Zur Wirksamkeit der „Förderung beruflicher Weiterbildung“ (FbW)

Wissenschaftliche Arbeit
zur Erlangung des Grades eines Diplom-Volkswirtes
im Fachbereich Wirtschaftswissenschaften
der Universität Konstanz

Verfasser: Marcus Jansen
Steinstr. 17
78467 Konstanz

Bearbeitungszeit: 08.12.2004 bis 15.02.2005

1. Gutachter: Prof. Ursprung
2. Gutachter: Prof. Fabel

Konstanz, den 11.02.2005

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AAMP	aktive Arbeitsmarktpolitik
AMP	Arbeitsmarktpolitik
AFG	Arbeitsförderungsgesetz
AMO	Arbeitsmarktmonitor Ost
AMSA	Arbeitsmarktmonitor Sachsen-Anhalt
ATET	average treatment effect on the treated
BA	Bundesagentur für Arbeit
CIA	bedingte Unabhängigkeitsannahme
DvD	Differenz-von-Differenz
FbW	Förderung beruflicher Weiterbildung
FuU	Fortbildung und Umschulung
GSOEP	deutsches sozio-ökonomisches-Panel
IAB	Institut für Arbeitsmarkt- und Berufsforschung
IABS	IAB-Beschäftigungsstichprobe
IV	Instrumentvariable
LATE	local avarage treatment effect
LED	Leistungsempfängerdatei
LMI	Leistungsmotivationsinventar
LMT	Leistungsmotivationstest
MPA	Motivationspotential der Arbeitssituation
OLS	ordinary least square
PH	proportionale Hazardraten
SGB	Sozialgesetzbuch
SUTVA	stable unit treatment value assumption
TAT	thematischen Apperzeptionstest
TDP	training participant data

I. Einleitung

Die Frage nach der Wirksamkeit und Effizienz von Maßnahmen der Förderung beruflicher Weiterbildung (FbW) der Bundesagentur für Arbeit gewinnt ihre Aktualität aus der seit Jahren prekären Lage am Arbeitsmarkt und den hohen fiskalischen Kosten der FbW bei gleichzeitig angespannter Lage der öffentlichen Haushalte.
Eine vollständige Evaluation erfordert die Messung der individuellen Effekte der Teilnahme an einer Weiterbildung und die Kenntnis der Wirkung auf aggregierte Größen wie Beschäftigung und Arbeitslosigkeit. Sollten diese Untersuchungen auf positive Effekte der FbW hinweisen, ist für eine abschließende Beurteilung mittels einer Kosten-Nutzen Analyse zu überprüfen, ob die Effekte die eingesetzten Mittel rechtfertigen.

Die Evaluationsanstrengungen in Deutschland haben in den letzten 10 Jahren deutlich zugenommen und konzentrieren sich insbesondere auf die mikroökonomische Bewertung der FbW. Die vorliegende Arbeit gibt einen Überblick über den Stand und die Perspektiven der aktuellen mikroökonomischen Evaluationsforschung.
Die Ermittlung individueller kausaler Effekte der Teilnahme an einer Maßnahme ist mit großen Schwierigkeiten behaftet, da nicht beobachtet werden kann, wie die Arbeitsmarktsituation der Teilnehmer ohne eine Fortbildung gewesen wäre. Die jüngsten Studien ermitteln die kausalen Effekte vorwiegend durch den Vergleich der geförderten Personen und einer Kontrollgruppe. Einfache deskriptive Vergleiche zeigen, dass sich beide Gruppen signifikant in ihren persönlichen Charakteristika unterscheiden. So nehmen zum Beispiel überdurchschnittlich hoch qualifizierte Personen an Maßnahmen der FbW teil. Die Gegenüberstellung der Wiederbeschäftigungswahrscheinlichkeiten beider Gruppen würde somit den Maßnahmeeffekt überschätzen, da höher Qualifizierte von vornherein bessere Arbeitsmarktchancen haben. Um diesem Selektionsproblem zu begegnen, werden sogenannte Matchingverfahren angewendet, die den geförderten Personen Nichtteilnehmer zuordnen, die hinsichtlich der beobachtbaren Merkmale identische Ausprägungen haben. Die Qualität der Schätzergebnisse hängt entscheidend von der Güte des Matches ab.

Im Rahmen dieser Studie soll gezeigt werden, dass Evaluationsansätze auf Basis von Matchingverfahren die Heterogenität der Beobachtungen nur unzureichend berücksichtigen und die Evaluationsergebnisse ohne zusätzliche Selektionskorrekturen verzerrt sind. Es ist zu vermuten, dass die in der Evaluationsforschung verwendeten Datensätze nicht alle entscheidungsrelevanten Variablen umfassen. Es fehlen insbesondere psychologische Merkmale wie zum Beispiel Selbstbewusstsein oder Motivation, die einen Einfluss auf die Ergebnisvariablen haben könnten. Werden diese Variablen im

Zuweisungsprozess nicht berücksichtigt kann eine unbeobachtbare Heterogenität existieren, die zu einer Selektionsverzerrung führt.
Um dies zu analysieren, wurde eine Befragung an ehemaligen Teilnehmern von Umschulungen durchgeführt. Anhand des modifizierten Leistungsmotivationstests von Hermans, Petermann und Zielinski (1978) wurde das Merkmal Motivation erhoben. Dieses ist in den üblich verwendeten Datensätzen nicht enthalten und somit eine unbeobachtbare Variable. Im Rahmen eines proportionalen Hazardratenmodells von Cox wird ein signifikant negativer Zusammenhang zwischen Motivation und der Dauer der Arbeitslosigkeit nach Ende der Maßnahme aufgedeckt. Eine kausale Interpretation erscheint zunächst problematisch, da Motivation zeitabhängig sein kann und somit ein Simultanitätsproblem besteht. In der weiteren Analyse werden jedoch Zusammenhänge aufgedeckt, die auf eine kausale Wirkung der Motivation auf die Arbeitslosigkeitsdauer hindeuten. Wird angenommen, dass dieser Zusammenhang als Regularität besteht, ist zu vermuten, dass Motivation auch kausal auf die Partizipationswahrscheinlichkeit wirkt und somit zu einer Selektionsverzerrung beiträgt. Diese Schlussfolgerungen lassen darauf schließen, dass matchingbasierte Evaluationsstudien einen Spielraum für inkonsistente Schätzergebnisse lassen.
Zur abschließenden Beurteilung sind jedoch weitere Untersuchungen bezüglich des Endogenitätsproblems und der kausalen Wirkung von Motivation auf die Maßnahmeteilnahme notwendig.

Die vorliegende Arbeit ist in folgender Weise gegliedert. Abschnitt II dient zur Einführung des Evaluationsgegenstandes. Zunächst wird die Arbeitsmarktentwicklung seit der Wiedervereinigung und anschließend die FbW als Instrument der aktiven Arbeitsmarktpolitik (AAMP) dargestellt. Im anschließenden Abschnitt III werden verschiedene Evaluationsansätze beschrieben, die in der deutschen Evaluationsforschung angewendet werden und Grundlage der weiteren Diskussion sind. Im Mittelpunkt dieser Darstellung steht das Matchingverfahren und die Erörterung ihrer Schwächen. Abschnitt IV fasst die Ergebnisse dieser Untersuchungen zusammen und gibt einen Überblick über die aktuellen Forschungstätigkeiten. In Abschnitt V wird schließlich der Zusammenhang zwischen Motivation und Dauer der Arbeitslosigkeit anhand eines eigenen Datensatzes analysiert. Zur Instrumentalisierung des Merkmals Motivation werden zwei verschiedene Motivationskonstrukte der Psychologie und mögliche Messmethoden vorgestellt. Nach der Beschreibung des Datensatzes und der methodischen Vorgehensweise werden die Ergebnisse präsentiert und abschließend werden Sensibilitätsanalysen durchgeführt. Die

Schlussfolgerungen und die Darstellung der wesentlichen Resultate dieser Arbeit werden im letzten Abschnitt VI zusammengefasst.

II. Hintergrund

II.1 Deutscher Arbeitsmarkt seit der Wiedervereinigung

Die Lage am deutschen Arbeitsmarkt wird vor allem für Ostdeutschland als kritisch eingeschätzt (vgl. Sachverständigenrat 2002: 7). Die zentrale Planwirtschaft der ehemaligen Deutschen Demokratischen Republik war 1989 nicht auf die Wiedervereinigung vorbereitet und die institutionellen Rahmenbedingung der westdeutschen Ökonomie, die Relativpreise und der internationale Wettbewerb wirkten wie ein Schock auf die ostdeutsche Ökonomie. Es setzte ein notwendiger Transformationsprozess ein, der aber 1996 ins Stocken geriet. Das Bruttoinlandsprodukt je Erwerbstätiger in Ostdeutschland konvergierte zwischen 1996 und 2001 lediglich um 3,5 Prozent auf 70,3 Prozent des westdeutschen Niveaus (Sachverständigenrat 2002: 7). Diese verhaltene Entwicklung mündete in einem Ungleichgewicht am Arbeitsmarkt. Während im letzten Jahr die Arbeitslosenquote, gemessen als der Anteil der registrierten Arbeitslosen an allen zivilen Erwerbspersonen, in Ostdeutschland bei 18,4 Prozent lag, betrug sie in Westdeutschland etwa 8,5 Prozent[1]. Insgesamt waren im Jahr 2004 durchschnittlich 4,38 Millionen Personen arbeitslos gemeldet. Damit lag die Arbeitslosenquote für das gesamte Bundesgebiet gegenüber dem Vorjahr unverändert bei 10,5 Prozent[2].

Trotz der hohen Arbeitslosigkeit im letzten Jahr war der Arbeitsmarkt durch eine beträchtliche Dynamik gekennzeichnet. So waren circa 8,1 Millionen Zugänge in und rund 8 Millionen Abgänge aus der Arbeitslosigkeit zu verzeichnen. Der Anteil der Langzeitarbeitslosen (Personen, die ein Jahr oder länger arbeitslos sind) war mit 40,3 Prozent an allen registrierten Arbeitslosen jedoch sehr hoch.

