make a big difference (Rubin, 2004). Also, the overall importance of balance checks depends on the further strategy for the estimation of the treatment effect (see Section 4.8.3). When balance is used as a pre-processing strategy for further parametric outcome analysis, this additional layer of robustness coming from the parametric analysis might compensate for some of the remaining imbalance. Rubin (2001) suggests three simple guidelines for assessing whether sufficient balance for regression analysis has been established: 1) the average distance in propensity scores is less than a standard deviation apart; 2) the ratio of propensity score variances between treatment groups is close to one; 3) the ratio of residual variance of the regression of X on a linear $\hat{p}(X)$ between treatment groups is also close to one.

In case the balancing tests detect remaining imbalance, a re-specification the propensity sore, the use an alternative neighborhood, the use of an exact matching schemes, etc. can be attempted to improve the balancing quality. The balancing weighs yielding the highest degree of balance should be selected. When balance cannot be achieved even after extensive re-specification attempts, efforts could be focussed on balancing a subset of characteristics. Alternatively, one might consider the reduction of the estimation sample to a subpopulation for which balance can be established in all characteristics.

Tests for differences in means The standardized bias (SB) and the two-sample t-test are most commonly used to assess the pre-balance and post-balance differences in variables means (see, e.g., Rosenbaum and Rubin, 1985a). In contrast to the parametric t-test, the SB is not a statistical test, and the calculated indicator does not follow a parametric distribution. However it has the advantage as being easily interpretable as the percentage of a standard deviations difference. Let $\bar{X}_U^D$ denote the mean of characteristic X across treated and controls respectively, in the unbalanced sample, and $\bar{X}_B^D$ the analogue in the balanced sample, using the weights assigned by matching or weighting. The respective standard

deviations of characteristics are represented by S_X^D. In the SB, the level of pre- and post-balancing similarity characteristics across treatment status are given by

$$\text{SB}_U = 100 \cdot \frac{\bar{X}_U^1 - \bar{X}_U^0}{\sqrt{(S_X^1 + S_X^1)/2}} \quad \text{and} \quad \text{SB}_B = 100 \cdot \frac{\bar{X}_B^1 - \bar{X}_B^0}{\sqrt{(S_X^1 + S_X^1)/2}}.$$

As the denominator remains the same, the pre-and post balancing values of the SB are directly comparable. A cut-off value of SB_B signifying a sufficient reduction in imbalance does not exist, it should be attempted to get a close to zero as possible.[9] Similarly, the p-value of the t-test, should be maximized without limit rather than using conventional significance values (see above). Both the SB-statistic and the t-test are sensitive to sample size, as an increase in the number of matched controls might add to the variability of the control characteristics (Smith and Todd, 2005b). To obtain a joint measure of balance, the Hotelling-test can be applied (Lee, 2013).

Regression Test Smith and Todd (2005b) propose a regression test, based on the idea that in a balanced sample the treatment indicator and the characteristics should be independent of each other. Hence, a regression of each characteristic X_l on a polynomial of a given order K of the propensity score and the interaction between a treatment dummy and the propensity scores should not have much explanatory power. The model to be fitted is given by

$$\begin{aligned} X_l \;=\; & \beta_0 + \beta_1 \hat{p}(X) + \beta_2 \hat{p}(X)^2 + \beta_3 \hat{p}(X)^3 + \ldots + \beta_k \hat{p}(X)^k \\ & + \alpha_0 D + \alpha_1 D\hat{p}(X) + \alpha_2 D\hat{p}(X)^2 + \ldots + \alpha_k D\hat{p}(X)^k + \varepsilon \end{aligned}$$

The F-test of a test on joint insignificance of all coefficients α_k involving the treatment dummy in the model should be as low as possible in the balanced sample, as D should not contain information on the treatment status after having controlled for the propensity score. Note, that the order of the polynomial has to be chosen by the researcher, the size of which might influence the results one gets. In a very similar approach, Sianesi (2004) suggests to refit the propensity score model on the balanced sample to see whether the X's have any remaining explanatory power. In a balanced sample, the model should not be able to distinguish between treated and controls anymore, so that an F-test on joint insignificance should lead very

[9] Rule-of-thumb values of five or higher that are often used in the literature can be used as indicator for extreme imbalance. Very good balance represent SB-values close around one, or even less than one.

large p-values.

Quantile-Quantile-Plots A visual approach to assessing differences in continuous variables is the comparison of quantile-quantile plots (eQQ-plots). Recall, that the empirical quantile function provides for each probability p the value of $X = x$, such that the cumulative distribution function $F(x) = P(X \leq x)$ is at most p, i.e., $Q_X(p) = \inf\{x : F(x) \geq p\}$. The quantile-function can hence be obtained as the inverse of empirical distribution function: $Q_X(p) = F_X^{-1}(p)$. Based on a non-parametric estimate of the distribution function (CDF) for treated and reweighed controls respectively, the quantile values $q_x^1(p)$ and $q_x^0(p)$ of treated and controls are calculated.[10] In eQQ-plots, they are then plotted against each other in the same graph. Figure (4.2) provides an exemplary eQQ-plot of a continuous pre-treatment variable for treated and controls, before and after kernel matching.[11] Circles denote the quantiles in the raw sample, and crosses denote balance in the reweighed sample. From the distance to the diagonal it can be seen that balance could be improved by reweighting, but that there remains imbalance for values of X above 5000.

The quantile difference can also be expressed in one summary measure, e.g., the average quantile difference: $\frac{1}{n}\sum_p |q_x^1(p) - q_x^0(p)|$; similarly, the median or the maximum of the differences can be calculated. Note, however, that the average quantiles may hide deviation in case of both positive and negative deviation from balance, as they might cancel out on average (Sekhon, 2007).

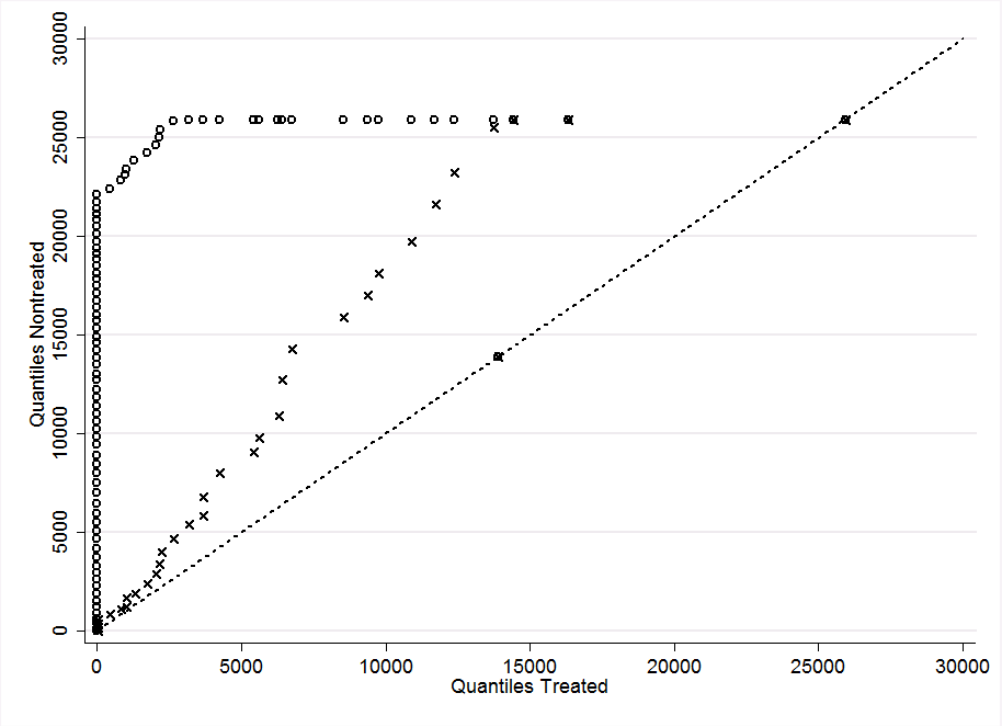
Tests for similar empirical distributions An alternative approach to test balance in continuous characteristics is to conduct non-parametric tests on the equality of two distributions can be used. Based on non-parametric estimates of the empirical distribution function (CDF), the Kolmogorov-Smirnov (KS) test compares the supremum of the differences in the empirical distribution functions $\hat{F}_N^D(X), D = 0, 1$

$$D_n^{KS} = \sqrt{\frac{N_0 \cdot N_1}{N}} \sup_x |\hat{F}_N^1(X) - \hat{F}_N^0(X)| \sqrt{\psi(\hat{F}(X)))}.$$

[10] The quantiles are often chosen as in $\tilde{p} = (i - 0.5)/N_q, i = 1, \ldots, N_q$ with N_q denoting the total number of quantiles.)

[11] Here we use the "cps1re74" data as constructed in LaLonde (1986), with the treatment being participation in a training program. The variable to be balanced is the pre-treatment earnings variables re74. We used kernel matching with an Epanechnikov kernel and a bandwidth size of 0.06 to calculate the PSM weights.

Figure 4.2: Quantile-Quantile plots of the pre-treatment earnings distribution before and after PSM matching



The two-sample Cramer-von-Mises-test (CMS) takes the integral over the differences, whereby the integral can also be replaced by a sum in case of discretized variable values,

$$W_n^2 = \frac{N_1 \cdot N_0}{N} \int_{-\infty}^{\infty} [\hat{F}_N^1(X) - \hat{F}_N^0(X)]^2 \psi(\hat{F}(X)))dF(X).$$

The $\psi(\hat{F}(X)))$ represents a weighting function that allows to vary the importance of different parts of the distribution.[12] Although the KS-test is much more common in applied work, (usually with $\psi(\cdot) = 1$), the CMS-test can be a more powerful alternative, (e.g., Stephens, 1974). Alternative to using the empirical distribution function for comparison in the above-mentioned tests, Huber (2011) proposes to compare quantiles functions (also see Koenker and Xiao, 2002; Chernozhukov and Fernandez-Val, 2005). As before, the balanced quantiles of controls are obtained by inversion of the reweighed CDF. Huber (2011) suggests that the size of the KS-test can be significantly improved when using quantiles rather than the CDF,

[12]Examples of weights are the inverse of the variance of the difference or in the case of balancing tests, the densities of propensity score distributions. Anderson and Darling (1952) show that weighting may improve the power of both tests (see, e.g., Büning, 2001), although it is not often applied.

and using weights that down-weigh areas with high variance or low support. He finds that differences between the KS and the CMS test are negligible.

For either test, equality of distributions is rejected when the statistics are too large. The distribution of the test-statistics under the null-hypothesis is usually derived via bootstrapping or permutation (see, e.g., Abadie, 2002). When using quantile functions, it is advised to re-center the bootstrap distribution to the mean of the original sample as this may improve the power of the test significantly (Chernozhukov and Fernandez-Val, 2005).

Multidimensional balance measures Iacus et al. (2011) suggest a single measure of the overall balance, which relies on the multivariate comparison of imbalance across subgroups. Their measure is based on the L_1 distance of the two groups which is represented by a multi-dimensional cross-tabulation of all characteristic-combinations $X_1 \times \ldots \times X_k$ of treated and controls. To reduce the dimensionality of the characteristic combinations, each characteristic is "coarsened", i.e., grouped into sensible bins. Within the cells, the cell-frequencies $f_{\ell_1 \ldots \ell_k}$ of treated and controls $g_{\ell_1 \ldots \ell_k}$ are calculated and compared,

$$\mathcal{L}(f, g) = \frac{1}{2} \sum_{\ell_1 \ldots \ell_k} |f_{\ell_1 \ldots \ell_k} - g_{\ell_1 \ldots \ell_k}|, \tag{4.6}$$

with $\mathcal{L}(f, g) = [0, 1]$. Any deviations from perfect congruence increases the measure, so that a value of zero is unlikely to be reached. A sufficient level of $\mathcal{L}(f, g)$ is not defined so that only relative comparisons of pre- and post-weighting balance can be made. Note, that the measure does not account for the relative importance of individual confounders, and should hence be used as an auxiliary balance indicator.

An alternative "omnibus" test for covariate balance following stratification is proposed by Hansen and Bowers (2008). Here, the standardized mean differences are calculated within the respective strata, the overall balance measures is then constructed by weighted aggregation of the strata-specific effects. Their approach hence accounts for the stratified nature of the data, but could also be used on the overall sample. They show that the test-statistic follows a χ^2-distribution. See Table A4.1 for statistical software.

4.8 Conditional Outcome Differences

Having performed steps 1 through 5 to estimate the balancing weights, these balancing weights can be further used to reweigh the control outcomes, and thereby construct the conditional control outcome Y_0^C that would occur had the controls the same characteristics as the treated. Many if not most empirical applications of balancing methods in, e.g., treatment evaluation or decomposition analysis, investigate the difference between the outcomes of treated and the reweighed controls as their main parameter of interest. When the focus is on the average difference in outcomes, the parameter of interest is estimated as the difference between the average of treated and reweighed control outcomes over the region of common support,

$$\Delta_{ATT} = \bar{Y}_1 - \bar{Y}_0^C = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{1i} - \frac{1}{N_0}\sum_{j=1}^{N_0} \omega_{0j} Y_{0j}. \tag{4.7}$$

whereby the weights are calculated by PSM or IPW as outlined in Section 4.2.

In case of a continuous outcome variable, one might be interested in estimating the differences over the whole of the distribution rather than at the averages. Let $\hat{f}_D(Y), D = 0, 1$ denote estimates of the unconditional density function for each treatment group. In practice this is done using a kernel density estimators $\hat{f}(Y_D) = 1/N_1 h \sum_i K(\frac{Y_D - Y_{iD}}{h})$. The counterfactual treatment distribution is obtained by multiplying the balancing weights in kernel weights, hence

$$\Delta_{f(X)} = 1/N_1 h \sum_i K(\frac{Y_1 - Y_{1i}}{h}) - 1/N_0 h \sum_{j=1}^{N_0} \bar{\omega}_{0j} K(\frac{Y_0 - Y_{0j}}{h}). \tag{4.8}$$

Similarly, the conditional differences may be obtained at specific quantiles of the distribution of Y_D. Recall, that quantile $Q_p(Y_D)$ with $p \in [0, 1]$ can be obtained by the inverse of the cumulative distribution function of $\hat{Q}_p(Y_D) = \hat{F}^{-1}(Y_D) = \inf\{Y_D : \hat{F}(Y_D) \geq p\}$, whereby the reweighed CDF of controls is simply obtained by $\hat{F}(Y_D) = 1/N_D \sum_i \omega_i \mathbb{1}(Y_{iD} < Y_D)$. For applications of either approach see, e.g.,Bitler et al. (2006); DiNardo et al. (1996); Frölich (2007b); Fortin et al. (2011); Firpo (2007).

By the non-parametric nature of these differences-in-distribution moments estimators, the variance of the outcome difference are most commonly estimated using resampling methods. However, for the differences-in-means estimators analytic standard errors are also available. In practice, the use of resampling methods

is conventionally preferred, as it allows to take account of the whole estimation process, including the estimation of the propensity scores, and the implementation of common support as this variability to the treatment effect estimate (Heckman et al., 1998). We address inference for the estimated parameters in the subsequent Section 4.8.1.

To increase the robustness of differences-in-means estimator, balancing can be combined with parametric regression analysis. The parametric estimation of outcome differences on the matched sample is expected to improve efficiency and robustness of the parameter estimate by taking care of any remaining imbalance (Rubin, 1973), and may mitigate the outlined sensitivity of balancing estimators to mild non-linearities in the area of low overlap (see Busso et al., 2014a,b, and also Section 4.5). Also, when introducing IPW weights in a regression estimate of outcome differences, the estimator may exhibit a doubly robust feature in the sense that it is consistent even when either the outcome or the treatment model are misspecified (e.g., Robins and Rotnitzky, 1995). We discuss these and further combinations of methods in Section 4.8.3.

4.8.1 Inference

Bootstrapping and Subsampling The most commonly used resampling method is bootstrapping (Efron, 1979), which is based on the assumption that the estimation sample consists of a random draw of size N of i.i.d observations from an unknown distribution F. Hence, when drawing a further random i.i.d sample of size N with replacement from the estimation sample, another random draw from the unknown distribution is obtained. By drawing a large number B of random samples, the so-called bootstrap distribution $\hat{F}_B(\cdot)$ approximates the true distribution of a parameter of interest. Aiming to estimate the distribution of a parameter τ, the bootstrapped standard error is obtained by estimating $\hat{\tau}_b^*$ on the bootstrapped sample at each draw $b = 1, ..., B$.

There are multiple ways to obtain confidence band in the $100(1-\alpha)$ - interval. The most intuitive way is to order the values of $\hat{\tau}_b^*, b = 1, ..., B$ and use the percentiles of this empirical bootstrap distribution $\hat{F}(\hat{\tau})$ directly as in $[\hat{F}^{-1}_{\alpha/2}, \hat{F}^{-1}_{1-\alpha/2}]$. This approach requires that the bootstrapped distribution of τ closely replicates the true, unknown distribution of τ, which is difficult to do in practice (Davison and Hinkley, 1997) and requires bootstrap replications in the magnitude of

$B = 1000$. Making the simplifying assumption that τ follows a normal or Student-t distribution, the variance estimate of τ can be used to estimate the confidence intervals in $[\hat{\sigma} \cdot (\hat{\tau}) \cdot z_{\alpha/2}, \hat{\sigma}(\hat{\tau}) \cdot z_{1-\alpha/2}]$, with the variance estimate being (Efron and Tibshirani, 1986)

$$\hat{\sigma}(\hat{\tau}) = \left(\frac{\sum_{b=1}^{B} (\hat{\tau}_b^* - \hat{\tau}^*)^2}{B-1} \right)^{\frac{1}{2}}, \tag{4.9}$$

and $\hat{\tau}^*$ representing the mean estimate of $\hat{\tau}$ over all bootstrap draws, i.e. $\hat{\tau}^* = 1/B \sum_{b=1}^{B} \hat{\tau}_b^*$. More involved methods account for strongly skewed distributions or the bias of estimates. See for further discussion Efron and Tibshirani (1986); DiCiccio and Efron (1996); Davison and Hinkley (1997), and the special case of confidence intervals for quantiles of the distribution Hall and Martin (1989) and Ho and Lee (2005).

In general, the bootstrap approach is not appropriate for all sample statistics (see, e.g., Horowitz, 2001), so that the bootstrapping approach should only be used if a formal justification of its use exists. Abadie and Imbens (2008) show that for k-NN matching with replacement the bootstrapped variance is not expected to estimate the true variance precisely, as the number of neighbors k is not data-adaptive and can hence not be approximated with bootstrapping. For other balancing methods (kernel, local linear, etc., matching without replacement and IPW), the frequency with which a control is used in a match is a direct function of the control sample size, so that resampling is expected to yield reliable variance estimates. Ham et al. (2011) provide evidence that bootstrapping yields consistent estimates of standard errors for local linear, local cubic and local linear ridge matching. To account for the additional variation of propensity score estimation, bootstrapping is done at the level of the estimation sample.[13]

As a resampling alternative for k-NN matching, Abadie and Imbens (2008) propose to use subsampling, i.e., permutation, rather than bootstrapping, as it bears the advantage of a higher robustness to certain data features that lead to the failure of bootstrapping. Subsampling is first proposed in Politis and Romano (1992, 1994), and bases on the idea of drawing a subsample $N_S \ll N$ that is smaller than the sample size N, without replacement, thereby mimicking a draw from the

[13]Recently de Luna et al. (2010) suggest a resampling strategy for k-NN matching based on block bootstrapping of the treatment effect, instead of the whole sample as in Abadie and Imbens (2008). Besides ignoring the estimation procedure, their simulations show that theses methods are less robust than the analytic standard errors provided by Abadie and Imbens (2006), so that the applicability of these methods have to be corroborated by further research.

original population sample. As more draws are taken, the confidence interval for the distribution parameters can be retrieved. This requires that $N_S \to \infty$. A downside of the approach is that it is less efficient than bootstrapping and hence requires large sample sizes and number of draws to achieve a reliable confidence interval. Furthermore it is required that the "convergence rate" is estimated from the sample; details of this procedure can be found in Politis et al. (1999).

Analytic standard errors Alternative to using resampling methods the standard errors of the treatment effect estimates can be calculated analytically, based on large sample theory. Note, that the marginal or overall variance $\mathbb{V}_\tau$ of the treatment effect can be expressed as the sum of variances in the two groups,

$$\mathbb{V}_\tau = \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \sigma_1^2(Y_i) + \sum_{i=1}^{N} (1 - D_i) \omega_i^2 \sigma_0^2(Y_i), \tag{4.10}$$

with $\sigma_D^2(Y_i)$ denoting the conditional outcome variance (Lechner, 2001; Imbens and Wooldridge, 2009). The estimation of the conditional outcome variance $\sigma_D^2(Y)$ for treated and controls respectively can be done using nonparametric kernel regressions. Abadie and Imbens (2006) propose to use an easier-to-implement alternative estimator, which is based on matching individuals within the groups of treated and within the group of controls respectively. In particular, they suggest the variance formula to take the form

$$\hat{\sigma}_D^2(Y_i) = \frac{J}{J+1} (Y_i - \frac{1}{J} \sum_{m=1}^{J} Y_{j(i)}), \quad D = 0, 1 \tag{4.11}$$

with $Y_{m(i)}$ denoting the m-th closest individual to i in terms of the propensity score, with a fixed number of matches J. While this is not a consistent estimate of the variance it is unbiased asymptotically. Based on an estimate of $\hat{\sigma}_D^2(Y_i)$, the variance of the treatment effect $\hat{\mathbb{V}}_\tau$ can be calculated (Imbens and Wooldridge, 2009).

4.8.2 Sensitivity Analysis

As the the conditional independence assumption cannot be tested, the assumption is best defended by good knowledge of the relevant observables X and an appropriate data source (see Section 4.3). Furthermore, several indirect assessments of the plausibility of the CIA assumption can be used to corroborate the claim of

causality, and/or assess the sensitivity of the estimated parameter to unobserved confounding.

One approach to sensitivity analysis aims to detect non-random unobserved confounders by the estimation of "pseudo" treatment effects that are *known* to be zero. By reweighing an outcome that is clearly not affected by the treatment, e.g., historical outcome values, no differences should emerge between treated and controls, provided that conditional independence holds. Alternatively, a zero "pseudo" treatment effect should arise when two non-treated outcomes are balanced. In particular, when there are non-eligible and eligible non-participants, the treatment effect could be estimated using one group as treatment group. A non-zero treatment effect estimate is clearly suggestive of unobserved confounding in one of the control groups. Note, however, that a zero estimated treatment effect may be indicative of either no confounding or the same unobserved confounding across non-treated groups. See Imbens and Wooldridge (2009) for a detailed analyses of these approaches.

An alternative strand of sensitivity analysis directly models the degree of influence of potential unobservable confounding required to invalidate the qualitative finding of treatment effect estimates. Here, the treatment effect estimate is re-estimated taking explicit account of one unobserved hypothetical confounder U, making specific assumptions of the relation of U with the treatment selection D or the outcome of interest Y. By varying the hypothetical influence of U, alternative treatment effect estimates are calculated. The variability of the estimates is then used as in indicator of the sensitivity of treatment effect estimate under conditional independence and the "degrees of unconfoundedness" necessary to invalidate the results. In the following we outline two of these approaches that use non-parametric sensitivity analyses. Further parametric sensitivity analyses of this type are found in Gastwirth et al. (1998); Imbens (2003); Altonji et al. (2005); Lee and Lee (2009). The relative performance of the different sensitivity analyses has not been subject of thorough investigation, further research is needed to compare the different approaches.

Rosenbaum bounds In the balanced sample, the odds of receiving treatment should be the same across treatment groups unless systematic unobserved confounding remains. Rosenbaum (1987, 2002) proposes a non-parametric bounding approach that is based on the assumption that a binary unobserved confounder U_i affects the conditional selection probability, i.e., $P(D = 1|X) = \beta X_i + \gamma U_i$. They

show that this set-up implies that the confounded conditional odds of receiving treatment are bounded. The size of bounds depends on the assumed size and direction of γ. Assuming a logistic regression of the treatment model, the bounds are given by

$$\frac{1}{\exp(\gamma)} \leq \frac{\exp(\beta X_i + \gamma U_i)}{\exp(\beta X_j + \gamma U_j)} \leq \exp(\gamma). \tag{4.12}$$

The degree of confounding can hence be expressed in terms of the odds-ratio: assuming a confounder of strength γ, the conditional odds of treatment differ with a factor of $\Gamma = \exp(\gamma)$, if $\Gamma > 1$. To assess the impact of different levels of confounding Γ on these bounds, Rosenbaum (1987, 2002) modifies non-parametric tests of no treatment effect to accommodate Γ. The type of non-parametric test to be used is determined by the outcome (continuous, binary, ordinal), and the structure of the data (number of matched controls, strata).[14] Based on a choice of Γ, the test-statistic calculates the rejection probability of a non-zero treatment effect. As the influence of U can be positive or negative, two one-sided tests are conducted with the respective opposing Γ^+ and Γ^--values. The sensitivity of estimates is then assessed by comparing the p-values across different scenarios. For example, when the p-value indicates a treatment effect of zero already for small deviations of Γ from one, the estimate is very sensitive to even very small degrees of confounding. For some types of tests, the Rosenbaum bounds have been implemented in statistical packages, compare Table A4.1.

Note, that the proposed test only makes assumptions about the strength of U on D, but not on Y. Gastwirth et al. (1998) extend this analysis, assuming a binary treatment and a binary outcome, and two sensitivity parameters γ and δ, influencing the treatment and the outcome probability, respectively. Only when $\gamma\delta \neq 0$, bias due to unobserved confounding will arise. A comparison of the results of the Rosenbaum bounds and the modification by Gastwirth et al. (1998) show that the Rosenbaum approach is often conservative in that it rejects unconfoundedness where the dual sensitivity approach does not (Lee and Lee, 2009).

The Rosenbaum-Rubin Approach The sensitivity test proposed by Rosenbaum and Rubin (1983a) assumes a binary confounder U that is both correlated with D and outcome Y. Based on the specification of the correlation structure

[14]Note, that these tests only rely on the assumption of random assignment, so in the absence of unconfoundedness $\gamma = 1$, they could also be used to calculate significance of the treatment effect estimate in the balanced sample, compare, e.g., Aakvik (2001).

between U and D and Y, respectively, a variable $\hat{U}$ can be simulated and the effect estimate is then recalculated including $\hat{U}$ in the set of confounders. The variability of these effect estimates is taken as indication sensitivity to unobserved confounding. Imbens (2003) offers a parametric version of this, Ichino et al. (2008) propose a non-parametric one that we outline in the following.

Ichino et al. (2008) assume a binary treatment and a binary outcome measure, so that the distribution of the binary confounder can be expressed via four conditional probabilities, capturing a different treatment-outcome-combination,

$$Pr(U = 1|D = i, Y = j, X) = Pr(U = 1|D = i, Y = j, X) = p_{ij}, \tag{4.13}$$

with $i, j \in \{0, 1\}$. This hence allows to model positive or negative selection into treatment. By the definition of these probabilities U is fully characterized and can be simulated. The effect of the simulated confounder can be expressed by the conditional odds-ratio of treatment participation and a positive outcome. While Γ reflects the "outcome effect" of U, the parameter Λ reflects the "selection effect",

$$\Gamma = \frac{\frac{P(Y_1=1|D=0,U=1,X)}{P(Y_1=0|D=0,U=1,X)}}{\frac{P(Y_1=1|D=0,U=0,X)}{P(Y_1=0|D=0,U=0,X)}} \quad \text{and} \quad \Lambda = \frac{\frac{P(D_1=1|U=1,X)}{P(D_1=0|U=1,X)}}{\frac{P(D_1=1|U=0,X)}{P(D_1=0|U=0,X)}}. \tag{4.14}$$

The definition of the parameters p_{ij}, Γ, Λ allows to concisely model the size and strength of the unobserved confounding without having to rely on parametric assumptions. To assess whether a particular degree of hypothetical confounding is reasonable in a given setting, Ichino et al. (2008) propose to model the probabilities p_{ij} similar to the conditional sample probabilities of some observed binary confounder (compare for a similar reasoning Imbens, 2003). They further provide an implementation in statistical software (see Table A4.1).

Trimming A further sensitivity approach suggests to restrict the estimation sample to a bias-minimizing subset of the whole sample. In the theory of sample selection (see, e.g., Heckman, 1979; Heckman and Navarro-Lozano, 2004), the omitted variables bias of treatment effect estimates is due to the correlation structure of the unobserved confounders, so that the bias of the ATT estimator can be shown to be minimized at $p(X) = 1/2$, under joint normality of the errors (compare Black and Smith, 2004; Millimet and Tchernis, 2012, for details). Black and Smith (2004) therefore suggest to restrict the sample to the center of the propensity score distribution, i.e., between $p(X_i) \in (0.33, 0.67)$, in order to obtain an estimated with the least bias. Stürmer et al. (2010) provide a similar intu-

itive reasoning suggesting to eliminate non-treated with high propensity scores, as treatment has then been probably withheld on purpose based on unobserved confounders, as well as treated with very low propensity scores, as treatment is then probably a "last-resort" measure for some unobserved condition.