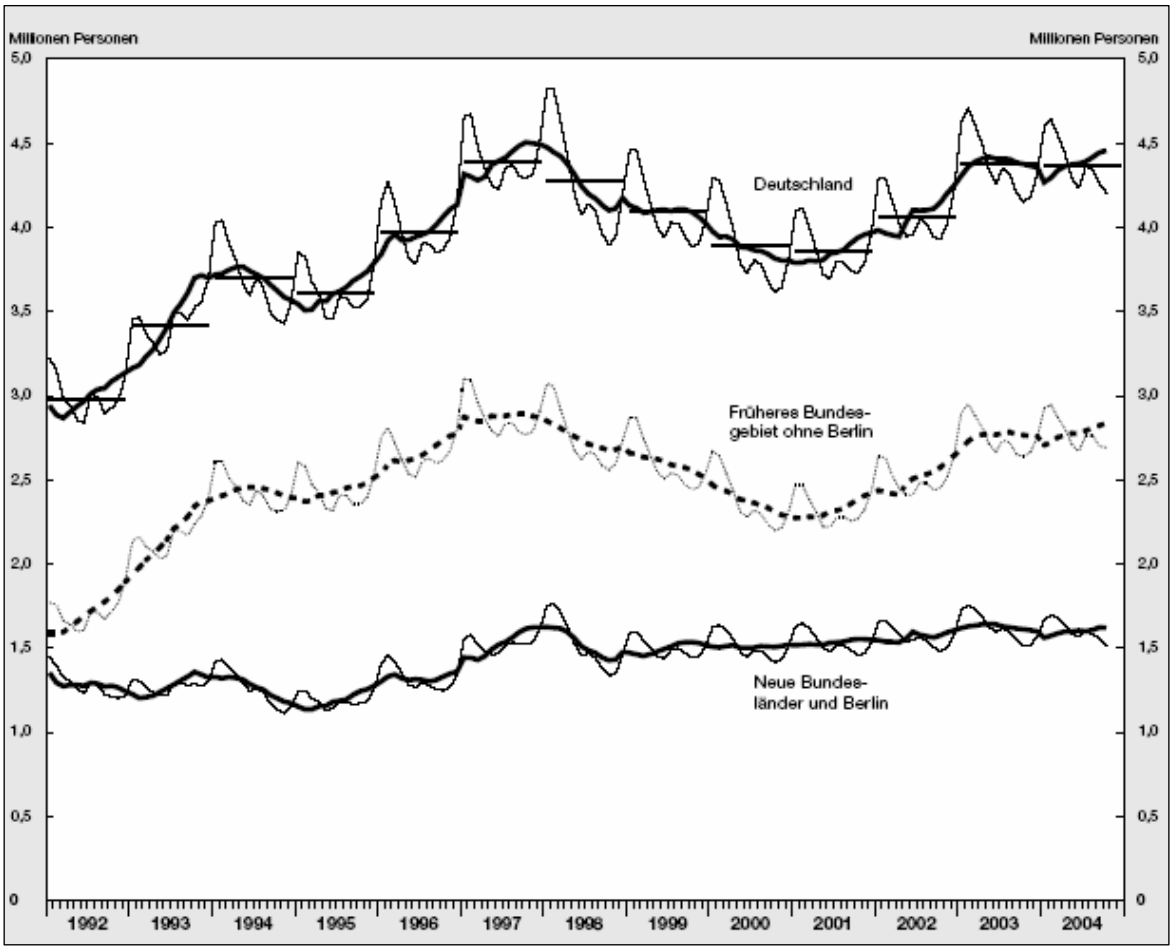
Anhand Abbildung 1 ist der Verlauf der registrierten Arbeitslosigkeit innerhalb der letzten 13 Jahre nachzuvollziehen. Seit 1992 ist ein Anstieg der Arbeitslosigkeit zu verzeichnen, der 1997 mit durchschnittlich 4,38 Millionen Arbeitslosen seinen ersten Höhepunkt erreichte. In den Folgejahren entspannte sich die Lage am Arbeitsmarkt etwas, blieb aber mit einer Arbeitslosigkeit um 4 Millionen weiterhin auf einem hohen Niveau.

[1] Die in diesem Abschnitt aufgeführten Daten sind, sofern nicht anders angemerkt, dem Jahresgutachten 2004/05 des Sachverständigenrats zur Begutachtung der gesamtwirtschaftlichen Entwicklung entnommen.

[2] Bis zum letzten Jahr wurden Teilnehmer an Eignungsfeststellungs- und Trainingsmaßnahmen noch zu den registrierten Arbeitslosen gezählt und werden nun, durch das Inkrafttreten des „Dritten Gesetztes für moderne Dienstleistung am Arbeitsmarkt (Hartz III), zu der verdeckten Arbeitslosigkeit gezählt. Die Arbeitslosenquote wäre nach der alten Gesetzgebung somit um einige Prozentpunkte gestiegen.

Um das tatsächliche Ausmaß der Unterbeschäftigung zu erfassen, ist es notwendig, neben den registrierten Arbeitslosen auch jene erwerbsfähigen Personen zu berücksichtigen, die zwar arbeitslos sind, aber nicht zu den registrierten Arbeitslosen gezählt werden. Zu diesem Zweck ermittelt der Sachverständigenrat die verdeckte Arbeitslosigkeit, die alle

Abbildung 1: Entwicklung der registrierten Arbeitslosigkeit



Quelle: Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2004: 178). Monatsendstände. Dünnen Linien: Grundzahlen; dicke Linien: Saisonbereinigte Werte; kurze Querstriche: Jahresdurchschnitte.

subventioniert beschäftigten Personen und alle Teilnehmer an AAMP-Maßnahmen umfasst[3]. Tabelle 1 veranschaulicht die Entwicklung der als arbeitslos registrierten Personen, der verdeckten Arbeitslosigkeit und der Teilnehmer der FbW. Demnach wurden im letzten Jahr 1,625 Millionen Personen der verdeckten Arbeitslosigkeit zugerechnet. Das tatsächliche Ausmaß der Unterbeschäftigung betrug somit cirka 6 Millionen beziehungs-

[3] Der IAB ermittelt zur Quantifizierung der Unterbeschäftigung die „Stille Reserve“ im engeren Sinne (nicht registrierte Arbeitssuchende und solche, die bei einer ungünstigen Arbeitsmarktlage nicht mehr arbeiten) und die Stille Reserve in Maßnahmen (Teilnehmer AAMP-Maßnahmen, die nicht erwerbstätig sind - zum Bespiel Maßnahmen der Arbeitsbeschaffung). In dieser Arbeit werden anstatt der Stillen Reserven in Maßnahmen die verdeckte Arbeitslosigkeit thematisiert, da für diese Kennziffern auch Daten zum Jahr 2004 vorliegen.
Eine ausführliche Abgrenzung der Stillen Reserven und verdeckten Arbeitslosigkeit findet sich im Jahresgutachten des Sachverständigenrats 2004 (181, 582).

Tabelle 1: Offene und verdeckte Arbeitslosigkeit (in Tausend)

Zeitraum	Registrierte und verdeckte Arbeitslosigkeit	Registrierte Arbeitslosigkeit	Verdeckte Arbeitslosigkeit	FbW in Vollzeit
1991	5198	2602	2596	406
1992	5646	2979	2667	675
1993	5998	3419	2579	583
1994	5937	3698	2239	467
1995	5791	3612	2179	500
1996	6104	3965	2139	505
1997	6340	4384	1956	400
1998	6175	4279	1896	322
1999	6042	4099	1943	333
2000	5700	3890	1810	324
2001	5620	3853	1767	322
2002	5820	4061	1759	307
2003	6015	4377	1638	232
2004	6003	4378	1625	165

Quelle: Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2000: 81; 2004: 180). Jahresdurchschnitte berechnet aus gerundeten Quartalswerten der BA.

weise 7,9 Millionen, wenn die Stillen Reserven im eigentlichen Sinne hinzugerechnet werden.[4]

Während Abbildung 1 noch suggeriert, dass sich die Arbeitslosigkeit seit 1993 rapide verschlechtert hat (Anstieg der registrierten Arbeitslosigkeit um etwa 28 Prozent) ist anhand Tabelle 1 auf keine wesentliche Veränderung der Unterbeschäftigung seit 1993 zu schließen. Die registrierte und verdeckte Arbeitslosigkeit betrug zu diesem Zeitpunkt, wie auch 2004, circa 6 Millionen. Es ist lediglich eine Umverteilung von der verdeckten Arbeitslosigkeit hin zur registrierten zu verzeichnen – dieses jedoch auf einem hohen Niveau. Obwohl das Ausmaß der Unterbeschäftigung durch die verdeckte und registrierte Arbeitslosigkeit charakterisiert werden sollte, stand bis vor kurzem gerade die Arbeitslosenquote und die Zahl der registrierten Arbeitslosen in den Medien im Vordergrund. Erst der Anstieg der Arbeitslosenzahlen auf über fünf Millionen Arbeitslose im Januar 2005, der primär auf Umklassifizierungen und Ersterfassungen im Zuge der Zusammenlegung der Arbeitslosenhilfe und Sozialhilfe sowie Saison-Effekte zurückzuführen ist, entfachte eine Diskussion um die „Ehrlichkeit“ der Arbeitslosenstatistiken. Da die Zahlen der verdeckten Arbeitslosigkeit beziehungsweise der Stillen Reserven seit Jahren von der BA veröffentlicht werden, kann und konnte sich jeder ein genaues Bild über die Unterbeschäftigung machen. Unverständlich ist daher die mit Bekanntwerden der

[4] Im Jahresgutachten beruft sich der Sachverständigenrat (2004: 181) auf eine Schätzung des IAB.

jüngsten Arbeitslosenzahlen einhergehende Kritik (vgl. Bovensiepen 2005), dass diese Zahlen das Versagen der Arbeitsmarktreform widerspiegeln.
Der Sachverständigenrat schrieb im letzten Jahresgutachten (2004: 345):

> „So beunruhigend diese Zahlen auch sind, als Beleg für eine Verschlechterung der Arbeitsmarktlage oder gar ein Versagen beim Übergang zum Arbeitslosengeld II sind sie gänzlich untauglich."

Von der prekären Arbeitsmarktlage sind vor allem unqualifizierte Arbeitskräfte betroffen. Während der Strukturwandel der Wirtschaft die Arbeitsnachfrage nach qualifizierten Tätigkeiten im Dienstleistung- und Produktionssektor steigert, ist keine gesteigerte Nachfrage nach gering qualifizierten Arbeitskräften zu verzeichnen (Spitznagel/Vogler-Ludwig 2004: 2). Die Arbeitslosenquote für Personen mit einem Hoch- beziehungsweise Fachhochschulabschluss lag in (West-) Ostdeutschland bei (3,3) 5,6 Prozent und betrug für Personen ohne Berufsabschluss in (West-) Ostdeutschland (19,8) 49,1 Prozent (Koch/Walwei 2003: 4).

II.2 Förderung der beruflichen Weiterbildung als Instrument der AAMP

II.2.1 Ziele der Förderung beruflicher Weiterbildung

Die staatliche Förderung der beruflichen Weiterbildung und Umschulung (FuU) ist seit der Verabschiedung des Arbeitsförderungsgesetzes (AFG) im Jahr 1969 ein wichtiges Instrument der AAMP. Das dritte Sozialgesetzbuch (SGB III) löste 1998 das AFG ab und bildet nun den ordnungspolitischen Rahmen der AAMP in Deutschland. Die Zielsetzung der Förderung der beruflichen Weiterbildung[5] lässt sich aus den Paragrafen 5, 77-96, 153-159 sowie 235c SGB III ableiten.
Demnach sollen berufliche Qualifikationen der Fortbildungsteilnehmer an veränderte Anforderungen am Arbeitsmarkt angepasst und die Möglichkeit geboten werden, einen fehlenden Berufsabschluss zu erwerben. Es wird angestrebt, Engpässe bestimmter Qualifikationen und den durch den strukturellen Wandel möglichen Überschuss an unqualifizierten Arbeitnehmern (mismatch) zu überwinden. Dadurch sollen sich die Arbeitsmarktchancen arbeitsloser Teilnehmer verbessern und drohende Arbeitslosigkeit abgewendet werden.

[5] Das SGB III nimmt die im AFG gebräuchliche Unterscheidung zwischen Fortbildung und Umschulungsmaßnahmen nicht vor und es wird zusammenfassend von der Förderung der beruflichen Weiterbildung gesprochen (FbW). Die vorliegende Arbeit folgt dieser Formulierung. Darüber hinaus werden für Maßnahmen der FbW synonym die Begriffe Qualifizierungsmaßnahme, Fortbildungen oder Weiterbildungen verwendet.

Neben diesen indirekten Effekten auf die Arbeitslosigkeit wirkt sich die AAMP auch direkt auf die Arbeitslosenquote aus, da Maßnahmeteilnehmer der AAMP nicht zu den registrierten Arbeitslosen gezählt werden. Dieses kann jedoch nicht als Erfolg der AAMP gewertet werden.

II.2.2 Entwicklung der FbW seit 1990

Mit der Wiedervereinigung wurde durch den massiven Einsatz von AAMP-Instrumenten in Ostdeutschland eine neue Dimension erreicht. So nahmen zwischen Ende 1989 und Ende 1993 mindestens 13 Prozent der Bevölkerung im erwerbsfähigen Alter in Ostdeutschland einmal an einer Maßnahme der FbW teil (Bielenski/Brinkmann/Kohler 1995: 14).

Die Teilnehmerzahl erreicht im gesamten Bundesgebiet 1992 mit etwa 675 Tausend den höchsten Stand und betrug 1998 nur noch 322 Tausend (vgl. Tabelle 1). Auf diesem Niveau verharrte die Teilnehmerzahl, und erst 2003 und 2004 sanken die Teilnehmerzahlen wieder deutlich. Im letzten Jahr nahmen nur noch 165 Tausend Personen an einer Maßnahme der FbW teil, was gegenüber dem Jahr 2002 ein Rückgang von 46 Prozent bedeutet. Maßgeblich dafür seien, gemäß der Bundesagentur für Arbeit (BA) eine stärkere Ausrichtung am Eingliederungserfolg sowie die Konzentration auf kürzere Maßnahmen und weniger die durch das Inkrafttreten der „Gesetze zur Modernisierung der Dienstleistung am Arbeitsmarkt“ (Hartz I und II) bedingten gesetzlichen Änderungen zum 1. Januar 2003.[6]

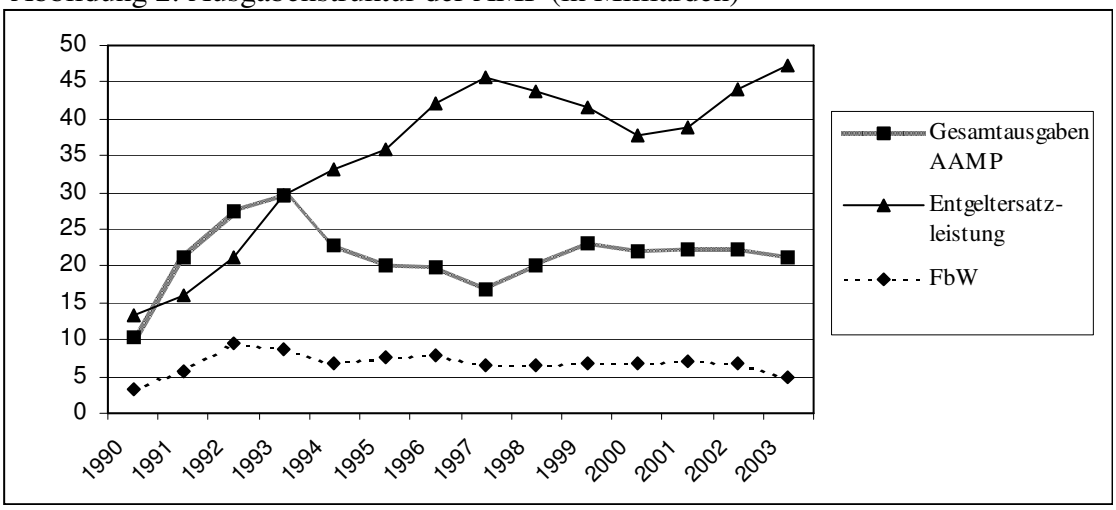
Abbildung 2 veranschaulicht die Entwicklung der Ausgaben für Maßnahmen der FbW und setzt sie ins Verhältnis zu den Ausgaben der gesamten AAMP sowie den Entgeltersatzleistungen[7]. Entsprechend dem Höchststand der Teilnehmerzahlen im Jahr 1992 erreichten auch die Ausgaben für Maßnahmen der FbW mit 9,41 Milliarden Euro zu diesem Zeitpunkt ihr Maximum. Die Ausgaben fielen in den Folgejahren und betrugen im

[6] Seit dem 01.01.2003 werden Teilnehmerberechtigte nicht mehr direkt von den Arbeitsämtern in Maßnahmen von Bildungsträger vermittelt, sondern bekommen Bildungsgutscheine ausgehändigt. Die Gutscheininhaber können diese bei einem Träger ihrer Wahl einlösen. Dieses soll den Wettbewerb unter den Bildungsträgern fördern, hat aber auch den Nachteil, dass sich mancher Bildungssuchender angesichts des großen Bildungsangebots überfordert sieht. Dieses und andere Gründe, wie vielleicht fehlende Motivation oder Eigeninitiative, führen dazu, dass nur ein Teil der Bildungsgutscheine eingelöst werden. Kühnlein und Klein (2003: 38) stellten fest, dass lediglich 55,2 Prozent der in Nordrhein-Westfalen zwischen Januar und Mai 2003 ausgegebenen Gutscheine eingelöst wurden.
Daher erscheinen weitere Wirkungsanalysen zur Einführung der Bildungsgutscheine notwendig, um diese Effekte quantifizieren zu können und die Aussage der BA zu bestätigen.

[7] In Anlehnung an den Statistiken der Amtlichen Nachrichten der BA umfassen die Entgeltersatzleistungen Arbeitslosengeld, Teilarbeitslosengeld, Insolvenzgeld und Arbeitslosenhilfe. Unterhaltszahlungen für Maßnahmeteilnehmer der AAMP werden zu den Ausgaben der AAMP und nicht zu den Entgeltersatzleistungen gerechnet.

Jahr 2003 ungefähr 5 Milliarden Euro, was etwa ein Viertel der Gesamtausgaben der AAMP entspricht. Dieser Anteil erreichte 1996 mit 40 Prozent seinen Höchststand. Mit

Abbildung 2: Ausgabenstruktur der AMP (in Milliarden)



Quelle: BA (1991-2004), Jahreszahlen

dem deutlichen Rückgang der Teilnehmerzahlen im Jahr 2003 ist auch ein Rückgang der Ausgaben im Jahr 2003 um 25 Prozent zu beobachten.

Der Vergleich der Ausgaben für aktive und passive AMP zeigt, dass sie sich bis 1993 parallel entwickelt haben und danach einem gegensätzlichen Verlauf folgten. Letzteres resultiert aus der Umverteilung der gesamten Unterbeschäftigung seit 1993, durch eine Abnahme der verdeckten und eine Zunahme der registrierten Arbeitslosigkeit. Lagen die Ausgaben der aktive und passive AMP 1993 bei jeweils 29,5 Milliarden Euro, wurden 2003 etwa 47,34 Milliarden Euro für die Entgeltersatzleistungen und ungefähr 21,20 Milliarden Euro für die aktive AMP ausgegeben. Seit 1991 werden, mit Ausnahme des Jahres 1996 und 1997, für die AAMP mehr als 20 Milliarden Euro aufgewendet.

Das hohe Niveau der Ausgaben der AAMP spiegelt die Orientierung der Arbeitsmarktpolitik wieder. Neben den hohen Transferzahlungen setzt der deutsche Staat zur Bekämpfung der Arbeitslosigkeit auf den Einsatz von AAMP-Maßnahmen.

Der hohe Ausgabenanteil der FbW unterstreicht die Bedeutung, die ihr als Instrument der AAMP in der Vergangenheit beigemessen wurde. Die letzten Reformen am Arbeitsmarkt lassen aber darauf schließen, dass die FbW zurückgehen wird. Dieses ist bereits an der Entwicklung der Teilnehmerzahlen seit 2002 zu erkennen und schlägt sich bereits in den Ausgaben für das Jahr 2003 nieder. Für 2004 dürften die Ausgaben für FbW noch einmal stark fallen und die Aufnahme dieser Daten in Abbildung 2 würde den fallenden Trend noch deutlicher hervorheben.

II.2.3 Definition der Maßnahmearten

In der Evaluationsliteratur wird in der Regel von der FbW als solcher gesprochen und es wird analysiert, welche Effekte aus der Teilnahme an einer Maßnahme der FbW resultieren. Durch diese Vorgehensweise wird jedoch außer acht gelassen, dass die Maßnahmen äußerst heterogen sind. Eine Unterscheidung der Maßnahmen nach ihrer Fristigkeit und ihrer sektoralen Ausrichtung scheint sinnvoll, da anzunehmen ist, dass sich die Maßnahmeeffekte unterscheiden.

Tabelle 2 fasst die verschiedenen Maßnahmearten zusammen, die in Deutschland zur FbW gehören[8].

Tabelle 2: Maßnahmearten der staatlich geförderten Weiterbildung

Programme	Description
Practice firm	Further training that simulates a job in a specific field of profession
Short training	Further training (i) with the aim of a general adjustment of working skills in the profession held; (ii) to obtain an additional qualification in the profession held; (iii) to obtain a first professional degree; planned duration $\leq$ 6 months.
Long training	Same types as short training with a planned duration > 6 months.
Retraining	Training to obtain a new professional degree in a field other than the profession currently held.

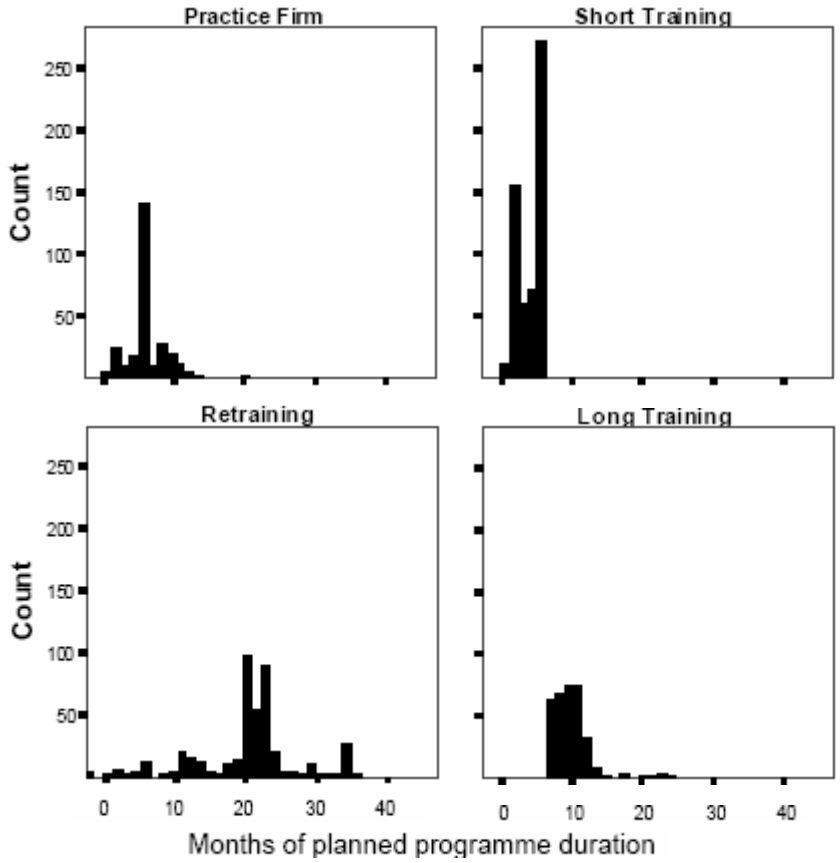
Quelle: Lechner, Miquel und Wunsch (2004a: 14)

Die Maßnahmen unterscheiden sich signifikant im Ausmaß der Investition in das Humankapital. So zielen manche Fortbildungen auf den Erwerb eines beruflichen Abschlusses (Umschulung), andere Maßnahmen wiederum dienen lediglich zur Anpassung beruflicher Fertigkeiten (Kurzmaßnahmen). Abbildung 3 zeigt die Verteilungen der Fortbildungsdauer nach Maßnahmeart differenziert, wie sie in einer Studie von Lechner, Miquel und Wunsch (2004a) dargestellt sind. Anhand der Verteilungen wird die Heterogenität der FbW deutlich unterstrichen. Umschulungen dauern typischerweise 21 Monate bis zwei Jahre, Teilzeitumschulungen können sogar bis zu drei Jahre dauern. Praxisorientierte Weiterbildungen (*practice firm*) und Kurzmaßnahmen hingegen erstrecken sich vorwiegend über einen Zeitraum von drei bis sechs Monaten.

Weiterhin zeigen Lechner, Miquel und Wunsch (2004a: 10), dass der überwiegende Teil der Umschulungen im Bereich der Dienstleistung (62 Prozent) und Fertigungssektor (27 Prozent) stattfindet. Umschulungen in der Agrarwirtschaft, im Ingenieurwesen oder im Bergbau spielen mit jeweils 3 bis 4 Prozent nur eine untergeordnete.

[8] Es gibt noch weitere Fortbildungsarten, die hier aber wegen ihres geringen Umfangs nicht erwähnt werden.