Millimet and Tchernis (2012) suggest some modifications to this approach. As very large sample reduction increase the variance of the estimate, they propose to put a limit to the reduction of the sample size, by imposing that at least a certain share $0 < \alpha < 1$ of treated and controls be retained in the radius around $p(X) = 1/2$. Further, they propose a "bias-correction" term for the estimator, that accounts for potential deviations from normality. Recall, however, that by restricting the estimation to a subset of the full sample, the estimated treatment effect changes its meaning. In particular in case of heterogenous treatment effects, the emergence of different findings between the full and the restricted sample might be either due to effect heterogeneity or the elimination of estimation bias.

4.8.3 Additional outcome analysis

A number of studies have outlined the benefits of combining non-parametric weighting and parametric outcome analysis. In the subsequent chapter we outline a number of ways to do so. While these combinations may assist in improving consistency and efficiency of the estimated parameters, they are not expected to be of any help in defending the conditional independence assumption. However, by combining balancing with difference-in-difference and instrumental variable approaches, alternative identification assumptions might help to remove any bias in the ATT due to remaining unobservable characteristics. We hence also briefly outline to suggestions to combine reweighing with these estimation approaches. In the absence of these additional identification assumptions, a number of sensitivity analysis has been suggested that help to assess the sensitivity of the estimator to potential violations of the CIA assumptions. We briefly outline the different approaches to sensitivity analysis.

It is insightful to briefly recall the assumptions and the practical implementation of the ATT estimator using parametric regression models. The regression estimator assumes linearity in it's regressors and separability between the observed and unobserved factors. The outcome equations presented in Section 4.2 (see equation 4.1) is hence assumed to be given by $Y_{Di} = \beta_D' X_i + U_{Di}, D = 0, 1$, with $E(U_{Di}|X_i, D_i) = 0$. The conventional ATT estimator based on regression

analysis is given by

$$\Delta_{ATT}^{reg} = E[X|D=1](\hat{\beta}_1 - \hat{\beta}_0), \tag{4.15}$$

which is implemented using the sample moments $E[X_i|D=1] = 1/N_1 \sum_{i:D=1} X_i$ and the coefficients obtained from the group-specific regression of outcomes Y_{Di} on the explanatory variables and an intercept. Note, that this estimator was made familiar by the linear decomposition analysis, (see,e.g., Oaxaca, 1973; Blinder, 1973), where this parameter arises as the "unexplained" part of the difference $E(Y|D=1,X) - E(Y|D=0,X) = \bar{X}_1(\hat{\beta}_1 - \hat{\beta}_0) + (\bar{X}_1 - \bar{X}_0)\hat{\beta}_0$.

The Δ_{ATT}^{reg} estimate is consistent under the assumption that $\bar{Y}_0^C = \hat{\beta}_0\bar{X}_1$ is a good approximation of the average conditional control outcome around $\bar{X}_1$. This assumption is not likely to hold if the mean value of characteristics of $\bar{X}_0$ in the control sample, on which $\hat{\beta}_0$ was estimated, is very different from $\bar{X}_1$ and the relationship between Y and X is not linear (Imbens and Wooldridge, 2009; Fortin et al., 2011). While the estimates based on non-parametric reweighing do not require this assumption, the combination of (pre-) balancing of characteristics in the two samples may also improve the robustness of the regression-based estimator (Ho et al., 2007). In the following we hence outline several practical approaches to combine non-parametric weighting and parametric regression analysis in practice.

Weighted Regression Analysis The most intuitive way to use the balancing weights obtained by PSM or IPW is to use them in a weighted outcome regression. Note, that in the reweighed sample the distribution of outcomes are rendered unconditionally independent of the treatment status. Hence, on the reweighed sample, the difference in conditional means estimator outlined in equation 4.7 can also be obtained by conducting a weighted regression of the Y on a constant and D, using weights that are equal to $\bar{\omega}_{0j}$ if $D = 0$ and 1 if $D = 1$ (Busso et al., 2014b). Clearly, further covariates can be added to this regression equation - the effect estimate is given by the coefficient $\tau = \Delta_{wreg}^{ATT}$ in the reweighed regression model (Imbens, 2004),

$$Y_i = \beta' X + \tau D_i + \varepsilon. \tag{4.16}$$

The benefits of the combination is twofold. First, the estimator is expected to improve in precision compared to estimation models using either method, although Busso et al. (2014b) note the increase is expected to be relatively small. Secondly, and more importantly, this type of estimator has a higher probability of being consistent, as it exhibits the so-called double robust property (Robins and Rotnitzky,

1995; Robins et al., 1995; Scharfstein et al., 1999) that maintains the consistency of the estimator even if either one of the models is mis-specified.

When both models are misspecified, the bias of the double robust estimator may be magnified over the simple reweighing estimators. This also pertains to mis-specification coming from a wrong parametrization of the outcome equation, i.e., choosing a linear model, when a logistic model is more appropriate (Wooldridge, 2007). Kang and Schafer (2007) show that even slight mis-specifications in both models lead to very strong biases that are significantly higher than the one arising from IPW2. To ensure that both models are correctly specified extensive specification checks should be conducted, Hirano and Imbens (2001) a systematic selection procedure for characteristics to be included in the treatment and the outcome model. Note, that Δ^{ATT}_{wreg} is expected to be plagued with similar problems of extreme weights as simple rebalancinng using IPW, so that long tails in the distribution need to be taken care of (Robins et al., 2007). However, due to the parametric regularization, the double robust estimator tends to perform better than simple reweighing when there are slight non-linearities in the relation between characteristics and controls (see Busso et al., 2014a, and also compare Section 4.5.5).

Matching and Regression Further suggestions have been made in the literature to accommodate the matched sampling structure in the regression analysis of outcomes. While this is relatively easy to obtain in pair matching (Rubin, 1979), Abadie and Imbens (2011) extend this approach to $k > 1$ matching, and propose to construct an artificial "pair-matched" sample of size N_1. Here, the characteristics and outcomes of the "paired" controls are constructed as the weighted average of the k matched controls for each individual separately. The regression adjustment is then conducted on this artificial pair-matched sample. They show that when combined with the bias-correction, this approach perform quite well in terms of bias, although it clearly loses in terms of precision by the reduced sample size.

Bias-corrected matching Abadie and Imbens (2011) propose a bias-correction for the matching estimate of the conditional outcome Y^C_0 using within-match regression analysis to reduce bias arising from inexact matches in k-NN or radius matching. In particular they suggest the estimation of a correction term $\bar{Y}^C_{0-BC} = \bar{Y}^C_0 + \nu_{BC}$, which captures in outcomes arising from differences in matched

propensity scores. The correction term is calculated by fitting a local regression model of outcomes on characteristics X within a match, using both treated and control observations — the differences in predicted values between treated and controls outcomes is the bias-correction. When distance-based weights are used in matching, they should be also used in the local regression. Abadie and Imbens (2011) suggest that bias-corrected matching is more robust to the choice of k than conventional k-NN matching estimators. While Huber et al. (2013) conclude that bias-corrected radius matching performs substantially better than most other reweighing estimates, it is also important to note that the estimator requires that the local regression model is correctly specified. Hence, when there are strong local non-linearities, and/or only a few neighbors are used, the bias-correction tends to aggravate the bias (Busso et al., 2014b).

Subclassification and Regression When balancing is established via subclassification, local regression models may be fit in the respective strata. If the strata are fairly large, there main remain imbalance across score values within strata so that the regression-based estimate is likely to substantially improve the difference-in-means-estimator (Lunceford and Davidian, 2004). Furthermore, compared to regression analysis on the full sample, the local treatment effect estimates are less prone to extrapolation (Imbens and Wooldridge, 2009). By the reduction in sample size within strata it may not be possible to include all X_i in the estimation, the included variables should however be chosen to best approximate the local outcome equation (Lunceford and Davidian, 2004).

Inverse Probability Tilting Graham et al. (2012) propose an alternative way to inverse probability weighting (called inverse probability tilting, PIT), where the propensity scores are not estimated via maximum likelihood but as the solution to a method of moments problem, that exploits the weighting equalities $E[DY/p(Z)] - E[Y_1] = 0$ and $E[DY/(1-p(Z))] - E[Y_0] = 0$. Provided that the outcome model can be expressed as a linear transformation of the propensity score model, and Z is the union of elements necessary to consistently estimate the outcome model and the treatment model, this condition ensures, that the parameters of the propensity score model are estimated to exactly balance characteristics across treatment groups. They show that under the above-mentioned assumptions, the Δ_{ipt}^{ATT} estimator is semi-parametrically efficient and has the double robust feature. As in the Δ_{wreg}^{ATT} estimator, efficiency is lost if either model is misspecified.

So far, the attractive theoretical features of IPT have not been subject to much empirical scrutiny. Simulation evidence by Busso et al. (2014b) remains somewhat inconclusive, noting that the moment estimator may fail to produce an estimate of the propensity score in the case of particularly disparate distributions of characteristics of treated and controls, and may lead to non-normalized weights that imply higher variability than the standard IPW2 estimator. In terms of bias, the IPT estimator seems comparable to IPW2.

Matching and difference-in-differences One approach to removing any remaining *time-invariant* differences in unobservable characteristics, is to combine matching with a difference-in-difference (DID) estimator (see, e.g., Heckman et al., 1997) The DID-approach estimates the treatment effect by the change in outcome differences between treatment groups, before the treatment $t = 0$ and after the treatment $t = 1$. It hence requires panel data or data from repeated cross sections where the outcomes of treated and controls (Y_t^0, Y_t^1) are observed at both points in time. The DID estimator removes any systematic differences in characteristics across treatment status are constant over time, the combination with matching might remove further pre-treatment differences that might affect a differential time-trend between the two points of observation (see,e.g., Heckman et al., 1997, 1998; Abadie, 2005; Buscha et al., 2012, for applications of the DID and the DIDID method).

Chabé-Ferret (2012) discusses the consistency of the DID estimator and standard matching estimator in the presence of transitory shocks to past outcome values. As outlined in Section 4.3.1 it is usually a good idea to include past outcomes in the propensity score specification as these are good predictors of future outcome values and likely to capture systematic differences between treatment group. Chabé-Ferret (2012) argues however, that transitory shocks might render these outcomes values uninformative and forcing balance on them might even exacerbate bias due to unbalanced unobserved characteristics. While the matching estimator is always inconsistent in these setting, he shows that under certain conditions (e.g., symmetry of the shock with respect to the points of measurement) the DID matching is consistent. However, when these conditions do not hold, the DID matching will lead to highly variable results, whereas simple matching tends to systematically underestimate the parameter. In the absence of more information on the persistence of the shock it is hence advised to compare the performance of both approaches.

Matching and Instrumental Variables Recently Costa-Dias et al. (2013) propose a combination of matching with an instrumental variable approach. Prerequisite of this approach is the existence of exogenous and discontinuous eligibility rule to be used as instrument Z that shifts the participation probability to zero for some value of Z, whereas for all remaining values self-selection is present. These types of instruments are very common, e.g., in the assignment to active labor market policies, as eligibility to participate in a program is based on strict age or regional cutoffs. In this application this cut-off serves to define a second control group (the non-eligible) that has not been affected by treatment *and* did not undergo any selection. Hence, assuming that the instrument and the potential outcomes are conditionally independent ($Y_0 \perp\!\!\!\perp Z|X$) the difference in conditional outcomes (conditional on the distribution of observables X in the distribution of treated) of eligible non-participants (control group 1) and ineligible non-participants (control group 2), serves to identify unobserved selection, and can be used as a bias-correction for the conventional treatment counterfactual based on the difference between participants and control group 1. Costa-Dias et al. (2013) note that the magnitude of the bias correction can also be used as a test for the CIA assumption, as the bias correction will tend towards zero in the absence of unobserved confounders (also compare Section 4.8.2).

4.9 Further Balancing Issues

4.9.1 Automated Balancing

As outlined in Section 4.7, balance-checking is a fundamental part of the implementation of matching and weighting that can become quite cumbersome as multiple iterations of balance-checking and re-specification have to be conducted. Two recent contributions suggest to circumvent the lengthy iteration procedure by using matching and weighting algorithms that automatically maximize balance across treatment groups by iteratively adjusting parameters of the balancing procedure until the highest possible balance level is achieved.

Entropy balancing The entropy balancing algorithm by Hainmueller (2011) is similar to the idea of IPW in that the controls are reweighed using normalized weight that balance the characteristics of treated and controls to create the counterfactual. Instead of deriving the weights from an estimate of the propensity score, entropy balancing directly calculates weights which create balance in the sample

moments of characteristics across treatment groups. This is achieved via solving an optimization problem that aims to minimize the deviations of the balance weights from some baseline weights (e.g., uniform weights, sampling weights) subject to a number R of balancing constraints that reflect the equality of sample moments across treatment group. When the distributions moments are far too dispersed the algorithms may fail to produce a set of weights to balance all relevant sample moments, or assign individual observations very large weights, which make any balanced sample highly variable. In this case, the conventional diagnostic checks and trimming methods used in IPW as outlined in Section 4.5.4 could be applied. See Table A4.1 for statistical software implementing entropy balancing.

Genetic Matching The GenMatch algorithm proposed by Diamond and Sekhon (2013) is an extension to Mahalanobis matching, whereby the Mahalanobis distance measure is adapted to minimize the post-matching imbalance of variables across treated and controls. In particular, the generalized Mahalonobis distance between the characteristics X of two subgroups is given by

$$d(X_i, X_j) = \left\{(X_i - X_j)(S^{-\frac{1}{2}})' W S^{-\frac{1}{2}}(X_i - X_j)\right\}^{\frac{1}{2}}, \tag{4.17}$$

with W representing a weight matrix that can be used to modify the optimal matches between individuals. Diamond and Sekhon (2013) propose an algorithm that uses the balance achieved in the matching process to modify the elements of W. The post-matching covariate balance is measured by the paired t-test and the Kolmogorov-Smirnov test to capture different aspects of imbalance. The p-value of these tests is used as harmonized indicator of balance. In particular, the algorithm aims to minimize the maximal difference in covariates across treatment groups, i.e. at each step the highest minimal p-value is used as loss. The algorithm finally uses the weight matrix that achieves the highest degree of balance. While clearly very attractive in theory, the method crucially depends on a good performance of the two balance tests to correctly detect imbalance across treatment groups. Against the background of the discussion in Section 4.7 further research is needed to corroborate the reliability of these indicators.