Abbildung 3: Verteilung der geplanten Fortbildungsdauer nach Maßnahmeart



Quelle: Lechner, Miquel und Wunsch (2004a: 15)

II.3 Evaluation der staatlichen Förderung beruflicher Weiterbildung

Mit der Einführung des Dritten Buches des Sozialgesetzbuches wurde erstmals gesetzlich verankert, dass die BA die Wirkung der aktiven Arbeitsförderung zu beobachten, zu untersuchen und auszuwerten hat (§280 SGB III). Weiterhin verlangt §7(1) SGB III, dass unter Anwendung der Grundsätze der Wirtschaftlichkeit und Sparsamkeit, der individuell am besten wirkende Maßnahmenmix dem jeweiligen Arbeitslosen zuteil werden soll. Dieses Ziel setzt implizit die Kenntnis der Wirksamkeit einzelner Maßnahmen voraus. Als Erfolgsindikator zieht die Agentur für Arbeit die Verbleibsquote heran. Diese Quote gibt an, wie viele Teilnehmer sechs Monate nach Maßnahmeende nicht mehr arbeitslos gemeldet sind.

Die Verbleibsquote ist jedoch ein unzureichender Erfolgsindikator (Sachverständigenrat 2002: 85), da zum einen Personen, die in die Stille Reserven oder in den frühzeitigen Ruhestand gewechselt sind, positiv auf die Verbleibsquote wirken und zum anderen manche Maßnahmeteilnehmer der FbW eine Arbeit aufnehmen, die nicht im Zusammenhang mit der beruflichen Weiterbildung steht und somit diese Arbeitsaufnahme auch nicht als Erfolg der Maßnahme zu werten ist. Desweitern werden Subsitutions-

(Verzerrung auf den Gütermärkten durch Veränderung der Lohnrelationen), Verdrängungs- (Maßnahmeteilnehmer besetzten Arbeitsplätze nicht geförderter Arbeitskräfte) und Mitnahmeeffekte außer acht gelassen. Letztere treten immer dann auf, wenn eine geförderte Person auch ohne Teilnahme an einer Maßnahme eine Arbeit gefunden hätte.
Um diese Defizite zu beheben, sollte ein idealer Evaluationsprozess drei Schritte umfassen (Hujer/Bellmann/Brinkmann 2000: 341):

1. Zunächst muss anhand von Erfolgskriterien geprüft werden, ob eine Maßnahmeteilnahme kausal für eine Verbesserung der Arbeitsmarktsituation nach Maßnahmeende ist.
2. Danach sollte eine Kosten-Nutzen-Analyse untersuchen, ob der individuelle Erfolg die eingesetzten Mittel rechtfertigt.
3. Schließlich muss analysiert werden, ob auf volkswirtschaftlicher Ebene positive Netto-Effekte der AAMP festzustellen sind.

Mitte der 90er Jahre setzte die Evaluationsforschung zur AAMP in Deutschland vermehrt ein und im Jahr 2002 sprach Lechner (2002a: 11) von einer beeindruckenden Anzahl an Evaluationsstudien.
Der Schwerpunkt der Forschung liegt auf mikroökonomischen Evaluationsstudien, die insbesondere den Erfolg von Maßnahmen der beruflichen Weiterbildung und der Arbeitsbeschaffung analysieren. Sie beurteilen die direkte Wirkung der AAMP auf eine Zielgröße. Als mögliche Zielvariable kommen der Lohn, Dauer der Arbeitslosigkeit nach der Maßnahme, beziehungsweise die Wiedereingliederung in den Arbeitsmarkt oder die Stabilität der Beschäftigung in Betracht.
Makroökonomische Studien beschäftigen sich mit der Wirkung der AAMP auf aggregierte Größen wie Arbeitslosigkeit, Beschäftigung, Löhne und Matching von Arbeitsnachfrage und -angebot. Unter Berücksichtigung von Substitutions- und Verdrängungseffekten soll der Nettoeffekt der AAMP auf den betrachteten Untersuchungsgegenstand ermittelt werden. Diese indirekten Effekte sind nur schwer zu quantifizieren, ihre Kenntnis ist jedoch notwendig, da die Effekte der AAMP sonst überschätzt würden.
Während die mikro- und makroökonomischen Evaluationsstudien Punkt 1 und 3 des beschriebenen Evaluationsprozess abdecken, gibt es keine Studien, die die fiskalischen Kosten der AAMP berücksichtigen (Fitzenberger/Hujer 2002: 155).

III. Methodik der Mikroökonomischen Evaluation
III.1 Modellrahmen der Kausalanalyse

Im Rahmen der mikroökonomischen Wirkungsanalyse von Maßnahmen der AAMP hat sich in der ökonometrischen und statistischen Literatur das kontrafaktische Verständnis der Kausalität durchgesetzt. Der formale Rahmen dieses Kausalitätsverständnisses orientiert sich am Roy-Rubin-Modell potentieller Ergebnisse (Roy 1951; Rubin 1974).

Dieses Modell geht von der Annahme aus, dass es zu jedem Zeitpunkt möglich ist, für jedes Individuum die potentiellen Ergebnisse Y^1 und Y^0 zu bestimmen. Y^1 beschreibt die Ergebnisvariable nach der Partizipation an einer Qualifizierungsmaßnahme und Y^0 das Ergebnis bei Nichtteilnahme. Somit wird, ungeachtet des Teilnehmerverhaltens der einzelnen Person, die Welt vollständig beschrieben.

Weiterhin wird unterstellt, dass die potentiellen Ergebnisse eines Individuums unabhängig von dem Teilnahmeverhalten anderer sind. Diese Annahme wird in der Literatur *stable unit treatment value assumption* (SUTVA)[9] genannt.

Der Effekt eines kausal wirkenden Faktors D auf Individuum i kann dann durch die Differenz der potentiellen Ergebnisse Y^1 und Y^0 nach einer Maßnahme gemessen werden. D bezeichnet hier eine binäre Indikatorvariable, die den Wert Eins bei Maßnahmeteilnahme und Null bei Nichtteilnahme annimmt. Entscheidend ist also der Ergebnisvergleich zweier Zustände, von denen aber nur einer beobachtbar ist.

Die Evaluationsliteratur konzentriert sich auf die Messung durchschnittlicher Effekte. Der damit verbundene Informationsverlust wird in Kauf genommen, um das Evaluationsproblem handhabbar zu machen. Im Mittelpunkt der meisten Evaluationsstudien steht der a*verage treatment effect on the treated* (ATET):

$$\theta_0 := E\left(Y^1 - Y^0 \mid D=1\right) = E\left(Y^1 \mid D=1\right) - E\left(Y^0 \mid D=1\right) \qquad (1)$$

Er misst den durchschnittlichen Effekt auf einen Teilnehmer, der sich durch die Teilnahme an einer Maßnahme der FbW verglichen zur Nichtteilnahme ergibt. Obschon es nun nicht mehr erforderlich ist, die individuellen kontrafaktischen Werte der Ergebnisvariable der

[9] Angesichts der enormen Ausmaße, die AAMP-Maßnahmen in den Neuen Bundesländern in den 90er Jahren annahmen, ist die Abwesenheit allgemeiner Gleichgewichtseffekte fraglich. In makroökonometrischen Evaluationen werden daher Rückwirkungen von Programmen auf Ergebnisse berücksichtigt.

geförderten Personen bei Nichtteilnahme zu bestimmen, besteht das Evaluationsproblem weiterhin in der Bestimmung des erwateten kontrafaktischen Werts $E\left(Y^0 \mid D=1\right)$[10].

III.2 Das Selektionsproblem

III.2.1 Quellen der Selektionsverzerrung

Aufgrund der Unmöglichkeit, am Ende einer Weiterbildung die erwartete Ergebnisvariable von Teilnehmern bei Nichtteilnahme zu bestimmen, könnte man versucht sein, diesen Wert durch den der tatsächlichen Nichtteilnehmer zu ersetzen. Im Folgenden soll analytisch dargestellt werden, dass diese Vorgehensweise unzulässig ist[11].

Sei Y_{it}^0 die Ergebnisvariable der Person i zum Zeitpunkt t, wenn sie nicht an einer Qualifizierungsmaßnahme teilgenommen hat. Findet die Maßnahme zum Zeitpunkt k statt, gilt:

$$Y_{it} = \begin{cases} Y_{it}^0 + D_i \theta_t & \text{wenn } t > k \\ Y_{it}^0 & \text{wenn } t < k \end{cases}, \quad [12] \tag{2}$$

wobei θ_t den Effekt der Weiterbildung auf einen Teilnehmer i zum Zeitpunkt t darstellt. Die erwarteten Ergebnisse einer geförderten Person und eines Nichtteilnehmers lassen sich dann durch

$$E\left(Y_{it} \mid D_i = 1\right) = E\left(\theta_t \mid D_i = 1\right) + E\left(Y_{it}^0 \mid D_i = 1\right) \tag{3}$$

und

$$E\left(Y_{it} \mid D_i = 0\right) = E\left(Y_{it}^0 \mid D_i = 0\right) \tag{4}$$

darstellen. Betrachtet man die Differenz der erwarteten Ergebnisvariablen wird offensichtlich, dass ein einfacher Vergleich dieser Werte nicht dazu ausreicht, den ATET zu bestimmen:

$$E\left(Y_{it} \mid D_i = 1\right) - E\left(Y_{it} \mid D_i = 0\right) = E\left(\theta_t \mid D_i = 1\right) + \left\{E\left(Y_{it}^0 \mid D_i = 1\right) - E\left(Y_{it}^0 \mid D_i = 0\right)\right\} \tag{5}$$

Die Differenz setzt sich aus dem Effekt des kausal wirkenden Faktors D und einem zweiten Term zusammen, der in der Literatur als Selektionsverzerrung bezeichnet wird.

---

[10] Rubin verweist darauf (1974: 690), dass neben der Messung des durchschnittlichen Effekts auch andere Lageparameter, wie der Median, als Schätzer in Betracht gezogen werden können. Der Median hat jedoch den Nachteil, dass er von der gemeinsamen Verteilung (Y^1, Y^0) abhängt und nicht, wie der Durchschnitt, alleine von den marginalen Verteilungen von Y^1 und Y^0 (Lechner 2002c: 7).

[11] Die Vorgehensweise folgt einem Ansatz von Heckman und Hotz (1989).

[12] Aus Gründen der Einfachheit nehmen Heckman und Hotz (1989: 864) an, dass der Weiterbildungseffekt zwischen den Individuen gleich ist.

Offensichtlich liegt eine Selektionsverzerrung vor, wenn sich der erwartete Wert der Zielvariablen von Teilnehmern und Nichtteilnehmern schon vor der Maßnahme unterscheiden.

Stammen zum Beispiel Teilnehmer vorwiegend aus Problemgruppen des Arbeitsmarktes, wird der Maßnahmeeffekt unterschätzt, wenn dieses nicht in der Kontrollgruppe berücksichtigt wird. Umgekehrt wird der Maßnahmeeffekt überschätzt, wenn Arbeitsvermittler unter den potentiellen Teilnehmer diejenigen aussuchen, die höhere Arbeitsmarktchancen haben.

Bei der Bewertung AAMP-Maßnahmen ist gerade zu bedenken,

> „dass sich aufgrund individueller Rationalität und/oder aufgrund der Zielgruppenorientierung der staatlichen Arbeitsmarktpolitik die Teilnehmer in ihren sozioökonomischen Charakteristika und in ihrer Arbeitsmarktsituation stark von den Nichtteilnehmern unterscheiden“

(Fitzenberger/Prey 1998: 45) und somit die durchschnittliche Ergebnisvariable von Teilnehmern wie Nichtteilnehmern sich von vornherein unterscheiden.

Nur wenn Maßnahmeteilnehmer durch experimentale Verfahren[13] bestimmt werden, ist

$$E\left(Y_{it}^{0} \mid D_i = 1\right) - E\left(Y_{it}^{0} \mid D_i = 0\right) = 0. \qquad (6)$$

Obwohl soziale Experimente in den USA durchgeführt wurden (vgl. Prey 1999: 106 f.), finden diese im europäischen Raum keine Anwendung, da sie mit ethischen Bedenken behaftet sind und sehr schnell auf forschungspraktische und methodologische Grenzen stoßen (Gangl/DiPrete 2004: 1). So müssen Evaluationsstudien zur deutschen AAMP auf nicht-experimentelle Datensätze wie den deutschen Sozio-ökonomischen-Panel (GSOEP)[14], den Arbeitsmarktmonitor Ost (AMO)[15] oder den Arbeitsmarktmonitor Sachsen-Anhalt (AMSA)[16] zurückgreifen. Daher wird sich die Analyse im folgendem auf die Evaluationsmethodik auf Grundlage nicht-experimenteller Daten konzentrieren.

[13] Beim experimentellen Verfahren werden Teilnehmer wie Nichtteilnehmer per Zufallsauswahl aus einem Pool von Teilnehmerberechtigten bestimmt. Potentielle Selektionseffekte hinsichtlich der Teilnahmeberechtigung treffen dann für Teilnehmer und die Kontrollgruppe zu und ein Vergleich der Ergebnisvariable beider Gruppen liefert einen unverzerrten durchschnittlichen Maßnahmeerfolg (Prey 1999: p.106).

[14] Eine ausführliche Beschreibung des GSOEP findet sich zum Beispiel bei Hanefeld (1987) oder bei Projektgruppe Sozio-ökonomisches Panel (1995). Siehe auch Tabelle A.1 im Anhang 1.

[15] Vergleiche Tabelle A.1 im Anhang 1

[16] Vergleiche Tabelle A.1 im Anhang 1

Um ein besseres Verständnis über die Ursachen der Selektionsverzerrung zu bekommen, sei im folgendem angenommen, dass Y_{it}^0 eine lineare Funktion eines beobachtbaren, X_{it}, sowie eines unbeobachtbaren Variablenvektors, U_{it}, ist. Somit ist

$$Y_{it}^0 = X_{it}'\beta + U_{it}, \tag{7}$$

wobei β ein Parametervektor ist. Aus Gleichung (2) und (7) ergibt sich

$$Y_{it} = X_{it}'\beta + D_i\theta_t + U_{it}. \tag{8}$$

Liegen keine experimentelle Daten vor, führt eine OLS-Regression der Gleichung 8 in der Regel zu verzerrten Schätzern von β und θ_t, da es eine Abhängigkeit zwischen U_{it} und D_i, $E(U_{it} \mid D_i, X_i) \neq 0$, geben kann. Die Teilnahme an einer Maßnahme kann somit von Eigenschaften einer Person abhängen, die gleichzeitig durch U_{it} einen Einfluss auf die Ergebnisvariable haben.

Um den Selektionsprozess für die Maßnahmeteilnahme zu modellieren, kann D_i als latente Variable D_i^* dargestellt werden, die linear von einem Vektor beobachtbarer Z_i und unbeobachtbarer V_i Variablen abhängt:

$$D_i^* = Z_i'\delta + V_i, \quad D_i = \begin{cases} 1 \text{ für } D_i^* > 0 \\ 0 \text{ für } D_i^* < 0 \end{cases} \tag{9}$$

wobei δ ein Parametervektor ist. Die Vektoren X_{it} und Z_i können unter Umständen die gleichen Variablen enthalten.

Ist eine Qualifizierungsmaßnahme nur einer bestimmten Personengruppe zugänglich (zum Beispiel Ausländer bei einem Sprachkurs), wird dieses durch Z_i aufgefangen. Als eine unbeobachtbare Variable (V_i) kann zum Beispiel Motivation interpretiert werden.

Selektion durch beobachtbare Variablen liegt vor, wenn die Abhängigkeit von U_{it} und D_i durch Z_i resultiert. Liegt hingegen eine Abhängigkeit von U_{it} und V_i vor, so spricht man in diesem Fall von einer Selektionsverzerrung aufgrund nicht beobachtbarer Variablen.

III.2.2 Deskriptive Evidenz des Selektionsproblems

Ein deskriptiver Vergleich von Charakteristika zwischen Teilnehmern und Nichtteilnehmern soll das Problem der Selektionsverzerrung verdeutlichen. Tabelle 3 zeigt durchschnittliche Merkmalsausprägungen von Personen, die zwischen 1990 und 1994 an Qualifizierungsmaßnahmen mit Unterhaltsgeldförderung (QS(mUG)) und solchen, die an keiner Maßnahme partizipiert (keine QM) haben. Die ausgewählten Merkmale weisen

signifikante Unterschiede der zwei Gruppen auf. Teilnehmer sind in der Regel jünger und besser ausgebildet. Der Frauenanteil der geförderten Personen beträgt 65.29 Prozent und ist damit deutlich höher als derjenige der Referenzgruppe. Außerdem ist die Beschäftigungsquote kurz vor der Maßnahme in der Teilnehmergruppe um cirka 34 Prozent niedriger als bei den Nichtteilnehmern.

Tabelle 3: Deskriptive Evidenz der Selektionsverzerrung

	QS (mUG)	keine QM	alle
Zahl der Personen	824	5248	8751
Frauenanteil	65.29	47.68	49.38
Beschäftigungsanteil vor Maßnahmebeginn	51.66	85.57	85.04
Alter (in Jahren)	37.12	43.35	41.51
Verheiratet (in %)	73.54	75.65	76.23
Teilfacharbeiter (in %)	1.35	5.46	3.63
Facharbeiter (in %)	55.68	58.01	51.14
Meister (in %)	5.54	7.88	7.83
Fachschulabschluss (in %)	19.46	16.65	20.67
Hochschulabschluss (in %)	15.41	7.70	13.97

Quelle: Prey (1999: 103). Anteile berechnet aus den Fallzahlen der 4. Welle (November 1992) Arbeitsmarktmonitor Ost.

Ähnliche Darstellungen finden sich auch bei Hujer, Maurer und Wellner (1997a: 31; 1998: 201), die außerdem die Nationalität berücksichtigen. Es zeigt sich, dass der Anteil deutscher Staatsbürger in der Referenzgruppe deutlich niedriger liegt als bei den geförderten Personen. An dieser Stelle sei aber darauf hingewiesen, dass diese deskriptiven Befunde keine generellen quantitative Aussagen zulassen. Der oben beschriebene Trend ist zwar konsistent, aber in seinem Ausmaß, je nach Datenlage und regionaler Betrachtung, verschieden stark ausgeprägt. So fällt zum Beispiel der Altersunterschied bei Eichler und Lechner (2000) kaum ins Gewicht.

III.3 Evaluationsmethoden

Es existiert bereits eine umfangreiche Literatur über die methodischen Aspekte der Evaluation AAMP. Zur mikroökonometrischen Evaluation bietet die Abhandlung von Heckman et al. (1999) einen ausführlichen Überblick.

Dieser Abschnitt erhebt nicht den Anspruch einer dezidierten und vollständigen Darstellung aller methodischen Aspekte, sondern skizziert die Evaluationsansätze, die bei

der Diskussion der empirischen Wirkungsanalyse zur AAMP in Deutschland notwendig sind. Zu diesen gehören:

i) Ökonometrische Verfahren mit Selektionskorrektur und Instrumentvariablen (IV)-Schätzer
ii) Matchingansätze
iii) Differenz-von-Differenz (DvD) Methoden
iv) Ereignisanalyse

Während rein ökonometrische Verfahren in jüngster Vergangenheit stark kritisiert wurden, ist die Anwendbarkeit der IV-Schätzer, mangels geeigneter Instrumente, stark begrenzt. Diese Methoden werden daher nur kurz präsentiert.
In der neueren Evaluationsliteratur zur AAMP werden vorwiegend Matchingschätzer, die DvD-Methoden und ereignisanalytischer Modelle verwendet. Daher werden diese Verfahren ausführlich erläutert und eingehend diskutiert.

III.3.1 Ökonometrische Modelle mit Selektionskorrektur

Zur Berücksichtigung von Selektionsverzerrung aufgrund unbeobachtbarer Variablen werden in der Literatur[17] als rein ökonometrische Verfahren Instrumentvariablen-Ansätze oder auch klassische ökonometrische Selektionsmodelle vorgeschlagen.

Vollständige simultane ökonometrische Selektionsmodelle schätzen den Maßnahmeeffekt mit der Maximum-Likelihood Methode, bei der die Ergebnisgleichung 8 und die Partizipationsneigung (Gleichung 9) gemeinsam modelliert werden. Durch dieses Verfahren wird der Selektionsverzerrung implizit Rechnung getragen und der Maßnahmeeffekt kann konsistent geschätzt werden.

Sei Y_{it} die Ergebnisvariable mit einer dichotomen Ausprägung, die den Wert Eins für eine beschäftigte Person i und den Wert Null annimmt, wenn sie arbeitslos ist[18]. In Anlehnung an Gleichung (9) sei durch die dichotome Variable D_{it}, die Partizipationsneigung einer Person zum Zeitpunkt t dargestellt. Durch die simultane Schätzung beider Variablen wird implizit angenommen, dass eine gegenseitige Abhängigkeit des Erfolgskriteriums und der Partizipationsneigung besteht. Das zu schätzende Modell ist dann wie folgt spezifiziert:

[17] Prey (1999: 118), Fitzenberger/Prey (1998: 51 f.), Lechner (2002c: 14)
[18] Die hier gewählte Darstellung folgt der Arbeit von Prey aus dem Jahr 1999.

$$Y_{it}^{*} = X_{it}'\beta + D_{it}\theta + U_{it} \tag{10}$$

$$D_{it}^{*} = Z_{it}'\delta + \lambda Y_{it} + V_{it} \tag{11}$$

$$\text{mit } Y_{it} = \begin{cases} 1 & \text{für } Y_{it}^{*} > 0 \\ 0 & \text{für } Y_{it}^{*} \le 0 \end{cases} \quad \text{und} \quad D_{it} = \begin{cases} 1 & \text{für } D_{it}^{*} > 0 \\ 0 & \text{für } D_{it}^{*} \le 0 \end{cases}. \tag{12}$$

Es wird angenommen, dass die Fehlerterme einer bivariaten Normalverteilung mit einer Korrelation von ρ, folgen. Außerdem dürfen die Vektoren X_{it} und Z_{it} nicht identisch sein und müssen sich mindestens in einem Element unterscheiden (Prey 1999: 121).

Die Evaluationsstudien von Prey (1997, 1999) sowie Fitzenberger und Prey (1997, 2000) basieren auf diesem Ansatz. Darüber hinaus lassen sie eine Zustandsabhängigkeit der Erfolgskriterien zu, indem in die Gleichungen (10) und (11) jeweils die Ausprägung der Ergebnisvariablen der Vorperiode mit einfließen. Dieses erfordert zusätzliche Annahmen über den Anfangszustand Y_{i0}^{*}. Daher formulieren Prey und Fitzenberger eine weitere Gleichung, durch die der Anfangszustand modelliert wird und nehmen diese in das zu schätzende Modell mit auf. Die Modellierung der Anfangszustände durch die tatsächlich realisierten Variablen ist nicht zulässig, da nicht angenommen werden kann, dass Y_{i0} exogen ist oder der Prozess in t = 0 im Gleichgewicht war.

Vollspezifische ökonometrische Selektionsmodelle sind in der Literatur schon früh kritisiert worden. Die Hauptprobleme sind zum einen in den zugrunde gelegten Verteilungsannahmen der Fehlerterme zu sehen (Prey 1999: 123) und zum anderen in den Annahmen über die funktionale Form der Modellgleichungen. Im Zuge dieser Methodenkritik werden oft die Arbeiten von LaLonde (1986) sowie Ashenfelter und Card (1985) zitiert.

LaLonde hatte Zugriff auf experimentelle Datensätze und nutzte die Evaluationsergebnisse auf Basis dieser Daten, um nichtexperimentellen Schätzer zu beurteilen. Gemäß LaLonde (1986: 614) sei dies der beste Spezifikationstest, da Ergebnisse aus experimentellen Studien den Maßnahmeeffekt prinzipiell unverzerrt wiedergeben würden. Während sich Ashenfelter und Card (1985: 659) dieser Sichtweise anschließen und die Notwendigkeit der Durchführung von Experimenten zur Bewertung nichtexperimenteller Evaluationsmethoden betonen, sind mittlerweile Zweifel an der Konsistenz dieser Evaluationsergebnisse aufgekommen[19].