4.9.2 Multi-valued treatments

While we have so far exclusively focussed on the binary treatment case, the idea of balancing can be extended the multi-valued treatment or even the continuous treat-

ment case. Imbens (2000), Lechner (2001) and Hirano and Imbens (2005) provide a theoretical motivation for the estimation of treatment effects for the multi-valued treatment case. Assume that the treatment D can take on $m = 1, \ldots, M$ values. The *generalized* propensity score (GPS) is defined the conditional probability of receiving a particular level of treatment m, $r_i(m, X) = Pr(D = m|X)$. . Imbens (2000) shows that the GPS has the same balancing property as the binary propensity score, so that within strata of the GPS, the assignment to treatment level m is independent of pre-treatment characteristics. He further shows that identification of treatment effects only relies on a "weak" form of conditional independence which requires only pair-wise conditional independence, rather than independence from the joint set of all treatment levels. Hence, $Y(m) \perp\!\!\!\perp D, |X, \forall m$, which can be shown to be equivalent to $Y(m) \perp\!\!\!\perp D, |r(m, X), \forall m$.

In case where the number of distinct treatments M is discrete and limited, the methods applied in the binary treatment case can be straightforwardly applied by dissecting the evaluation problem in $M - 2$ binary evaluation problems. By estimating M propensity scores for all individuals, one for each treatment value $\hat{r}(1, X) = Pr(D = 1|X), ..., \hat{r}(M, X) = Pr(D = M|X)$, the previously outlined matching or weighting estimators can be applied. Depending on the characteristics of the treatment, multinomial logit models, nested logit models or multinomial probit models may be used to estimate the choice probabilities for qualitatively unordered treatments. Alternatively, ordered response models can be estimated when the treatment levels represent different intensities of exposure. Following this, standard matching techniques can be used to estimate the treatment effect of treatment m' vs treatment m'', whereby matching is conducted on the one-dimensional vector of $\hat{r}_{m'|m''} = \hat{r}(m', X)/(\hat{r}(m', X) + \hat{r}(m'', X))$, or the two-dimensional vector of $(\hat{r}(m', X), \hat{r}(m'', X))$ (Lechner, 2001). Analogous to the binary treatment case, a region of common support needs to be established using the methods outlined in Section 4.6. Note, that the region of overlap is now defined over all treatment levels jointly, i.e., it is required that for an individual i in treatment m a similar individual exists in all other treatment states. Clearly, in practice, this may be reduce the sample size substantially if there are a relatively large number of treatments, that are very heterogenous in the composition of participants.

4.9.3 Dynamic treatment assignment

A further extension of the standard balancing problem is the case where the treatment assignment occurs dynamically rather than instantly and does not occur at fixed point in time. Prominent examples are the assignment to labor market programs in the course of the unemployment, or the prescription of a specific drug over the course of a disease. These settings cannot be dealt with in a standard fashion for two reasons. First, while all individuals are likely to face a similar conditional propensity to be treated when entering the baseline state (i.e., unemployment, sickness), a dynamic selection out of this state over time prevents some individuals from getting treated. In our examples, individuals with better labor market characteristics or health conditions are likely to leave the baseline state before receiving treatment. In consequence, if we were to define treated and controls based on ever observing them in treatment, we would get a negatively selected treatment group (Fredriksson and Johansson, 2003; Sianesi, 2004), which would result in a downward bias of the estimates. A second problem arises due fact that treatment can occur at any time rather than at a fixed point in time, which implies that any control who remains in the baseline state may be treated at a later point in time. Hence, the treatment effect estimator may suffer from attenuation bias, as some controls have also participated in treatment. Note, that in some applications it may be insightful to estimate this "treated" vs. "not-yet-treated" treatment effect (Sianesi, 2004), it needs to be made clear however, that this is a downward biased estimate of the conventional ATT.

A straightforward solution to the first issue is to stratify the estimation problem by elapsed duration in the baseline state and only compare treated and controls who have remained in the baseline state until the (potential) treatment start. More specifically, define the treatment indicator as a function of the elapsed duration in the baseline state $t = 1, \ldots, T$. At a given treatment entry $t = t_D$, we compare treated who entered treatment at t_D with those who did not enter at t_D but who have remained in the baseline state until at least $t_D - 1$. At each t_D balancing and outcome analyses are conducted separately. The overall treatment effect is obtained by weighted averaging of the time-specific effects $\sum_t N_{t_D}/N_1 \Delta^{ATT}(t_D)$, whereas the weights are given by the share of entries at t_D from the group of overall entries. Note, that if the treatment is continuous, t_D may need to be discretized to ensure that sufficient observations are available in each interval.

In case where the parameter of interest is the effect of treatment on the

timing of exiting the baseline scenario t_u, the principles of non-parametric duration analysis can be used to tackle the problem of attenuation bias (Fredriksson and Johansson, 2008; Crepon et al., 2009; Vikström, 2014). The basic idea is to censor the outcomes of control observations once they enter treatment. As the decision to participate in treatment not expected to be random, Vikström (2014) proposes a weights correction that adjusts the initial PSM or IPW weights to account for this non-random censoring in each period.

4.10 Conclusion

Semi-parametric matching and weighting on the propensity score provide intuitive and transparent methods for establishing balance in covariates between two population or treatment groups. Based on an estimate of the propensity score as the conditional probability to be in the treatment group, the members of the control group are reweighed as a function of the predicted propensity score values. In propensity score matching (PSM), the members of the control group are reweighed based on the similarity of propensity score values. In inverse probability weighting (IPW), the members of the control group are reweighed inversely proportional to the value of propensity score. While the theoretical justification of the two reweighing schemes differ, the practical implementation is very similar, in that it consists of a multi-step implementation procedure that has the ultimate objective of minimizing imbalance in the covariates across the two sub-populations.

Practical difficulties in implementing these methods arise due the sensitivity of the balancing success to extreme values of the predicted propensity score, areas of low of lacking overlap in propensity score distributions, and uncertainty about the effects of tuning parameter choice on the final balancing quality. The vast amount of papers addressing issues of practical implementation shows that there is substantial need for increasing the knowledge about practical benefits and pitfalls of using these methods. Against this background, this chapter provides a comprehensive overview of the current practical guidance available in this literature regarding the implementation of IPW and PSM for the purpose of achieving covariate balance. Similar to Caliendo and Kopeinig (2008), the balancing challenge is dissected into five consecutive steps, outlining their role in the balancing process and suggesting ways to conduct the implementation. In contrast to them, we emphasize the distinctiveness of the balancing questions and questions related

to using balancing for the identification of causal outcome differences. Hence, the first part of this chapter focusses exclusively on the estimation of balancing weights, irrespective of outcome analysis. In the second part, we outline how to use the balancing weights in the estimation of conditional outcome differences, and address challenges specific to the identification of causal effects. We further present multiple ways so combine the non-parametric balancing with parametric outcome analysis, as the combination of these methods may increases robustness of estimates. Finally, we provide a detailed listing of the currently available statistical software for the implementation of matching and weighting and related methods.

Our summary of the state-of-the-art balancing tools also points to several issues that remain to be resolved in order to increase the reliability of these methods. A particular important issue is the definition of meaningful balancing tests, as conventional tests for equality of means or equality of distributions often fail to provide reliable guidance on whether to accept or to reject a given set of balancing weights (Imai et al., 2008; Ho et al., 2007; Lee, 2013). While it is currently advised to maximize balance of multiple balancing tests as far as possible, future research could be aimed at identifying meaningful, data-driven cut-off values, similar to significance levels, that can be used as benchmark for a sufficiently level of balance. A further practical problem is the intricate process of manual and iterative balance maximization, which is complicated due to the multitude of tuning parameters that can be influenced, e.g., the specification of the propensity score, the selection of caliper or bandwidth size, the trimming level, etc. Several promising approaches aiming to automatize the calculation of balance maximizing weights have been made in the literature, independent of outcomes analysis, see, e.g., Sekhon (2011) and Hainmueller (2011), or aiming to identifying robust treatment effect estimates, see, e.g., Graham et al. (2012) and Imai and Ratkovic (2014). Future research could complement and extend these advances by testing their robustness under the outlined problematic data-settings. Interdisciplinary research might speed up the development of this literature, as different areas of research deal with very similar challenges.

Appendix

A4.1 Tables

Table A4.1: Statistical Software packages in `stata` and `R`.

Steps	Stata	R	Matlab
1. Propensity Score Estimation			
Covariate balancing propensity scores		*cpbs*[1]	
Boosted CART	*boost*[2]	*gbm*[3]	
Boosted CART with balance optimization		*twang*[4]	
2. Matching and Weighting using the Propensity Score			
Compound packages	*psmatch2*[5] *nnmatch*[6]	*MatchIt*[7]	
Radius Matching	*radiusmatch*[23]	*radiusmatching*[23]	*radiusmatch*[23]
Optimal matching		*optmatch*[8]	
Full Optimal matching		*optmatch*[8]	
IPW1	*teffects ipw*[21]	*ipw*[22]	
3. Common Support			
Min-Max	*psmatch2*[5]		
Optimal Support	*optselect*[9]		*optselect*[9]
Convex Hull		*WhatIf*[10]	-
4. Balancing tests			
Two-sample KS-test with $\psi(\cdot) = 1$	*ksmirnov*	*ks.test*	
CMS-test	-	*CvM2SL2Test*[11]	
Multi-dimensional balance	*cem*[12]	*cem*[13]	
Omnibus test		*RItools*[16]	
Further Issues			
Sensitivity Analysis			
Rosenbaum bounds	*rbounds*[17] *mhbounds*[18]		
Rosenbaum-Rubin	*sensatt*[19]		
Automated Balance			
Automated Entropy Balancing	*ebalance*[14]	*baltest.collect*[14]	
Genetic Matching		*Matching*[15]	
Further outcome analysis			
Bias-correction for NN-matching	*nnmatch*[6]		
Regression adjusted IPW	*teffects ipwra*[21]		
Multi-dimensional treatment			
Generalized Propensity Score	*DRF*[20]		

[1]Imai and Ratkovic (2014), [2]Schonlau (2005), [3]Ridgeway et al. (2012), [4]Ridgeway (2007), [5]Leuven and Sianesi (2003), [6]Abadie et al. (2004) [7]Ho et al. (2011), [8]Hansen (2007), [9]Crump et al. (2009), [10]King and Zeng (2006), [11]Xiao et al. (2007), [12]Blackwell et al. (2009), [13]Iacus et al. (2009) [14]Hainmueller and Xu (2011) [15] Sekhon (2011) [16]Hansen and Bowers (2008),[17]DiPrete and Gangl (2004), [18]Becker and Caliendo (2007) [19]Nannicini (2007) [20]Bia et al. (2013) [21] from Stata version 13.1. [22] van der Wal and Geskus (2011) [23] Huber et al. (2012).

Bibliography

Aakvik, A. (2001). Bounding a Matching Estimator: The Case of a Norwegian Training Program. *Oxford Bulletin of Economics and Statistics 63*(1), 115–43.

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American Statistical Association 97*(457), 284–292.

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies 72*(1), 1–19.

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens (2004). Implementing Matching Estimators for Average Treatment Effects in Stata. *Stata Journal 4*(3), 290–311.

Abadie, A. and G. W. Imbens (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica 74*(1), 235–267.

Abadie, A. and G. W. Imbens (2008). On the Failure of the Bootstrap for Matching Estimators. *Econometrica 76*(6), 1537–1557.

Abadie, A. and G. W. Imbens (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business and Economic Statistics. 29*(1), 1–11.

Abbring, J. H. and G. J. van den Berg (2003). The Nonparametric Identification of Treatment Effects in Duration Models. *Econometrica 71*(5), 1491–1517.

Abbring, J. H. and G. J. van den Berg (2004). Analyzing the Effect of Dynamically Assigned Treatments Using Duration Models, Binary Treatment Models, and Panel Data Models. *Empirical Economics 29*, 5–20.

Abbring, J. H., G. J. van den Berg, and J. C. van Ours (2005). The Effect of Unemployment Insurance Sanctions on the Transition Rate from Unemployment to Employment. *Economic Journal 115*(505), 602–630.

Addison, J. T. and P. Portugal (2002). Job Search Methods and Outcomes. *Oxford Economic Papers 54*(3), 505–533.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Anderson, T. W. and D. A. Darling (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics 23*(2), 193–212.

Andrisani, P. J. (1977). Inter-External Attitudes, Personal Initiative, and the Labor Market Experience of Black and White Men. *Journal of Human Resources 12*(3), 308–328.

Armstrong, C. S., A. D. Jagolinzer, and D. F. Larcker (2010). Chief Executive Officer Equity Incentives and Accounting Irregularities. *Journal of Accounting Research 48*(2), 225–271.

Arni, P., R. Lalive, and J. C. Van Ours (2013). How Effective are Unemployment Benefit Sanctions? Looking Beyond Unemployment Exit. *Journal of Applied Econometrics 28*(7), 1153–1178.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics 60*(1), 47–57.

Augurzky, B. and C. M. Schmidt (2001). The Propensity Score: A Means to An End. IZA Discussion Papers 271, Institute for the Study of Labor (IZA).

Austin, P. C. (2011). Optimal Caliper Widths for Propensity-Scorecore Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies. *Pharmaceutical Statististics 10*, 150–161.

Austin, P. C. and M. M. Mamdani (2006). A Comparison of Propensity Score Methods: A Case-Study Estimating the Effectiveness of Post-AMI Statin Use. *Statistics in Medicine 25*(12), 2084–2106.

Autorengruppe Bildungsberichterstattung (2006). *1. Bildungsbericht - Bildung in Deutschland. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration.* Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland und BMBF.

Baiocchi, M., D. S. Small, S. Lorch, and P. R. Rosenbaum (2010). Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants. *Journal of the American Statistical Association 105*(492), 1285–1296.

Basu, A., D. Polsky, and W. G. Manning (2008). Use of Propensity Scores in Non-linear Response Models: The Case for Health Care Expenditures. NBER Working Paper 14086, National Bureau of Economic Research.

Bayer, P., S. L. Ross, and G. Topa (2008). Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes. *Journal of Political Economy 116*(6), 1150–1196.

Becker, S. O. and M. Caliendo (2007). Sensitivity Analysis for Average Treatment Effects. *Stata Journal 7*(1), 71–83.

Behncke, S., M. Frölich, and M. Lechner (2008). Public Employment Services and Employers: How Important Are Networks with Firms? *Zeitschrift für Betriebswirtschaft 1*, 151–178.

Behncke, S., M. Frölich, and M. Lechner (2010a). A Caseworker Like Me - Does The Similarity Between The Unemployed and Their Caseworkers Increase Job Placements? *Economic Journal 120*(549), 1430–1459.

Behncke, S., M. Frölich, and M. Lechner (2010b). Unemployed and their caseworkers: should they be friends or foes? *Journal of the Royal Statistical Society Series A 173*(1), 67–92.

Bell, D. N. and D. G. Blanchflower (2010). Youth Unemployment: Déjà Vu? IZA Discussion Papers 4705, Institute for the Study of Labor (IZA).

Bergemann, A., M. Caliendo, G. J. van den Berg, and K. F. Zimmermann (2011). The Threat Effect of Participation in Active Labor Market Programs on Job Search Behavior of Migrants in Germany. *International Journal of Manpower 32*(7), 777–795.

Bergemann, A., B. Fitzenberger, and S. Speckesser (2009). Evaluating the Dynamic Employment Effects of Training Programs in East Germany using Conditional Difference-in-Differences. *Journal of Applied Econometrics 24*(5), 797–823.

Bergemann, A. and G. J. van den Berg (2008). Active Labor Market Policy Effects for Women in Europe – A Survey. *Annals of Economics and Statistics / Annales d'Économie et de Statistique* (91/92), 385–408.

Bhattacharya, J. and W. Vogt (2012). Do Instrumental Variables Belong in Propensity Scores? *International Journal of Statistics & Economics 9*(A12), 107–127.

Bia, M., C. A. Flores, A. Flores-Lagunes, and A. Mattei (2013). A Stata Package for the Application of Semiparametric Estimators of Dose-Response Functions. Working Paper Series 2013-07, CEPS/INSTEAD.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review 96*(4), 988–1012.

Black, D. A. and J. A. Smith (2004). How Robust is the Evidence on the Effects of College Quality? Evidence from Matching. *Journal of Econometrics 121*(1-2), 99–124.

Blackwell, M., S. Iacus, and G. King (2009). cem: Coarsened Exact Matching in Stata. *The Stata Journal 9*(4), 524–546.

Blau, D. M. and P. K. Robins (1990). Job Search Outcomes for the Employed and Unemployed. *Journal of Political Economy 98*(3), 637–655.

Blinder, A. S. (1973). Wage Discrimation: Reduced Form and Structural Estimates. *Journal of Human Resources 7:4*, 436–455.

Büning, H. (2001). Kolmogorov-Smirnov and Cramer-von-Mises Type Two-Sample Tests with Various Weight Functions. *Communications in Statistics - Simulation and Computation 30*(4), 847–865.

Boockmann, B., C. Osiander, M. Stops, and H. Verbeek (2013). Effekte von Vermittlerhandeln und Vermittlerstrategien im SGB II und SGB III (Pilotstudie): Abschlussbericht an das IAB durch das Institut für Angewandte Wirtschaftsforschung e. V. (IAW), Tübingen. IAB-Forschungsbericht 201307, Institute for Employment Research (IAB), Nürnberg.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. t. Weel (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources 43*(4), 972–1059.

Boyd, C. L., L. Epstein, and A. D. Martin (2010). Untangling the Causal Effects of Sex on Judging. *American Journal of Political Science 54*(2), 389–411.

Bramoullé, Y. and G. Saint-Paul (2010). Social Networks and Labor Market Transitions. *Labour Economics 17*(1), 188–195.

Brodaty, T., B. Crépon, and D. Fougère (2001). Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986-1988. In M. L. und Friedhelm Pfeiffer (Ed.), *Econometric Evaluation of Labour Market Policies*, Number 13 in ZEW Economic Studies. Physica-Verlag.

Brookhart, A., S. Schneeweiss, K. Rothman, R. Glynn, J. Avorn, and T. Stürmer (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology 163*(12), 1149–1156.

Browning, M. and E. Heinesen (2012). Effect of Job Loss Due to Plant Closure on Mortality and Hospitalization. *Journal of Health Economics 31*(4), 599–616.

Bundesagentur für Arbeit (2007). *Arbeitsmarkt 2007*. Bundesagentur für Arbeit.

Bundesministerium für Arbeit und Soziales/Bundesministerium für Bildung und Forschung (1999). Eckpunkte für ein Sofortprogramm zum Abbau der Jugendarbeitslosigkeit. Technical report, In: Chronik er Artbeitsmarktpolitik, IAB.

Bundesministerium für Bildung und Forschung (2009). *Datenreport zum Berufsbildungsbericht 2009 - Informationen und Analysen zur Entwicklung der beruflichen Bildung*. Bundesinstitut für Berufsbildung.

Burgard, S. A., J. E. Brand, and J. S. House (2007). Toward a Better Estimation of the Effect of Job Loss on Health. *Journal of Health and Social Behavior 48*(4), 369–384.

Burgess, S., C. Propper, H. Rees, and A. Shearer (2003). The Class of 1981: The Effects of Early Career Unemployment on Subsequent Unemployment Experiences. *Labour Economics 10*(3), 291–309.

Buscha, F., A. Maurel, L. Page, and S. Speckesser (2012). The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Differences Approach. *Oxford Bulletin of Economics and Statistics 74*(3), 380–396.

Busso, M., J. DiNardo, and J. McCrary (2014a). Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. *Journal of Business & Economic Statistics*. forthcoming.

Busso, M., J. DiNardo, and J. McCrary (2014b). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *The Review of Economics and Statistics*. forthcoming.

Caliendo, M., D. Cobb-Clark, and A. Uhlendorff (2014). Locus of Control and Job Search Strategies. *Review of Econmics and Statistics*. forthcoming.

Caliendo, M., A. Falk, L. C. Kaiser, H. Schneider, A. Uhlendorff, G. van den Berg, and K. F. Zimmermann (2011). The IZA Evaluation Dataset: Towards Evidence-based Labor Policy Making. *International Journal of Manpower 32*(7), 731–752.

Caliendo, M. and R. Hujer (2006). The Microeconometric Estimation of Treatment Effects - An Overview. *Allgemeines Statistisches Archiv 90*(1), 197–212.

Caliendo, M., R. Hujer, and S. L. Thomsen (2008). The Employment Effects of Job Creation Schemes in Germany: A Microeconometric Evaluation. In D. Millimet, J. Smith, and E. Vytlacil (Eds.), *Advances in Econometrics*, Volume 21. Emerald Group Publishing Limited.

Caliendo, M., S. Künn, and R. Schmidl (2011). Fighting Youth Unemployment: The Effects of Active Labor Market Policies. IZA Discussion Papers 6222, Institute for the Study of Labor (IZA).

Caliendo, M. and S. Kopeinig (2008). Some Practical Guidance for fhe Implementation of Propensity Score Matching. *Journal of Economic Surveys 22*(1), 31–72.

Caliendo, M. and R. Schmidl (2011). Youth Unemployment and ALMP in Europe. mimeo.

Caliendo, M., R. Schmidl, and A. Uhlendorff (2011). Social Networks, Job Search Methods and Reservation Wages: Evidence for Germany. *International Journal of Manpower 32*(7), 796–824.

Calmfors, L. (1994). Active Labour Market Policy and Unemployment - a Framework for the Analysis of Crucial Design Features. *OECD Economic Studies 22*.

Calvo-Armengol, A. and M. O. Jackson (2007). Networks in Labor Markets: Wage and Employment Dynamics and Inequality. *Journal of Economic Theory 132*(1), 27–46.

Cappellari, L. and K. Tatsiramos (2010). Friends' Networks and Job Finding Rates. IZA Discussion Papers 5240, Institute for the Study of Labor (IZA).

Card, D., J. Kluve, and A. Weber (2010). Active Labour Market Policy Evaluations: A Meta-Analysis*. *The Economic Journal 120*(548), F452–F477.

Centeno, L., M. Centeno, and l. A. Novo (2009). Evaluating Jobsearch Programs for Old and Young Individuals: Heterogeneous Impact on Unemployment Duration. *Labour Economics 16*(1), 12–25.

Cepeda, M. S., R. Boston, J. T. Farrar, and B. L. Strom (2003). Comparison of Logistic Regression Versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. *American Journal of Epidemiology 158*(3), 280–287.

Chabé-Ferret, S. (2012). Matching vs Differencing When Estimating Treatment Effects with Panel Data: the Example of the Effect of Job Training Programs on Earnings. TSE Working Paper 12-356, Toulouse School of Economics.

Chabé-Ferret, S. (2014, March). Symmetric Difference in Difference Dominates Matching in a Realistic Selection Model. mimeo.

Chernozhukov, V. and I. Fernandez-Val (2005). Subsampling Inference on Quantile Regression Processes. *Sankhya : The Indian Journal of Statistics 67, Part 2*, 253–276. Special Issue on Quantile Regression and Related Methods.

Christensen, B. (2003). Mismatch-Unemployment Among the Low-Skilled. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 34*(4), 506–514.

Clarke, K. A., B. Kenkel, and M. R. Rueda (2011). Misspecificationcation and the Propensity Score: The Possibility of Overadjustment. mimeo.

Cochran, W. G. (1957). Analysis of Covariance: Its Nature and Uses. *Biometrics 13*(3), pp. 261–281.

Cockx, B. and M. Dejemeppe (2012). Monitoring Job Search Effort: An Evaluation Based on a Regression Discontinuity Design. *Labour Economics 19*(5), 729 – 737.

Cockx, B., M. Dejemeppe, A. Launov, and B. Van der Linden (2011). Monitoring, Sanctions and Front-Loading of Job Search in a Non-Stationary Model. IZA Discussion Papers 6181, Institute for the Study of Labor (IZA).

Costa-Dias, M., H. Ichimura, and G. J. van den Berg (2013). Treatment Evaluation with Selective Participation and Ineligibles. *Journal of the American Statistical Association* (tbc).

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment. *The Quarterly Journal of Economics 128*(2), 531–580.

Crepon, B., G. Jolivet, M. Ferracci, and G. van den Berg (2009). Active Labor Market Policy Effects in a Dynamic Setting. *Journal of the European Economic Association 7*(2-3), 595–605.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika 96*(1), 187–199.

Darity, William, J. and A. H. Goldsmith (1996). Social Psychology, Unemployment and Macroeconomics. *The Journal of Economic Perspectives 10*(1), pp. 121–140.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*, Volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

de Luna, X., P. Johansson, and S. Sjöstedt-de Luna (2010). Bootstrap Inference for k-Nearest Neighbour Matching Estimators. IFAU Working Paper Series 2010: 13, IFAU - Institute for Labour Market Policy Evaluation, Uppsala.

Dehejia, R. H. and S. Wahba (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Dehejia, R. H. and S. Wahba (2002). Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics 84*(1), 151–161.

Diamond, A. and J. Sekhon (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics 95*(3), 932–945.

DiCiccio, T. J. and B. Efron (1996). Bootstrap Confidence Intervals. *Statistical Science 11*(3), 189–228.

Dietrich, H. (2001). Wege aus der Jugendarbeitslosigkeit - Von der Arbeitslosigkeit in die Maßnahme? *Mitteilungen aus der Arbeitsmarkt und Berufsforschung 34*(4), 419–439.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica 64*(5), 1001–1044.

DiPrete, T. A. and M. Gangl (2004). Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments. *Sociological Methodology 34*, 271–310.

Dornette, J. and M. Jacob (2006). Zielgruppenerreichung und Teilnehmerstruktur des Jugendsofortprogramms. *IAB Forschungsbericht 16*, 3–48.

Dorsett, R. (2006). The New Deal for Young People: Effect on the Labor Market Status of Young Men. *Labor Economics 13*, 405–422.

Edin, P.-A. and M. Gustavsson (2008, January). Time out of Work and Skill Depreciation. *Industrial and Labor Relations Review 61*(2), 163–180.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics 7*(1), 1–26.

Efron, B. and R. Tibshirani (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science 1*(1), 54–75.

Eggers, A. C. and J. Hainmueller (2009). MPs for Sale? Returns to Office in Postwar British Politics. *American Political Science Review 103*(4), 513.

Ehlert, C., J. Kluve, and S. Schaffner (2012). Temporary Work as an Active Labor Market Policy – Evaluating an Innovative Activation Program for Disadvantaged Youths. *Economics Bulletin 32*(2), 1765–1773.

Eichhorst, W. and B. Ebbinghaus (2009). *The Labour Market Triangle Employment Protection, Unemployment Compensation and Activation in Europe*, Chapter Employment Regulation and Labor Market Policy in Germany, 1991-2005, pp. 119–144. Cheltenham: Edward Elgar.

Ellwood, D. T. (1983). Teenage Unemployment: Permanent Scars or Temporary Blemishes? NBER Working Papers 0399, National Bureau of Economic Research, Inc.

Engström, P., P. Hesselius, and B. Holmlund (2012). Vacancy Referrals, Job Search, and the Duration of Unemployment: A Randomized Experiment. *LABOUR 26,*(4), 419–435.

Fan, J. (1992). Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association 87*(420), pp. 998–1004.

Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics 21*(1), pp. 196–216.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1997). Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency. *Annals of the Institute of Statistical Mathematics 49*(1), 79–99.

Fan, J. and I. Gijbels (1995). Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction. *Journal Of Computational And Graphical Statistics 4*(3), 213.

Firpo, S. (2007). Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica 75*(1), 259–276.

Fitzenberger, B. and S. Speckesser (2007). Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany. *Empirical Economics 32*(2-3), 529–573.

Fortin, N., T. Lemieux, and S. Firpo (2011). *Decomposition Methods in Economics*, Volume 4 of *Handbook of Labor Economics*, Chapter 1, pp. 1–102. Elsevier.

Fougère, D., J. Pradel, and M. Roger (2009). Does the Public Employment Service Affect Search Effort and Outcomes? *European Economic Review 53*(7), 846–869.

Franzen, A. and D. Hangartner (2006). Social Networks and Labour Market Outcomes: The Non-Monetary Benefits of Social Capital. *European Sociological Review 22*(4), 353–368.

Fredriksson, P. and P. Johansson (2003). Program Evaluation and Random Program Starts. IFAU Working Paper Series 2003:1, IFAU - Institute for Labour Market Policy Evaluation.