[19] Bei der Durchführung von Experimenten können sich praktische Schwierigkeiten ergeben, die zu möglichen Verzerrungen („substitution bias“ oder „randomization bias“) der geschätzten Maßnahmeeffekte führen (vgl. Prey 1999: 107 f.).

In seiner Arbeit stellt LaLonde (1986: 617) fest, dass sich die Ergebnisse auf Basis ökonometrischer Selektionsmodellen von den experimentellen Ergebnissen signifikant unterscheiden. Dieses ließ LaLonde erkennen, dass rein ökonometrische Verfahren zur Wirkungsanalyse eher ungeeignet sind.
Andere Spezifikationstests testen die Annahmen, die über die Verteilung der Fehlerterme getroffen werden oder die funktionale Form der Gleichungen. Im Rahmen dieser Tests zeigt sich, dass Schätzergebnisse höchst sensitiv auf die zugrundegelegten stochastischen Annahmen reagieren (Ashenfelter/Card 1985: 659; LaLonde 1986: 613 f.). Fitzenberger und Speckesser (2000: 362) fassen zusammen, „dass meist auf Basis fehlspezifizierter und restriktiver funktionaler Form eine Vielzahl von Schätzergebnissen möglich sind."

Zu einem ähnlichen Ergebnis kamen auch Fitzenberger und Prey (1998: 51) bei der Bewertung ihrer oben erwähnten Arbeiten. In vielen Fällen würden die geschätzten Modelle keine ausreichende Selektionskorrektur bezüglich unbeobachtbarer Charakteristika ermöglichen. Sie stellen fest, dass die geschätzten Maßnahmeeffekte äußerst sensitiv auf die parametrische Spezifikation der Qualifizierungsgleichung reagieren.

Instrumentvariablenansätze werden angewendet, wenn die vorhandenen Daten nicht dazu ausreichen, die Qualifizierungsteilnahme in der Erfolgsgleichung (8) exogen zu modellieren. Dieses setzt aber die Verfügbarkeit von Variablen, S, voraus, die:

i) die Partizipationsneigung beeinflussen ($\mathrm{cov}(S_i, D_i) \neq 0$) und

ii) nicht mit dem Fehlerterm der Erfolgsgleichung korreliert ($E(S_i U_{it}) = 0$)

beziehungsweise nicht explizit in der Ergebnisgleichung enthalten sind.

Betrachtet man die zeitunabhängige Darstellung der Erfolgsgleichung (8)

$$Y_i = X_i'\beta + D_i\theta + U_i \qquad (8')$$

mit

$$D_i^* = Z_i'\delta + V_i, \quad D_i = \begin{cases} 1 \text{ für } D_i^* > 0 \\ 0 \text{ für } D_i^* < 0 \end{cases} \qquad (9)$$

impliziert Annahme (ii), dass die Variable keinen direkten Einfluss auf das Ergebnis hat, sondern nur durch D auf Y wirkt (Ausschlussrestrektion). Gibt es ein Element S in Z aus Gleichung (9) oder kann eine Transformation *S* gebildet werden, die die Annahmen (i) und (ii) erfüllt, ist S (*S*) eine gültige Instrumentvariable (IV) (Angrist/Imbens/Rubin 1996: 445). Bei Kenntnis und Verfügbarkeit von S ist es möglich, den Maßnahmeeffekt

konsistent zu schätzen. Der IV-Schätzer nimmt im Falle eines binären Instruments mit den möglichen Ausprägungen 0 und 1 folgende Form an (Durbin 1954):

$$\theta^{IV} = \frac{\hat{cov}(Y_i, S_i)}{\hat{cov}(D_i, S_i)} = \frac{\sum_{i=0}^{N} Y_i S_i / \sum_{i=1}^{N} S_i - \sum_{i=1}^{N} Y_i (1-S_i) / \sum_{i=1}^{N} (1-S_i)}{\sum_{i=0}^{N} D_i S_i / \sum_{i=1}^{N} S_i - \sum_{i=1}^{N} D_i (1-S_i) / \sum_{i=1}^{N} (1-S_i)}. \quad (13)$$

Da S und D(S) nur den Wert 1 oder 0 annehmen, geben die Komponenten der Differenz des Zählers jeweils die durchschnittlich Höhe der Ergebnisvariable der Individuen mit $S_i = 1$ beziehungsweise $S_i = 0$ wieder. Die Komponenten des Nenners sind, bedingt der Ausprägung von S, jeweils als die Wahrscheinlichkeiten, dass ein Individuum an einer Maßnahme teilnimmt, interpretierbar:

$$\theta_0^{IV} = \frac{E(Y|S=1) - E(Y|S=0)}{P(D=1|S=1) - P(D=1|S=0)} \quad (14)$$

Die Anwendbarkeit dieses Schätzansatzes ist jedoch begrenzt, da sie einige restriktive Annahmen voraussetzt (Imbens/Angrist 1994: 446). Daher führten Imbens und Angrist 1994 einen neuen Parameter in die Evaluationsliteratur ein: den durchschnittlichen Effekt eines kausal wirkenden Faktors derjenigen, die aufgrund einer Veränderung des Instruments ihren Teilnehmerstatus verändern. Dieser Effekt wird *local avarage treatment effect* (LATE) genannt und misst im Allgemeinen nicht den ATET.

Die Voraussetzung zur Schätzung des LATE ist die *monotonicity*-Annahme. Demnach müssen alle Individuen auf eine Änderung des Instruments in die gleiche Richtung reagieren. Sei D(S) der Teilnahmestatus in Abhängigkeit des Instruments und S = 0 die Ausprägung des Instruments zum Ausgangszeitpunkt, sowie S = 1 nach einer Veränderung des Instruments zu einem späteren Zeitpunkt. Unter der monotonicity-Annahme gilt daher:

$$D_i(1) \leq D_i(0), \forall i \text{ oder } D_i(1) \geq D_i(0), \forall i. \quad (15)$$

Bei einer Veränderung der IV darf es also nicht vorkommen, dass einige Individuen zur Maßnahmeteilnahme übergehen, während andere Maßnahmeteilnehmer sich zur Aufgabe der Maßnahme entschließen oder umgekehrt. Der LATE hat die Form

$$\theta_{LATE}^{IV} = E\left[\left(Y^1 - Y^0\right) \middle| |D(1) - D(0)| = 1\right]. \quad (16)$$

Unterschiedliche IVs führen somit auch zu unterschiedlichen Interpretationen vom LATE. Ist die IV zum Beispiel eine Indikatorvariable für verschiedene Politiksysteme oder wird durch sie unterschiedliche Intensitäten der Arbeitsmarktpolitik widergespiegelt, identifiziert LATE den Effekt der Politikveränderung für diejenigen, die aufgrund der

veränderten Rahmenbedingungen ihren Teilnahmestatus an einer Maßnahme verändern. Werden andererseits persönliche Merkmale als IV verwendet, mag LATE weniger politikrelevant sein. Würde das Haushaltseinkommen als IV gewählt, so würde LATE die Änderung der Ergebnisvariable für die Personen angeben, die bedingt einer Haushaltseinkommensveränderung ihr Teilnahmeverhalten ändern würden.

Ein Vorteil der IV-Ansätze gegenüber den vollspezifizierten Selektionsmodellen ist darin zu sehen, dass keine Annahmen über die Verteilungseigenschaften von V_i gemacht werden müssen. Weiterhin gibt es keine Einschränkung bezüglich der Transformation von S_i oder dessen Bestimmungsgröße, so lange die Bedingungen (i) und (ii) erfüllt sind (Prey 1999: 120). Hujer, Maurer und Wellner (1997a) bestimmten zum Beispiel den propensity-score (hierzu Abschnitt III.3.2) als Instrument für die diskrete Indikatorvariable der Partizipationsneigung in der Schätzung der Erfolgsgleichung.

Problematisch am IV-Schätzer ist es jedoch, Variablen zu finden, die nicht mit dem Fehlerterm der Erfolgsgleichung korreliert sind. Die Annahme über die Exogenität der IV in der Erfolgsgleichung lässt sich nicht testen und „we must maintain this assumption by appealing to economic behavior or a gut feeling“ (Wooldridge 2000: 463). Eine IV als Indikatorvariable der institutionellen Rahmenbedingung gilt im Allgemeinen als exogene IV der Erfolgsgleichung 8 und deren Wirkung ist von zentraler ökonomischer Relevanz (Heckman 1997: 451). Die Verwendung persönlicher Merkmale als IV erscheint hingegen fragwürdig, da sie möglicherweise im Vektor X der Erfolgsg-leichung enthalten sind oder aber mit dem Fehlerterm korrelieren, da sie als direkt erfolgswirksame Variablen in der Erfolgsgleichung fallengelassen worden sind und somit vom Fehlerterm berücksichtigt werden.

Ein Nachteil der IV-Ansätze ist in der Interpretation der Evaluationsergebnisse zu sehen. Der LATE bezieht sich in Abhängigkeit des gewählten Instruments nur auf einen Teil der Population. Handlungspolitische Empfehlungen sollten jedoch auf Evaluationsstudien begründet werden, die die Wirkung beruflicher Weiterbildung auf alle Maßnahmeteilnehmer messen.

Die hier gewählte Darstellung soll dem Leser eine kurze Einführung in die Methodik der IV geben. Zur inhaltlichen Vertiefung bietet sich die Arbeit von Heckman aus dem Jahr 1997 an. Im Rahmen eines *dummy endgenous variable model* werden IV eingeführt und in den Kontext des Kausalitätsmodels von Roy-Rubin gesetzt.

Der LATE wird in einer Arbeit von Imbens und Angrist (1994) vorgestellt und in einer weiteren Studie (Angrist/Imbens/Rubin 1996) werden die Folgen der Verletzung verschiedener Annahmen bei der Anwendung des LATEs diskutiert.

III.3.2 Matchingansätze

Matchingverfahren haben in den letzten Jahren in der angewandten Statistik und in ökonometrischen Arbeiten zur Ermittlung kausaler Effekte deutlich an Bedeutung gewonnen (Gangl/DiPrete 2004: 3). Das Matching ist ein nichtparametrisches Verfahren und muss daher, im Gegensatz zu Regressionsanalysen, keine Annahmen über die funktionale Zusammenhänge von Kovariablen oder Verteilungseigenschaften der Fehlerterme treffen (Prey 1999: 115).

Ziel der Matchingverfahren ist die Bildung einer Kontrollgruppe, die mit der Gruppe der Maßnahmeteilnehmer im höchsten Maß vergleichbar ist. Der Matchingprozess weist jedem Teilnehmer $m > 0$ Personen aus der Menge der Nichtteilnehmer zu, die idealerweise die gleichen Ausprägung aller zu berücksichtigen Charakteristika aufweisen und sich nur hinsichtlich der Maßnahmeteilnahme unterscheiden.

Gelingt die adäquate Zusammenstellung einer Vergleichsgruppe, sind die Ergebnisvariablen direkt vergleichbar und man erhält einen unverzerrten Schätzer für den hypothetischen Wert der Ergebnisvariable bei Nichtteilnahme, wenn ein Individuum an einer Maßnahme teilgenommen hat.

Die Schätzung des ATETs erfolgt in der Regel durch einfache deskriptive Verfahren, wie dem Mittelwertvergleich der Ergebnisvariablen der Teilnehmer und der gematchten Kontrollgruppe.

III.3.2.1 Propensity-score Matching

Matchingbasierte Evaluationsstudien stützen sich auf die methodischen Ansätze von Rubin (1977) sowie Rosenbaum und Rubin (1983). Danach ist unter Berücksichtigung aller für die Ergebnisvariable und das Partizipationsverhalten relevanten Einflussfaktoren das potentielle Niveau von Y unabhängig von der Maßnahmeteilnahme:

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X,$$

wobei „$\perp\!\!\!\perp$“ für stochastische Unabhängigkeit steht. Diese Annahme wird in der Literatur „bedingte Unabhängigkeitsannahme“ (CIA) genannt.