Fredriksson, P. and P. Johansson (2008). Dynamic Treatment Assignment: The Consequences for Evaluation Using Observational Data. *Journal of Business and Economic Statistics 26*(4), 435–445.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics 29*(5), 1189–1232.

Frölich, M. (2004). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics 86*(1), 77–90.

Frölich, M. (2006). Nonparametric Regression for Binary Dependent Variables. *The Econometrics Journal 9*(3), 511–540.

Frölich, M. (2007a). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics 139*(1), 35–75.

Frölich, M. (2007b). Propensity Score Matching Without Conditional Independence Assumption– With an Application to the Gender Wage Gap in the United Kingdom. *Econometrics Journal 10*(2), 359–407.

Galdo, J. C., J. J. Smith, and D. Black (2008). Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data. *Annals of Economics and Statistics / Annales d'Économie et de Statistique 91/92*, 189–216.

Galeotti, A. and L. Merlino (2014). Endogenous Job Contact Networks. *International Economic Review*. forthcoming.

Gastwirth, J. L., A. M. Krieger, and P. R. Rosenbaum (1998). Dual and Simultaneous Sensitivity Analysis for Matched Pairs. *Biometrika 85*(4), 907–920.

Gautier, P., P. Muller, B. van der Klaauw, M. Rosholm, and M. Svarer (2012, July). Estimating Equilibrium Effects of Job Search Assistance. IZA Discussion Papers 6748, Institute for the Study of Labor (IZA).

Gaynor, J. J., E. J. Feuer, C. C. Tan, D. H. Wu, C. R. Little, D. J. Straus, B. D. Clarkson, and M. F. Brennan (1993). On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data. *Journal of the American Statistical Association 88*(422), pp. 400–409.

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist 48*(1), 26.

Goldsmith, A., J. Veum, and W. Darity (1997). Unemployment, Joblessness, Psychological Wellbeing and Self-Esteem: Theory and Evidence. *The Journal of Socio-Economics 26*, 133–158(26).

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics.

Graham, B. S., C. Campos de Xavier Pinto, and D. Egel (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *The Review of Economic Studies 79*(3), 1053–1079.

Granovetter, M. S. (1995). *Getting a Job: A Study of Contacts and Careers*. University of Chicago Press, Chicago.

Gregg, P. and E. Tominey (2005). The Wage Scar from Male Youth Unemployment. *Labour Economics 12*(4), 487–509.

Gregg, P. and J. Wadsworth (1996). How Effective are State Employment Agencies? Jobcentre Use and Job Matching in Britain. *Oxford Bulletin of Economics and Statistics 58*(3), 443–467.

Gregory, M. and R. Jukes (2001). Unemployment and Subsequent Earnings: Estimating Scarring Among British Men 1984–94. *The Economic Journal 111*, 607– 625.

Hainmueller, J. (2011). Entropy Balancing for Causal Effects: A multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 1–22.

Hainmueller, J., B. Hofmann, G. Krug, and K. Wolf (2011). Do Lower Caseloads Improve the Effectiveness of Active Labor Market Policies? New Evidence from German Employment Offices. Research Paper 2011-22, MIT Political Science Department.

Hainmueller, J. and Y. Xu (2011). Ebalance: A Stata Package for Entropy Balancing. Research paper, MIT Political Science Department.

Hall, P. and M. A. Martin (1989). A Note on the Accuracy of Bbootstrap Percentile Method Confidence Intervals for a Quantile. *Statistics & Probability Letters 8*(3), 197–200.

Ham, J. C., X. Li, and P. B. Reagan (2011). Matching and Semi-Parametric IV Estimation, a Distance-based Measure of Migration and the Wages of Young Men. *Journal of Econometrics 161*, 208–227.

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association 99*, 609–618.

Hansen, B. B. (2007). Optmatch: Flexible, Optimal Matching for Observational Studies. *New Functions for Multivariate Analysis 7*(2), 18–24.

Hansen, B. B. and J. Bowers (2008). Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science 23*(2), 219–236.

Haviland, A., S. Nagin, Daniel, and P. Rosenbaum (2007). Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study. *Psychological Methods 12*(3), 247–267.

Heckman, J., C. Heinrich, and J. Smith (1997). Assessing the Performance of Performance Standards in Public Bureaucracies. *The American Economic Review*, 389–395.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–61.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5), 1017–1098.

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

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies 64*(4), 605–54.

Heckman, J. J. and S. Navarro-Lozano (2004). Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *The Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J., J. A. Smith, and C. Taber (1996). What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA Program. Technical report, National Bureau of Economic Research.

Hielscher, V. and P. Ochs (2009). *Arbeitslose als Kunden? Beratungsgespräche in der Arbeitsvermittlung zwischen Druck und Dialog*. Berlin: edition sigma.

Hirano, K. and G. W. Imbens (2001). Estimation of Causal Effects using Propensity Score Weighting: An application to Data on Right Heart Catherization. *Health Services & Outcomes Research Methodology 2*, 259–278.

Hirano, K. and G. W. Imbens (2005). *The Propensity Score with Continuous Treatments*, pp. 73–84. John Wiley & Sons, Ltd.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica 71*(4), 1161–1189.

Ho, D., K. Imai, G. King, and E. A. Stuart (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis 15*(3), 199–236.

Ho, D., K. Imai, G. King, and E. A. Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software 42*(i08).

Ho, Y. H. and S. Lee (2005). Iterated Smoothed Bootstrap Confidence Intervals for Population Quantiles. *The Annals of Statistics 33*(1), 437–462.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association 81*(396), 945–960.

Holzer, H. J. (1988). Search Method Use by Unemployed Youth. *Journal of Labor Economics 6*(1), 1–20.

Horowitz, J. L. (2001). The Bootstrap. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5 of *Handbook of Econometrics*, pp. 3159 – 3228. Elsevier.

Horvitz, D. G. and D. J. Thompson (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association 47*(260), pp. 663–685.

Huber, M. (2011). Testing for Covariate Balance Using Quantile Regression and Resampling Methods. *Journal of Applied Statistics 38*(12), 2881–2899.

Huber, M., M. Lechner, and A. Steinmayr (2012). Radius Matching on the Propensity Score With Bias Adjustment: Finite Sample Behaviour, Tuning Parameters and Software Implementation. Economics Working Paper Series 1226, University of St. Gallen, School of Economics and Political Science.

Huber, M., M. Lechner, and C. Wunsch (2010). How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score. Discussion Paper 5268, IZA.

Huber, M., M. Lechner, and C. Wunsch (2013). The Performance of Estimators Based on the Propensity Score. *Journal of Econometrics 175*(1), 1–21.

IAB (2008). Entwicklung des gesamtwirtschaftlichen Stellenangebots vom IV. Quartal 2005 bis zum III. Quartal 2008 in Deutschland. IAB-Erhebung des gesamtwirtschaftlichen Stellenangebots 2005-2008, Bundesagentur für Arbeit.

Iacus, S., G. King, and G. Porro (2009). cem: Software for Coarsened Exact Matching. *Journal of Statistical Software 30*(9), 1–27.

Iacus, S. M., G. King, and G. Porro (2011). Multivariate Matching Methods That are Monotonic Imbalance Bounding. *Journal of the American Statistical Association 106*, 345–361.

Ichino, A., F. Mealli, and T. Nannicini (2008). From Temporary Help Jobs to Permanent Employment: What Can we Learn from Matching Estimators and their Sensitivity? *Journal of Applied Econometrics 23*(3), 305–327.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings Between Experimentalists and Observationalists about Causal Inference. *Journal Of The Royal Statistical Society Series A 171*(2), 481–502.

Imai, K. and M. Ratkovic (2014). Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Imbens, G. W. (2000). The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika 87*(3), 706–710.

Imbens, G. W. (2003). Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review 93*(2), 126–132.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. and J. Wooldridge (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature 47*(1), 5–86.

Ioannides, Y. M. and L. D. Loury (2004). Job Information Networks, Neighborhood Effects, and Inequality. *Journal of Economic Literature 42*(4), 1056–1093.

Jacobi, L. and J. Kluve (2007). Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany. *Zeitschrift für ArbeitsmarktForschung 40, H. 1*, 45–64.

Kahn, L. M. and S. A. Low (1988). Systematic and Random Search: A Synthesis. *Journal of Human Resources 23*(1), 1–20.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying Double Robstness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science 4*(4), 523–539.

Kaplan, E. L. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 53*(282), 457–481.

Khan, S. and E. Tamer (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica 78*(6), 2021–2042.

King, G. and L. Zeng (2006). The Dangers of Extreme Counterfactuals. *Political Analysis 14*(2), 131–159.

Kluve, J. (2010, December). The Effectiveness of European Active Labor Market Programs. *Labour Economics 17*(6), 904–918.

Kluve, J. and B. Augurzky (2007). Assessing the performance of matching algorithms when selection into treatment is strong. *Journal of Applied Econometrics 22*(3), 533–557.

Koenker, R. and Z. Xiao (2002). Inference on the Quantile Regression Process. *Econometrica 70*(4), 1583–1612.

Koenker, R. and J. Yoon (2009). Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy. *Journal of Econometrics 152*(2), 120–130.

Koning, P., G. J. van den Berg, and G. Ridder (1997). A Structural Analysis of Job Search Methods and Subsequent Wages. Tinbergen Institute Discussion Papers 97-082/3, Tinbergen Institute.

Koppel, O. (2008). Ingenieurarbeitsmarkt in Deutschland - Gesamtwirtschaftliches Stellenangebot und regionale Fachkräftelücken. IW-Trends 35, Institut der Deutschen Wirtschaft Köln.

Korpi, T. (1997). Is Utility related to Employment Status? Employment, Unemployment, Labor Market Policies and Subjective Well-Being among Swedish Youth. *Labour Economics 4*(2), 125 – 147.

Kroft, K., F. Lange, and M. J. Notowidigdo (2013). Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment*. *The Quarterly Journal of Economics 128*(3), 1123–1167.

Kultusministerkonferenz Germany and the EURIDYCE Unit (2009). The Education System in the Federal Republic of Germany 2007. Technical report, Bundesministerium für Forschung und Bildung.

Lagerström, J. (2011). How Important Are Caseworkers - and Why? New Evidence from Swedish Employment Offices. IFAU Working Paper Series 2011: 10, IFAU - Institute for Labor Market Policy Evaluation.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review 76*(4), 604–20.

Larrson, L. (2003). Evaluation of Swedish Youth Labor Market Programs. *The Journal of Human Resources 38, 4*, 891–927.

Layard, R. and S. Nickell (1986). Unemployment in Britain. *Economica 53*(210), pp. S121–S169.

Layard, R., S. Nickell, and R. Jackman (2005). *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford University Press.

Lechner, M. (2001). Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption. *Econometric Evaluation of Labour Market Policies*, 43–58.

Lechner, M. (2008). A Note on the Common Support Problem in Applied Evaluation Studies. *Annales d'Économie et de Statistique*, 217–235.

Lechner, M. and R. Miquel (2010). Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions. *Empirical Economics 39*(1), 111–137.

Lechner, M., R. Miquel, and C. Wunsch (2011). Long-run Effects of Public Sector Sponsored Training in West Germany. *Journal of the European Economic Association 9*(4), 742–784.

Lechner, M. and S. Wiehler (2011). Kids or Courses? Gender Differences in the Effects of Active Labor Market Policies. *Journal of Population Economics 24*, 783–812.

Lechner, M. and C. Wunsch (2013). Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables. *Labour Economics 21*, 111–121.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving Propensity Score Weighting using Machine Learning. *Statistics in Medicine 29*(3), 337–346.

Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight Trimming and Propensity Score Weighting. *PLoS ONE 6*(3), 6.

Lee, M. J. and S. J. Lee (2009). Sensitivity Analysis of Job-Training Effects on Reemployment for Korean Women. *Empirical Economics 36*(1), 81–107.

Lee, W.-S. (2013). Propensity Score Matching and Variations on the Balancing Test. *Empirical Economics 44*, 47–80.

Lehrer, S. and G. Kordas (2004). Matching Using Semiparametric Propensity Scores. Econometric Society 2004 North American Summer Meetings 441, Econometric Society.

Leuven, E. and B. Sianesi (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.

Liu, C. (2005). *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*, pp. 227–238. John Wiley & Sons, Ltd.

Lunceford, J. K. and M. Davidian (2004). Stratification and Weighting Via the Propensity Score in the Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine 23*, 2937–2960.

Martin, J. P. and D. Grubb (2001). What Works and for Whom: A Review of OECD Countries' Experiences with Active Labour Market Policies. *Swedish Economic Policy Review 8*(2), 9–56.

Mattei, A. (2009). Estimating and Using Propensity Score in Presence of Missing Background Data: An Application to Assess the Impact of Childbearing on Wellbeing. *Statistical Methods and Applications 18*, 257–273.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods 9*(4), 403–425.

Meyer, B. D. (1990). Unemployment Insurance and Unemployment Spells. *Econometrica*, 757–782.

Miguel, E. and M. Kremer (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica 72*(1), 159–217.

Millimet, D. L. and R. Tchernis (2009). On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies. *Journal of Business & Economic Statistics 27*(3), 397–415.

Millimet, D. L. and R. Tchernis (2012). Estimation of Treatment Effects Without an Exclusion Restriction: With an Application to the Analysis of the School Breakfast Program. *Journal of Applied Econometrics 28*(6), 982–1017.

Ming, K. and P. R. Rosenbaum (2000). Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls. *Biometrics 56*(1), 118–124.

Müller, A., M. Rebien, and M. Stops (2011). Einschaltungspotenzial für den Arbeitgeber-Service der Bundesagentur für Arbeit. Ergebnisse aus der IAB-Erhebung des Gesamtwirtschaftlichen Stellenangebots. Stellungnahme 10, IAB.

Müller, P. and B. Kurtz (2003). Active Labour Market Policy and Gender Mainstreaming in Germany. Gender-specific Aspects of Participation and Destination in Selected Instruments of the Federal Employment Service. *IAB Labour Market Research Topics 50*, 1–28.

Montgomery, J. D. (1991). Social Networks and Labor-Market Outcomes: Toward an Economic Analysis. *The American Economic Review 81*(5), 1408–1418.

Mortensen, D. T. (1986). Job Search and Labor Market Analysis. In O. Ashenfelter and R. Layard (Eds.), *Handbook of Labor Economics*, Volume 2 of *Handbook of Labor Economics*, Chapter 15, pp. 849–919. Elsevier.

Mortensen, D. T. and T. Vishwanath (1995). Personal Contacts and Earnings: It Is Who You Know! *Labour Economics 2*(1), 103–104.

Mouw, T. (2003). Social Capital and Finding a Job: Do Contacts Matter. *American Sociological Review 68*(6), 868–898.

Myers, J. A., J. A. Rassen, J. J. Gagne, K. F. Huybrechts, S. Schneeweiss, K. J. Rothman, M. M. Joffe, and R. J. Glynn (2011). Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates. *American Journal of Epidemiology 174*(11), 1213–1222.

Nannicini, T. (2007). A Simulation-Based Sensitivity Analysis for Matching Estimators. *The Stata Journal 7*(3), 334–350.

Neumann, M., J. Schmidt, and D. Werner (2010). *Die Integration Jungendlicher in Ausbildung und Beschäftigung*. Institut der deutschen Wirtschaft Köln.

Nicoletti, C. and C. Rondinelli (2010). The (Mis-)Specification of Discrete Duration Models with Unobserved Heterogeneity: A Monte Carlo Study. *Journal of Econometrics 159*(1), 1–13.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review 14*(3), 693–709.

OECD (1993). *Employment Outlook*. Paris: OECD Publishing.

OECD (1994). *OECD Job Study - Pushing Ahead With the Strategy*. OECD Publishing.

OECD (2001). *Labour Market Policies and the Public Employment Service*. Paris: OECD Publishing.

OECD (2004). *Employment Outlook*. Paris: OECD Publishing.

OECD (2007). *Employment Outlook*. Paris: OECD Publishing.

OECD (2011). *Employment Outlook*. Paris: OECD Publishing.

OECD (2013). *Employment Outlook*. Paris: OECD Publishing.

Ñopo, H. (2008). Matching as a Tool to Decompose Wage Gaps. *The Review of Economics and Statistics 90*(2), 290–299.

Osberg, L. (1993). Fishing in Different Pools: Job Search Strategies and Job-Finding Success in Canada in the Early 1980's. *Journal of Labor Economics 11*(2), 348–386.

Osborne Groves, M. (2005). How Important Is Your Personality? Labor Market Returns to Personality for Women in the US and UK. *Journal of Economic Psychology 26*(6), 827–841.

Patrick, A. R., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, A. J. Rothman, K. J., and T. Stürmer (2011). The Implications of Propensity Score Variable Selection Strategies in Pharmacoepidemiology: An Empirical Illustration. *Pharmacoepidemiology and Drug Safety 20*, 551–559.

Pavoni, N., O. Setty, and G. L. Violante (2013). Search and Work in Optimal Welfare Programs. Working Paper 18666, National Bureau of Economic Research.

Pavoni, N. and G. L. Violante (2007). Optimal Welfare-to-Work Programs. *Review of Economic Studies 74*(1), 283–318.

Pellizzari, M. (2010). Do Friends and Relatives Really Help in Getting a Good Job? *Industrial and Labor Relations Review 63*(3), 494–510.

Petrongolo, B. (2009). The long-term effects of job search requirements: Evidence from the UK JSA reform. *Journal of Public Economics 93*(11), 1234–1253.

Politis, D. N. and J. P. Romano (1992). A General Theory for Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions. Technical Report 399, Department of Statistics, Stanford University.

Politis, D. N. and J. P. Romano (1994). Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions. *Annals of Statistics 22*, 2031–2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer series in statistics. New York: Springer.

Pregibon, D. (1982). Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics 38*(2), pp. 485–498.

Quintini, G., J. P. Martin, and S. Martin (2007). The Changing Nature of the School-to-Work Transition Process in OECD Countries. IZA Discussion Papers 2582, Institute for the Study of Labor (IZA).

Raphael, S. and R. Winter-Ebmer (2001). Identifying the Effect of Unemployment on Crime. *Journal of Law and Economics 44*(1), 259–283.

Rees, A. (1966). Information Networks in Labor Markets. *The American Economic Review 56*(1/2), 559–566.

Reinberg, A. and M. Hummel (2005). Höhere Bildung schützt auch in der Krise vor Arbeitslosigkeit. *IAB Kurzbericht 9*, 1–6.

Ridgeway, G. (1999). The State of Boosting. *Computing Science and Statistics 31*, 172–181.

Ridgeway, G. (2007). Generalized Boosted Models: A Guide to the `gbm` package. Available at http://cran.r-project.org/web/packages/gbm/index.html.

Ridgeway, G., D. Mccaffrey, A. Morral, L. Burgette, and B. A. Griffin (2012). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the `twang`-package. Available at http://cran.r-project.org/web/packages/twang/index.html.

Robins, J., A. Rotnitzky, and L. P. Zhao (1995). Analysis of Semiparametric Regresion Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association 90*(429), 106–121.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science 22*(4), pp. 544–559.

Robins, J. A. and A. Rotnitzky (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association 90*(429), 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association 89*(427), 846–866.

Rosenbaum, P. R. (1987). Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika 74*(1), 13–26.

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association 84*(408), 1024–1032.

Rosenbaum, P. R. (1991). A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society. Series B (Methodological) 53*(3), pp. 597–610.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R., R. N. Ross, and J. H. Silber (2007). Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer. *Journal of the American Statistical Association 102*, 75–83.

Rosenbaum, P. R. and D. B. Rubin (1983a). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological) 45*(2), pp. 212–218.

Rosenbaum, P. R. and D. B. Rubin (1983b). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association 79*(387), 516–524.

Rosenbaum, P. R. and D. B. Rubin (1985a). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician 39*(1), pp. 33–38.

Rosenbaum, P. R. and D. B. Rubin (1985b). The Bias Due to Incomplete Matching. *Biometrics 41*(1), pp. 103–116.

Rosholm, M. and M. Svarer (2008). The Threat Effect of Active Labour Market Programmes. *Scandinavian Journal of Economics 110*(2), 385–401.

Rotter, J. (1966). Generalized Expectancies for Internal Versus External Control of Reinforcement. *Psychological Monographs 80*.

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

Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics 29*(1), pp. 185–203.

Rubin, D. B. (1974). Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Policy 66*, 688–701.

Rubin, D. B. (1979). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association 74*, 318–328.

Rubin, D. B. (1980). Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics 36*(2), pp. 293–298.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology 2*(3), 169–188.

Rubin, D. B. (2004). On Principles for Modeling Propensity Scores in Medical Research. *Pharmacoepidemiology and Drug Safety 13*(12), 855–857.

Rubin, D. B. (2007). The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels With the Design of Randomized Trials. *Statistics in Medicine 26*(1), 20–36.

Rubin, D. B. and N. Thomas (1996). Matching Using the Estimated Propensity Scores: Relating Theory to Practice. *Biometrics 52*(1), 249–264.

Saposnik, G., M. Kapral, R. Cote, P. Rochon, J. Wang, S. Raptis, M. Mamdani, and S. Black (2012). Is Pre-Existing Dementia an Independent Predictor of Outcome After Stroke? A Propensity Score-Matched Analysis. *Journal of Neurology 259*, 2366–2375.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association 94*, 1096–1146.

Schonlau, M. (2005). Boosted Regression (Boosting): An Introductory Tutorial and a Stata Plugin. *The Stata Journal 5*(3), 330–354.

Schreyögg, J., T. Stargardt, and O. Tiemann (2011). Costs and Quality of Hospitals in Different Health Care Systems: A Multi-level Approach with Propensity Score Matching. *Health Economics 20*(1), 85–100.

Schütz, H., J. Steinwede, H. Schröder, B. Kaltenborn, N. Wielage, G. Christe, and P. Kupka (2011). *Vermittlung und Beratung in der Praxis - eine Analyse von Dienstleistungsprozessen am Arbeitsmarkt*. IAB-Bibliothek, Biefeld, Bertelsmann.

Seifert, B. and T. Gasser (1996). Finite-Sample Variance of Local Polynomials: Analysis and Solutions. *Journal of the American Statistical Association 91*(433), pp. 267–275.

Seifert, B. and T. Gasser (2000). Data Adaptive Ridging in Local Polynomial Regression. *Journal of Computational and Graphical Statistics 9*(2), 338–360.

Sekhon, J. S. (2007). Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference. Technical report, UC Berkeley.

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software 42*(7), 1–52.

Setoguchi, S., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety 17*(6), 546–555.

Shaikh, A. M., M. Simonsen, E. J. Vytlacil, and N. Yildiz (2009). A Specification Test for the Propensity Score Using its Distribution Conditional on Participation. *Journal of Econometrics 151*(1), 33–46.

Sianesi, B. (2002). Swedish Active Labour Market Programmes in the 1990s: Overall Effectiveness and Differential Performance. IFS Working Papers W02/03, Institute for Fiscal Studies.

Sianesi, B. (2004). An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s. *The Review of Economics and Statistics 86*(1), 133–155.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. CCr Pr Inc.

Smith, J. and P. Todd (2005a). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics 125*(1-2), 305–353.

Smith, J. and P. Todd (2005b). Rejoinder. *Journal of Econometrics 125*, 365–375.

Spinnewijn, J. (2013). Training and Search During Unemployment. *Journal of Public Economics 99*, 49–65.

Stephens, M. A. (1974). EDF Statistics for Goodness-of-Fit and Some Comparisons. *Journal of the American Statistical Association 69*, 730–737.

Stephens, M. J. (2001). The Long-Run Consumption Effects of Earnings Shocks. *Review of Economics and Statistics 83*(1), 28–36.

Stürmer, T., K. J. Rothman, J. Avorn, and R. J. Glynn (2010). Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution - A Simulation Study. *American Journal of Epidemiology 172*(7), 843–854.

Stuart, E. and K. Green (2008). Using Full Matching to Estimate Causal Effects in Non-Experimental Studies: Examining the Relationship between Adolescent Marijuana Use and Adult Outcomes. *Developmental Psychology 44*, 395–406.

Sullivan, J. X. (2008). Borrowing During Unemployment Unsecured Debt as a Safety Net. *Journal of human resources 43*(2), 383–412.

Topa, G. (2001). Social Interactions, Local Spillovers and Unemployment. *Review of Economic Studies 68*(2), 261–295.

van den Berg, G. J. (2001). Duration models: Specification, Identification and Multiple durations. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5 of *Handbook of Econometrics*, Chapter 55, pp. 3381–3460. Elsevier.

van den Berg, G. J., B. Hofmann, and A. Uhlendorff (2013). The Role of Sickness in the Evaluation of Job Search Assistance and Sanctions. mimeo.

van den Berg, G. J. and B. van der Klaauw (2006). Counseling And Monitoring Of Unemployed Workers: Theory And Evidence From A Controlled Social Experiment. *International Economic Review 47*(3), 895–936.

van der Wal, W. M. and R. B. Geskus (2011). ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software 43*(13), 1–23.

Verick, S. (2011). Who is Hit Hardest during a Financial Crisis? The Vulnerability of Young Men and Women to Unemployment in an Econonomic Downturn. In I. Islam and S. Verick (Eds.), *From the Great Recession to Labour Market Recovery: Issues, Evidence and Policy Options*. ILO/Palgrave Macmillan.

Vikström, J. (2014). IPW estimation and related estimators for evaluation of active labor market policies in a dynamic setting. Working paper 1, IFAU.

Vikström, J., P. Johansson, and I. Waernbaum (2012). Identification and Estimation of Causal Effects of Treatment Regimens with Duration Outcomes. mimeo.

Waernbaum, I. (2012). Model Misspecification and Robustness in Causal Inference: Comparing Matching with Doubly Robust Estimation. *Statistics in Medicine 31*(15), 1572–1581.

Wahba, J. and Y. Zenou (2005). Density, Social Networks and Job Search Methods: Theory and Application to Egypt. *Journal of Development Economics 78*(2), 443–473.

Weber, A. and H. Mahringer (2008). Choice and Success of Job Search Methods. *Empirical Economics 35*(1), 153–178.

Westreich, D., J. Lessler, and M. J. Funk (2010). Propensity Score Estimation: Neural Networks, Support Vector Matchines, Decision Trees (CART) and Meta-Classifiers as Alternatives to Logistic Regression. *Journal of Clinical Epidemiology 63*, 826–833.

Wooldridge, J. M. (2005). Violating Ignorability Of Treatment By Controlling For Too Many Factors. *Econometric Theory 21*(05), 1026–1028.

Wooldridge, J. M. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics 141*(2), 1281–1301.

Wooldridge, Jeffrey, M. (2002). Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal 1*, 117–139.

Wunsch, C. (2013). Optimal Use of Labor Market Policies: The Role of Job Search Assistance. *Review of Economics and Statistics 95*(3), 1030–1045.

WZB and infas (2006). Evaluation der Maßnahmen zur Umsetzung der Vorschläge der Hartz-Kommission Modul 1a Neuausrichtung der Vermittlungsprozesse. Technical report, Research report of the Federal Ministry for Labour and Social Affairs, BMAS Berlin.

Xiao, Y., A. Gordon, and A. Yakovlev (2007). A C++ program for the Cramer-von Mises Two Sample Test. *Journal of Statistical Software 17*(8).

Zhao, Z. (2004). Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. *The Review of Economics and Statistics 86*(1), 91–107.

Zhao, Z. (2008). Sensitivity of Propensity Score Methods to the Specifications. *Economics Letters 98*(3), 309–319.

German Summary

Die zielgerichtete und nachhaltige Integration von Arbeitslosen auf den ersten Arbeitsmarkt ist seit den Hartz-Reformen ein zentraler Bestandteil der Arbeitsmarktpolitik in Deutschland. Aktivierende Maßnahmen der Arbeitsmarktpolitik haben zum einen das Ziel, die Eigenbemühungen von Arbeitslosen zu unterstützten und zu stärken, beispielsweise durch den Ausbau von Kontroll- und Beratungsmechanismen in der Arbeitssuche. Zum anderen sollen durch innovative Förderprogramme die Chancen von Arbeitslosen mit strukturelleren Vermittlungshemmnissen erhöht werden. Seit der Einführung der Hartz-Gesetze wird die Arbeitsmarktpolitik zudem verstärkt durch die empirische Evaluationsforschung begleitet, um Art und Ausgestaltung von Arbeitsmarktmaßnahmen durch einen Kreislauf von Kontrolle, Feedback und Anpassung möglichst effektiv zu gestalten. Neben der empirischen Evaluation von aktiver Arbeitsmarktpolitik ist auch die Analyse von individuellen Determinanten des Suchverhaltens von Arbeitslosen ein zentraler Bestandteil der Arbeitsmarktforschung. Die vorliegende Dissertation hat das Ziel, neuartige Erkenntnisse für die optimale Ausgestaltung von Aktivierungsmaßnahmen generieren.