Bei Kenntnis aller für die Selektionsverzerrung verursachenden Kovariablen ist es also unerheblich, ob man Y^0 auf Basis der Teilnehmer mit den Charakteristika X oder Nichtteilnehmer mit den gleichen Merkmalsausprägungen determiniert. Somit folgt aus CIA:

$$E\left(Y^0 \middle| D=1, X=x\right) = E\left(Y^0 \middle| D=0, X=x\right) \tag{17}$$

und der ATET ist in Anlehnung an Gleichung 1 wie folgt bestimmbar:

$$\theta_0^M = E\left(Y^1 \middle| D=1, X=x\right) - E\left(Y^0 \middle| D=0, X=x\right). \tag{18}$$

Ein exaktes Matching von Teilnehmern und Nichtteilnehmern über den Kovariablenvektor X ist mit zwei erheblichen Problemen behaftet. Einerseits könnten Individuen innerhalb der Testgruppe Merkmalsausprägungen haben, die nicht in der Kontrollgruppe vorkommen. Andererseits tritt ein Dimensionalitätsproblem auf, wenn eine große Anzahl an Merkmalen berücksichtigt wird, da für jede Ausprägung ein Vergleich vorgenommen werden muss[20]. Eine Reduzierung des X-Vektors ist nicht zulässig, da jede Variable im Zuweisungsprozess berücksichtigt werden muss, die einen Einfluss auf Y oder D hat. Andernfalls wäre CIA verletzt.

Die Arbeit von Rosenbaum und Rubin aus dem Jahr 1983 bietet einen Lösungsansatz um Matchingverfahren praktisch anwendbar zu machen. Wenn $P(x) = P\left(D=1 \middle| X=x\right)$ die bedingte Partizipationswahrscheinlichkeit in Abhängigkeit der beobachtbaren Variabeln ist und $0 < P(x) < 1$ sowie CIA erfüllt ist, dann gilt die bedingte Unabhängigkeitsannahme auch für $P(x)$ und folglich (Rosenbaum/Rubin 1983: Theorem 2 und 3) ist:

$$E\left(Y^0 \middle| D=1, P(x)\right) = E\left(Y^0 \middle| D=0, P(x)\right). \tag{19}$$

$P(x)$ wird in der Literatur „propensity-score“ genannt und wird „aus einer logistischen oder einer Probit-Regression mit D als abhängiger und X als unabhängiger Variable bestimmt“ (Gangl/DiPrete 2004: 17). Analog zu Gleichung (18) ist:

$$\theta_0^{PM} = E\left(Y^1 \middle| D=1, P(x)\right) - E\left(Y^0 \middle| D=0, P(x)\right). \tag{20}$$

Kennt man die bedingte Verteilung der Partizipationswahrscheinlichkeit, können Nichtteilnehmer Teilnehmern zugeordnet werden, die einen gleichen propensity-score haben. Somit reduziert sich das Dimensionalitätsproblem, da nur noch hinsichtlich einer skalaren Größe gematcht wird (Hujer/Maurer/Wellner 1998: 205).

---

[20] Wenn alle Kovariablen nur binäre Variablen wären, ergäbe sich bei p verschieden Merkmalen 2^p unterschiedliche Ausprägungskonstellationen (bei p = 20 immerhin über eine Millionen).

Eine allgemeine Darstellung von Gleichung 20 ist:

$$\theta_1^{MP} = \sum_{i \in I_1} \omega_{N_0,N_1}(i)\left[Y_i^1 - \sum_{j \in I_0} W_{N_0,N_1}(i,j) Y_j^0 \right], \qquad (21)$$

wobei $\sum_{j \in I_0} W_{N_0,N_1}(i,j) Y_j^0 = E\left(Y^0 \mid D=1, P(x) \right)$ ist (vgl. Heckman/Ichimura/Todd 1998: 262; Smith/Todd 2004: 8). Der ATET ist der mit $\omega_{N_0,N_1}(i)$ gewichtete Durchschnitt der Differenz der Ergebnisvariablen der Teilnehmer und der mit $W_{N_0,N_1}(i,j)$ gewichteten Ergebnisse Y_j^0 der Kontrollbeobachtungen, wobei $\sum_{j \in I_0} W_{N_0,N_1}(i,j) = 1$ und $\omega_{N_0,N_1}(i)$ eine Gewichtung zur Berücksichtigung von Hetroskedastie ist (Heckman/Ichimura/Todd 1997: 629). Dabei bezieht sich i auf den Personenkreis der Teilnehmer, $i \in I_1$, und j auf Nichtteilnehmer, $j \in I_0$. Durch die Indexierung der Gewichte kommt zum Ausdruck, dass diese, abhängig von der Anzahl der Teilnehmer (N_1) und Nichtteilnehmer (N_o), variieren können.

Weiterhin bezeichne $C(P_i)$ den Geltungsbereich der Partizipationswahrscheinlichkeit der Person i, so dass jede Person j mit $P_j \in C(P_i)$ der Testperson i zugeordnet wird.

Die verschiedenen Matchingansätze unterscheiden sich in der Determinierung von $C(P_i)$ und der Spezifizierung der Gewichte $W_{N_0,N_1}(i,j)$. Aus der angewandten Statistik sei hier die Stratifikation, das Nearest-neighbor- und Caliper-Verfahren erwähnt, die jeweils nur einen Teil der Kontrollgruppe zur Konstruktion der kontrafaktischen Ergebnisvariablen berücksichtigen. Das Kernel-Verfahren bildet hingegen einen gewichteten Durchschnitt aus der gesamten Kontrollgruppe.

III.3.2.2 Nearest-neighbor und Caliper-Verfahren

Eine ausführliche Erörterung des Nearest-neighbor-Verfahren und dem verwandten Caliper-Verfahren bietet eine Arbeit von Dehejia und Wahba (2002).

Demnach stellt sich zunächst die Frage, wie viele (m = ?) Kontrollbeobachtungen einem Teilnehmer zugeordnet werden und ob eine Kontrollperson öfters als Match herangezogen werden kann (mit oder ohne „Zurücklegen“).

Matching mit Zurücklegen minimiert die propensity-score Distanz zwischen Kontroll- und Testpersonen und verringert somit die Verzerrung. Problematisch kann sein, dass unter Umständen nur wenige Vergleichspersonen für das Matching genutzt werden und andere, sehr ähnliche Nichtteilnehmer, nicht berücksichtigt werden. Dieses kann zu einer Erhöhung der Varianz des geschätzten Maßnahmeeffekts führen.

Bezüglich der Frage, wie groß m sein soll, ist es offensichtlich, dass die propensity-score Distanz für m = 1 minimiert wird. Eine Erhöhung von m hat jedoch den Vorteil, dass die Genauigkeit des Schätzers steigt. Die Varianz sinkt, da mehr Informationen genutzt werden, um die kontrafaktischen Ergebnisvariablen der jeweiligen Teilnehmer zu konstruieren (Smith/Todd 2004: 9). Die Matchqualität nimmt jedoch ab, da im Durchschnitt schlechtere Matches erzielt werden (Rosenbaum/Rubin 1983: 49).

Welcher Vorgehensweise der Vorzug zu geben ist, hängt entscheidend vom Datenmaterial ab. Gibt es nur eine minimale Überlappung in der Verteilung der propensity-scores, muss beim Matching ohne Zurücklegen auf Kontrollpersonen zurückgegriffen werden, die von den Testpersonen sehr verschieden sind, wenn „ähnliche" Personen erst einmal fürs Matching herangezogen worden sind. Daher sollte ein Matching mit Zurücklegen und m = 1 durchgeführt werden. Liegen der Analyse andererseits Daten zugrunde, in denen sich die Verteilungen der propensity-scores größtenteils überlappen, kann zugunsten der Genauigkeit der Schätzer Matching mit Zurücklegen und m > 1 angewendet werden. Die resultierende Verzerrung fällt umso geringer aus, je mehr sich die Verteilungen ähneln.

Das **nearest-neighbor-Verfahren** nimmt traditionell eine paarweise Zuordnung vor, wobei

$$C(P_i) = \min_j \left\| P_i - P_j \right\| , \quad j \in I_o \tag{22}$$

gilt. In der einfachsten Anwendung diesen Verfahrens wird jedem Teilnehmer i der Nichtteilnehmer j zugeordnet, der sich am geringsten im propensity-score von ihm unterscheidet. Die meisten Evaluationsstudien nutzen jedoch einen von Rosenbaum und Rubin (1985) sowie Rubin (1991) eingeführten Matching-Algorithmus oder Zuweisungsmodelle, die eine leicht abgewandelte Form von diesem darstellen[21]. Dieser Algorithmus nutzt als „Ähnlichkeitsindex" nicht nur den propensity-score, sondern berücksichtigt darüber hinaus zusätzliche Variablen. Üblicherweise wird auf Basis der geschätzten Partizipationswahrscheinlichkeit der Teilnehmer ein Caliper gebildet und die Nichtteilnehmer ermittelt, die in diesem Bereich liegen. Gibt es keine Vergleichspersonen, die in diesem Bereich liegen, wird der Teilnehmer nicht weiter berücksichtig. Gibt es mehrere Nichtteilnehmer, die in diesem Caliper liegen, werden die Individuen einander zugeordnet, für die die Mahalanobis-Distanz minimal ist. Zur Ermittlung der Mahalanobis-Distanz

[21] Der Matching-Algorithmus von Rosenbaum und Rubin, den zum Beispiel Hujer und Wellner (2000a: 43, 2000b: 419) verwendeten, ist im Anhang 2 aufgeführt. Ähnliche Algorithmen finden sich auch bei Fitzenberger und Prey (1998: 89 f.), Bergemann et al. (2000: 206) oder Eichler und Lechner (2002: 175,177).

werden neben den Partizipationswahrscheinlichkeiten zusätzliche Variablen, wie zum Beispiel Variablen zur Erwerbsgeschichte, berücksichtigt[22].
Ist der Zuweisungsprozess abgeschlossen, ergibt sich der ATET durch die Differenz der jeweiligen Werte der Ergebnisvariablen:

$$\theta_N^{PM} = \frac{1}{N_1} \sum_{i \in I_1} \left[Y_i^1 - Y^0_{j \in I_0 \mid P_j \in C(P_i)} \right], \tag{23}$$

wobei $\omega_{N_0,N_1}(i) = 1/N_1$ und $W_{N_0,N_1}(i,j) = 1$ ist[23].

Das **Caliper-Verfahren** berücksichtigt die Tatsache, dass die Distanz zwischen i und j sehr groß sein kann und schließt die Zuordnung von schlechten Matches aus. Der Person i werden alle m Personen aus I_0 zugeordnet, die in dem Bereich

$$C(P_i) = \left\{ P_j \,\middle|\, \left\| P_i - P_j \right\| < \varepsilon \right\} \tag{24}$$

liegen. Dehejia und Wahba (2002) wählten in einer Analyse die Gewichtung $\omega_{N_0,N_1}(i) = 1/N$ und $W_{N_0,N_1}(i,j) = 1/m_i$, wobei m_i die Anzahl der Personen ist, die auf i gematcht werden. Demnach ergibt sich:

$$\theta_C^{PM} = \frac{1}{N_1} \sum_{i \in I_1} \left[Y_i^1 - \frac{1}{m_i} \sum_{j \in I_0 \cap P_j \left\| P_i - P_j \right\| < \varepsilon} Y_j^0 \right]. \tag{25}$$

Naheliegender wäre aber eine Gewichtung von $W_{N_0,N_1}(i,j)$ in Abhängigkeit der propensity-score Distanz. Vergleichspersonen die dem Teilnehmer näher liegen, sollten stärker gewichtet werden als weiter Entfernte.

III.3.2.3 Stratifikation

Stratifikation ist eine grobe Approximation der bereits genannten Matching-Verfahren. Die propensity-scores werden in z verschiedene Intervalle (u = 1,....,z) unterteilt[24]. Für jedes Intervall wird über alle Teilnehmer wie Nichtteilnehmer des zugehörigen Intervalls die mittlere Differenz der Ergebnisvariablen ermittelt. Der ATET

---

[22] Eine hinreichende Erläuterung der Mahalanobis-Distanz ist innerhalb der Beschreibung des Matching-Algorithmus in Anhang 2 gegeben.