Die empirischen Studien der Kapiteln 1 bis 3 basieren auf dem *IZA Evaluationsdatensatz S*. Der administrative Teil des Datensatzes basiert auf den Integrierten Erwerbsbiographien (IEB) des Instituts für Arbeitsmarkt- und Berufsforschung (IAB), und besteht aus einer 900.000 Individuen umfassenden Zufallsstichprobe von monatlichen Eintritten in Arbeitslosigkeit zwischen 2001 und 2008, wobei die Erwerbsbiographien der Individuen im Zeitverlauf verfolgt werden. Dieser Datensatz wird in Kapitel 3 verwendet. Der Survey des *IZA Evaluationsdatensatz S* basiert auf einer rund 17.000 Personen umfassenden, repräsentativen Stichprobe der monatlichen Eintritte in Arbeitslosigkeit zwischen Juni 2007 und Mai 2008.

Der Datensatz ist als Panel konzipiert, wobei die eingangs arbeitslosen Individuen ein und drei Jahre nach ihrem Eintritt erneut befragt wurden. Basierend

auf der zweiten und dritten Befragung kann über die Zeit ein detaillierter Erwerbsverlauf erstellt werden (siehe Caliendo et al., 2011). Dieser Teil des Datensatzes wird in den Kapiteln 1 und 2 verwendet. In Kapitel 1 der Dissertation wird die Rolle von sozialen Netzwerken als wichtige Determinante im Suchverhalten von Arbeitslosen analysiert. Obwohl soziale Netzwerke zu den am häufigsten verwendeten und effektivsten Informationsquellen während der Arbeitssuche gehören (Pellizzari, 2010) und eine umfangreiche Literatur der Frage nachgeht wie sich diese Netzwerke auf Löhne und Beschäftigungsstabilität auswirken (Ioannides and Datcher Loury, 2004, Mouw, 2003), existiert vergleichsweise wenig direkte Evidenz über die Rolle dieser Netzwerke im eigentlichen Suchprozess. Ziel der Analyse ist es, das Zusammenspiel zwischen Netzwerken sowie formeller und informeller Jobsuche besser zu verstehen, und somit die erwarteten Effekte von sozialen Netzwerken auf den Arbeitsmarkterfolg besser einordnen zu können. Innerhalb eines theoretischen Modells der Arbeitssuche wird folgender Zusammenhang hergestellt. Basierend auf der Hypothese, dass Arbeitslose durch ihr soziales Netzwerk Informationen über Stellenangebote generieren, sollten Personen großen sozialen Netzwerken eine erhöhte Produktivität der informellen Suche erfahren, und ihre Suche in formellen Kanälen reduzieren. Durch die höhere Produktivität der Suche sollten Arbeitslose mit größerem Netzwerk zudem einen höheren Reservationslohn haben, als Arbeitslose mit kleinem Netzwerk.

Die modelltheoretischen Vorhersagen werden empirisch getestet, wobei die Netzwerkinformationen durch die Anzahl guter Freunde, sowie Kontakthäufigkeit zu früheren Kollegen approximiert wird. Diese werden dann als erklärende Variablen in linearen Regressionsmodellen in Bezug zur Suchintensität, zur Art der Suche, und zum Reservationslohn gesetzt. Die Ergebnisse zeigen, dass das Suchverhalten der Arbeitslosen in der Tat durch das Vorhandensein sozialen Kontakte signifikant beeinflusst wird. Insbesondere finden sich für größere Netzwerke Substitutionseffekte Personen zu informelle Suche zu Lasten formeller Suche. Die Substitution ist besonders stark für passive formelle Suchmethoden, d.h. Informationsquellen die eher unspezifische Arten von Jobangeboten bei niedrigen relativen Kosten erzeugen. Im Einklang mit den Vorhersagen des theoretischen Modells finden sich auch deutlich positive Auswirkungen einer Erhöhung der Netzwerkgröße auf den Reservationslohn.

Kapitel 2 befasst sich mit den Arbeitsmarkteffekten von Vermittlungsangeboten in der frühzeitigen Aktivierungsphase von Arbeitslosen. Obwohl individuali-

sierte Informationen über verfügbare Stellenangebote ein wichtiger Bestandteil der Aktivierungsstrategie in OECD-Ländern sind (OECD, 2007), wurde diese Komponente der frühen Aktivierung bisher nicht umfangreich untersucht. Die Nutzung von Vermittlungsangeboten könnte dabei eine "doppelte Dividende" versprechen. Zum einen reduziert die frühe Aktivierung die Dauer der Arbeitslosigkeit, und somit auch die Notwendigkeit späterer Maßnahmenteilnahme. Zum anderen ist die Aktivierung durch Arbeitsmarktinformation mit geringeren "locking-in" Effekten verbunden als alternative Programme der frühzeitigen Aktivierung. Ziel der Analyse ist es, die Effekte von frühen Vermittlungsangeboten auf die Eingliederungsgeschwindigkeit in Arbeit zu messen, und die kurz- und mittelfristigen Effekte auf Maßnahmenteilnahme der aktiven Arbeitsmarktpolitik zu analysieren. Zudem werden mögliche Effekte auf die Qualität der Beschäftigung untersucht.

Diese Ergebnisse zeigen, dass Vermittlungsangebote die Beschäftigungswahrscheinlichkeit signifikant erhöhen, und dass gleichzeitig die Wahrscheinlichkeit an aktiven Arbeitsmarktprogrammen teilzunehmen signifikant reduziert wird. Für die meisten betrachteten Subgruppen kann die langfristige Reduktion der Teilnahmewahrscheinlichkeit als Konsequenz der schnelleren Beschäftigungseintritts gesehen werden. Für Arbeitslose in Ostdeutschland zeigt sich jedoch bereits früh eine signifikante und temporäre Reduktion der Teilnahmewahrscheinlichkeit was darauf hinweist, dass Maßnahmen mit hohen und geringen "locking-in" Effekten aus Sicht der Sachbearbeiter austauschbar sind, was jedoch aus Effizienzgesichtspunkten fraglich ist. Es wird ein geringer negativer Effekt auf die Beschäftigungsqualität, in Form einer Reduktion der wöchentliche Stundenanzahl beobachtet.

In Kapitel 3 schließlich werden die Langzeiteffekte von Maßnahmen der aktiven Arbeitsmarktpolitik für arbeitslose Jugendlichen unter 25 Jahren ermittelt. Komplementär zu den Ergebnissen in Kapitel 2 werden hier die Effekte der Teilnahme in zeit- und kostenintensiveren Maßnahmen der aktiven Arbeitsmarktpolitik untersucht. Jugendarbeitslosigkeit wird besonders durch langfristige "scarring effects" als sehr problematisch in Bezug auf spätere Arbeitsmarktintegration gesehen (Ellwood, 1983, Burgess et al., 2003, Gregg and Tominey, 2005). Zum Zeitpunkt dieser Untersuchung sind jedoch noch keine umfassenden quantitativen Analysen der Wirksamkeit der aktiven Arbeitsmarktpolitik für Jugendliche in Deutschland durchgeführt worden, was unter anderem auf Einschränkungen in der Datenverfügbarkeit zurückzuführen ist. Die untersuchten ALMP Programme sind ABM-Maßnahmen, Lohnsubventionen, kurz-und langfristige Maßnahmen der be-

ruflichen Bildung sowie Maßnahmen zur Förderung der Teilnahme an Berufsausbildung. Ab Eintritt in die Maßnahme werden Teilnehmer und Nicht-Teilnehmer für einen Zeitraum von sechs Jahren beobachtet. Als Zielvariable wird die Wahrscheinlichkeit regulärer Beschäftigung, sowie die Teilnahme in Ausbildung untersucht.

Die Ergebnisse zeigen, dass alle Programme, bis auf ABM, positive und langfristige Effekte auf die Beschäftigungswahrscheinlichkeit von Jugendlichen haben. Kurzfristig finden wir jedoch nur für kurze Trainingsmaßnahmen positive Effekte, da lange Trainingsmaßnahmen und Lohnzuschüsse mit signifikanten "locking-in" Effekten verbunden sind. Maßnahmen zur Förderung der Berufsausbildung erhöhen zudem die Wahrscheinlichkeit der Teilnahme an eine Ausbildung, während alle anderen Programme keinen oder einen negativen Effekt auf die Ausbildungsteilnahme haben. Effektheterogenität nach Ausbildungsniveau zeigen, dass Jugendlichen mit höherem Ausbildungsniveau stärker von der Programmteilnahme profitieren. Jedoch zeigen sich für längerfristige Lohnsubventionen ebenfalls starke positive Effekte für Jugendliche mit geringer Vorbildung. Der relative Nutzen von Trainingsmaßnahmen ist höher in West- als in Ostdeutschland.

In den Evaluationsstudien der Kapitel 2 und 3 werden die semi-parametrischen Gewichtungsverfahren Propensity Score Matching (PSM) und Inverse Probability Weighting (IPW) verwendet um den Einfluss verzerrender Faktoren die sowohl die Maßnahmenteilnahme als auch die Zielvariablen beeinflussen zu beseitigen, und kausale Effekte der Programmteilahme zu ermitteln. Während PSM and IPW intuitiv und methodisch sehr attraktiv sind, stellt die Implementierung der Methoden in der Praxis oft eine große Herausforderung dar. Ein weiteres Ziel dieser Dissertation ist es daher, die Wissenslücke zwischen der methodischen und der angewandten Literatur hinsichtlich beider Schätzverfahren zu reduzieren und praktische Implementierungshinweise zu geben. Zu diesem Zweck werden in Kapitel 4 neue Erkenntnisse der empirischen und statistischen Literatur zusammengefasst und praxisbezogene Richtlinien für die angewandte Forschung abgeleitet.

Basis hierfür sind wissenschaftliche Veröffentlichungen der letzten Jahre, die mittels statistisch-theoretischen Analysen, methodischen Simulationen oder empirischen Studien neue Erkenntnisse hinsichtlich der praktischen Anwendung von PSM und IPW liefern. Das Kapitel beginnt mit einer theoretischen Motivation, die die statistische Balancierung der beobachtbaren Charakteristika im Rahmen der kontrafaktischen Ergebnisanalyse mit und ohne konditionale Unabhängigkeitsan-

nahme diskutiert, und einen Überblick über die praktischen Voraussetzungen der Angleichung mit Propensity Score Methoden gibt. Nach einer Skizzierung der praktischen Implementierungsschritte von PSM und IPW werden diese Schritte chronologisch dargestellt, wobei praxisrelevante Erkenntnisse aus der methodischen Forschung dargestellt werden. Im Anschluss werden die Themen Effektschätzung, Inferenz, Sensitivitätsanalyse und die Kombination von IPW und PSM mit parametrischen Analysemethoden diskutiert. Abschließend werden aktuelle Erweiterungen der Methodik dargestellt.

English Summary (Abstracts)

Abstract Chapter 1 In this paper we analyze the relationship between social networks and the job search behavior of unemployed individuals. It is believed that networks convey useful information in the job search process such that individuals with larger networks should experience a higher productivity of informal search. Hence, job search theory suggests that individuals with larger networks use informal search channels more often and substitute from formal to informal search. Due to the increase in search productivity, it is also likely that individuals set higher reservation wages. We analyze these relations using a novel data set of unemployed in Germany which contains extensive information on their job search behavior and direct measures for their social network. Furthermore, the data contain an unusually rich set of personality traits, which allows us to justify an identification approach based on observable characteristics. Our findings confirm theoretical expectations. Individuals with larger networks use informal search channels more often and shift from formal to informal search. We find that informal search is mainly considered a substitute for passive, less cost-intensive search channels. In addition to that, we find evidence for a positive relationship between the network size and reservation wages.

Abstract Chapter 2 In most countries, unemployment activation schemes are used progressively: following low intensity job broking services shortly after unemployment entry, more intensive active labor market programs (ALMP) are used later if unemployment persists. We study the effects of early vacancy information (VI) for unemployed in Germany, considering the effects on unemployment exit, participation in ALMP, and quality of accepted employment. Controlling for endogeneity arising from overall labor market conditions and caseworker heterogeneity, we show that VI significantly increase the unemployment exit rate, while decreasing ALMP participation. The latter occurs as a consequence of the former,

but we also find that ALMP may be used as substitute for lacking VI in the activation process. We hence show that the quality of early job-broking affects both individual outcomes, and cost-effectiveness of the overall activation process.

Abstract Chapter 3 A substantial number of young unemployed participate in active labor market programs (ALMP) in Germany each year. While the aims of these programs are clear—a fast re-integration into employment or enrollment in further education—a comprehensive analysis of their effectiveness has yet to be conducted. We fill this gap using administrative data on youth unemployment entries in 2002 and analyze the short- and long-term impacts for a variety of different programs. With informative data at hand we apply inverse probability weighting, thereby accounting for a dynamic treatment assignment and cyclical availability of programs. Our results indicate positive long-term employment effects for nearly all measures aimed at labor market integration. Measures aimed at integrating youths in apprenticeships are effective in terms of education participation, but fail to show any impact on employment outcomes until the end of our observation period. Public sector job creation is found to be harmful for the medium-term employment prospects and ineffective in the long-run. Our analysis further indicates that the targeting of German ALMP systematically ignores low-educated youths as neediest of labor market groups. While no employment program shows a positive impact on further education participation for any subgroup, the employment impact of participation is often significantly lower for low-educated youths.

Abstract Chaper 4 Matching and weighting on the propensity score are commonly used balancing methods that find ample application in the empirical evaluation of interventions, decomposition analysis and the design of surveys or field-experiments. Despite the heterogenous areas of application, the underlying balancing challenge is very similar, requiring the transformation of the estimated propensity scores into balancing weights that balance the distribution of characteristics across subgroups. The optimal estimation of weights requires a number of choices that need to be taken in view of the data setting at hand. The objective of paper is to summarize the state-of-the art knowledge on the implementation of propensity score matching (PSM) and inverse probability weighting (IPW) and give advice on the practical implementation of the balancing methods, outlining

their practical benefits and limits. The first part of the chapter focusses on the balancing challenge, i.e., the estimation of the propensity score, the choice of the balancing method and the measurement of balance. The second part of the chapter deals with the estimation of conditional outcome differences using the balancing weights, and provides practical guidelines on the combination of the weights with parametric outcome analysis in order to improve stability of the estimates.

Curriculum Vitae–Ricarda Schmidl

Aus datenschutzrechtlichen Gründen
wird in der publizierten Version
auf einen Lebenslauf verzichtet.

Aus datenschutzrechtlichen Gründen
wird in der publizierten Version
auf einen Lebenslauf verzichtet.