[23] Werden jedem Teilnehmer m > 1 Personen zugeordnet ist $W_{N_0,N_1}(i,j) = 1/m$ und

$$\theta_N^{PM} = \frac{1}{N_1} \sum_{i \in I_1} \left[Y_i^1 - \frac{1}{m} \sum_{j \in I_0 \mid P_j \in C(P_i)} Y_j^0 \right].$$

[24] Die Intervallbreite bestimmten Dehejia und Wahba in einer Arbeit aus dem Jahr 1999 derart, dass sich die durchschnittlich geschätzten Werte für P_i und P_j innerhalb der Intervalle nicht signifikant unterschieden.

ergibt sich dann durch die gewichtete Aufsummierung der mittleren Differenzen. Als $W_{N_0,N_1}(i,j)$ wird die Anzahl (N_u) der Personen i in den jeweiligen Intervallen verwendet:

$$\theta_S^{PM} = \sum_{u=1}^{z} \frac{1}{N_u} \left(\sum_{i \in I_1 | u} Y_i^1 - \sum_{j \in I_0 | u} Y_j^0 \right). \tag{26}$$

III.3.2.4 Kernel-Matching

Beim Kernel-Matching ergibt sich die Vergleichsgröße zu Y_i^1 durch einen gewichteten Durchschnitt der Ergebnisvariablen aller Kontrollpersonen. Nichtteilnehmer, die einen ähnlichen propensity-score wie die Vergleichsperson haben, werden stärker gewichtet als jene, die einen stark abweichenden propensity-score haben. Vollständigkeitshalber sind im folgendem der ATET und die Gewichtungsbestimmung aufgeführt[25]:

$$\theta_K^{PM} = \frac{1}{N_1} \sum_{i \in I_1} \left\{ Y_i^1 - \frac{\sum_{j \in I_0} Y_j^0 G\left(\frac{P_j - P_i}{a_n} \right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n} \right)} \right\}, \tag{27}$$

wobei

$W_{N_0,N_1}(i,j) = G\left(\frac{P_j - P_i}{a_n} \right) / \sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n} \right)$ ist. G(.) ist eine Kernel-Funktion und a_n ein sogenannter *bandwidth* Parameter (Heckman/Ichimura/Todd 1997: 630).

III.3.2.5 Selektionsverzerrung beim propensity-score Matching

Selektionsverzerrung tritt bei der Evaluation kausaler Effekte durch Matching auf Basis von propensity-scores immer dann auf, wenn die bedingte Unabhängigkeitsannahme und somit Gleichung 19 nicht erfüllt ist. Die resultierende Verzerrung

$$B\left(P(X)\right) = E\left(Y^0 \mid D=1, P(X)\right) - E\left(Y^0 \mid D=0, P(X)\right) \tag{28}$$

kann in drei Komponenten zerlegt und geschätzt werden (Heckman et al. 1996: 13418, 1998: 1042):

$$B\left(P(X)\right) = E\left(Y^0 \mid D=1, P(X)\right) - E\left(Y^0 \mid D=0, P(X)\right) = B_1 + B_2 + B_3. \tag{29}$$

Die ersten beiden Komponenten sind auf Unterschiede in den Verteilungen der Charakteristika zwischen der Test- und Kontrollgruppe zurückzuführen. Zum einem auf

[25] Zur ausführlichen Erörterung siehe Smith und Todd (2004) oder Heckman, Ichimura und Todd (1997, 1998).

mangelnde Überlappung der propensity-score Verteilungen (B_1) und zum anderem auf eine unterschiedliche Verteilung innerhalb des Überlappungsbereichs (B_2). Als dritte Komponente (B_3) wird die eigentliche Selektionsverzerrung durch unbeobachtbare Variablen genannt:

$$B_1 = \frac{1}{N_1} \sum_{\substack{i \in I_1 \\ P_i(X) \in S_{1P} \setminus S_p}} Y^0\left(P_i(X)\right) - \frac{1}{N_0} \sum_{\substack{j \in I_0 \\ P_j(X) \in S_{0P} \setminus S_p}} Y^0\left(P_j(X)\right), \tag{30}$$

$$B_2 = \frac{1}{N_1} \sum_{\substack{i \in I_1 \\ P_i(X) \in S_p}} E\left(Y^0 \mid D=0, P_i(X)\right) - \frac{1}{N_0} \sum_{\substack{j \in I_0 \\ P_j(X) \in S_p}} Y^0\left(P_j(X)\right), \tag{31}$$

$$B_3 = \frac{1}{N_1} \sum_{\substack{i \in I_1 \\ P_i(X) \in S_p}} \left[Y^0\left(P_j(X)\right) - E\left(Y^0 \mid D=0, P_i(X)\right) \right], \tag{32}$$

wobei $S_{1P} = \text{Support}\left\{P(X) \mid D=1\right\}$ und $S_{0P} = \text{Support}\left\{P(X) \mid D=0\right\}$ ist. Von den möglichen Ausprägungen der propensity-scores zwischen Eins und Null, bezeichnet Support den Bereich, in dem Ausprägungen der Partizipationswahrscheinlichkeiten für $i \in I_1$ oder $j \in I_0$ beobachtet werden. $S_p = S_{1P} \cap S_{0P}$ ist der Bereich (Support), in dem sich die beiden Verteilungen überlappen und $S_{1P} \setminus S_p$ ($S_{0P} \setminus S_p$) ist, gegeben D = 1 (D = 0), der Bereich, in dem sich die Verteilungen nicht überlappen.

Matchingverfahren, die ausschließlich Beobachtungen aus dem gemeinsamen Support berücksichtigen, können die Verzerrung B_1 und B_2 eliminieren.

Verzerrungen aufgrund von Matching von Teilnehmern und Nichtteilnehmern aus dem nichtüberlappenden Bereich (B_1) werden ausbleiben, wenn nur über den Bereich des gemeinsamen Supports gematcht wird.

„The bias due to different density weighting is eliminated because matching on participant propensity-scores effectively reweights the nonparticipant data“ (Heckmann et al. 1996: 13418).

Nur die letzte Komponente, die eigentliche Selektionsverzerrung, bleibt nach einem Matching weiterhin bestehen.

Heckman et al. (1998: 1042 f.) zeigen im Rahmen ihrer Datenanalyse, dass die auf B_1 und B_2 zurückzuführende Verzerrung den größten Teil der durchschnittlichen Unterschiede beider Gruppen ausmacht. Die eigentliche Verzerrung (B_3) trägt zwar nur in geringem Maße zur Gesamtverzerrung bei, ist aber, verglichen mit dem geschätzten Maßnahmeeffekt, noch immer relativ hoch (Heckman et al. 1998: Tabelle 5).

III.3.2.6 Anforderungen an Datenmaterial

Die Validität matchingbasierter kausaler Schlussfolgerungen hängt entscheidend von der bedingte Unabhängigkeitsannahme ab. Um gegen die CIA nicht zu verstoßen, ist ein sehr informativer Datensatz erforderlich. Der Evaluator muss alle Charakteristika kontrollieren, die einen Einfluss auf die Ergebnisvariable und das Partizipationsverhalten haben, um vergleichbare Kontextbedingungen zwischen teilnehmenden und nichtteilnehmenden Personengruppen herzustellen.

Bei der Analyse kausaler Effekte sieht man sich daher einem Trade-off, zwischen der Notwendigkeit, CIA zu erfüllen und einem ausführlichen Datensatz gegenüber. Wird zur Bestimmung der Partizipationsneigung ein umfangreicher X-Vektor bestimmt, müssen alle Personen des Datensatzes fallengelassen werden, für die nicht alle Variablen erhoben worden sind. Würde man andererseits die Anzahl der relevanten Kovariablen reduzieren, läuft man Gefahr, die bedingte Unabhängigkeitsannahme zu verletzen.

Eine weitere Verkürzung des Datensatzes ergibt sich aus den Vorüberlegungen des Abschnittes III.3.2.5. Demnach hängt die Genauigkeit eines Matchingschätzers von den berücksichtigten Beobachtungen ab. Werden nur Personen aus dem gemeinsamen Support gematcht, erhöht dieses die Validität des Schätzers. Somit steht der Evaluator vor dem Dilemma, zwischen einem validen Schätzergebnis oder einer allgemeinen Bewertung von Maßnahmeprogrammen wählen zu müssen. Wird nur der gemeinsame Support berücksichtigt, ist der gemessene Effekt nicht als allgemeiner Kausaleffekt zu interpretieren, da ein Teil der Versuchspersonen nicht für die Analyse herangezogen werden konnte.

Durch diese Vorgehensweise mussten manche Arbeiten bis zu 27 Prozent der Teilnehmer unberücksichtigt lassen (Bryson/Dorsett/Purdon 2002: 12). In diesem Fall wird das Matching bestenfalls einen partielle Beschreibung des Maßnahmeeffekts auf Teilnehmer beschreiben können. Dieses muss im Hinblick auf die politische Relevanz der Ergebnisse beachtet werden. Schließlich sollten arbeitsmarktpolitische Handlungsempfehlungen aus Analysen abgeleitet werden, die sich auf die Messung kausaler Effekte aller Teilnehmer beziehen und nicht nur auf Effekte einer Teilgruppe.

Die Wahl, das Matching auf Basis der gesamten Teilnehmergruppe oder nur einer Teilgruppe durchzuführen, kann erhebliche Auswirkungen auf die Qualität der Evaluationsergebnisse haben. Dieses verdeutlichen die Evaluationsstudien von LaLonde (1986) und Deheija und Wahba (2002), die gleiches Datenmaterial nutzten.

Während LaLonde in seiner Analyse die gesamte Teilnehmergruppe berücksichtigt und nicht-experimentellen Schätzern jegliche Aussagekraft abspricht (LaLonde 1986: 617), kommen Deheija und Wahba (2002) unter Berücksichtigung einer Teilpopulation zu der Erkenntnis, dass (Nearest-neighbor und Caliper) Matching adäquate Ergebnisse erzielen kann.

III.3.2.7 Kritik am Matchingansatz

Matchingverfahren sind intuitiv einleuchtend und aufgrund ihrer soliden konzeptionellen Fundierung durch das Roy-Rubin-Modell attraktive Instrumente der Kausalanalyse. Verglichen mit parametrischen Modellen der Ökonometrie, treffen Matchingansätze nur wenige einschränkende Annahmen und sind als komplementäre Verfahren zu den rein ökonometrischen Verfahren zu verstehen (Deheija/Wahba 2002: 160).

Trotzdem ist das Matching mit zwei Kernproblemen behaftet. Zum einem ist die Güte der Ergebnisse stark von dem verfügbaren Datenmaterial abhängig und zum anderen führt der Matchingschätzer zu inkonsistenten Ergebnissen, wenn eine Selektionsverzerrung aufgrund unbeobachtbarer Variablen vorliegt. Diese hat einen signifikanten Einfluss auf die geschätzten Effekte (Heckman et al. 1996, 1998) und daher sollte Matching in diesem Kontext nicht angewendet werden (Dehejia/Wahba 2002: 160). Dieses ist aber nicht als verfahrensspezifische Schwäche der Matchingansätze, denn vielmehr als allgemeines Problem der Evaluationsforschung zu verstehen.

Die Validität von Matchingschätzer kann nicht per se beurteilt werden. Vielmehr hängen sie gravierend von der Güte der Datenlage ab. Matchingergebnisse können nur unter Berücksichtigung von Sensibilitätsanalysen interpretiert werden. So kann die verbleibende Selektionsverzerrung innerhalb der Matchinganalyse zwar gering ausfallen, aber sensibel auf die Wahl der berücksichtigten Teilgruppe sowie der Variablen, die zur Schätzung der propensity-scores herangezogen werden, reagieren[26]. Dieses variiert je nach dem zugrundegelegten Datenmaterial. Nur wenn sich die propensity-score Spezifikation robust gegenüber Veränderungen der beachteten Variablen erweist, kann eine niedrig geschätzte Selektionsverzerrung auf eine hohe Genauigkeit des Schätzergebnis hindeuten.

Als Fazit der Methodendiskussion um das Matchingverfahren hielt Deheija (2004: 2f.) fest, dass das Matching keine „*magic bullet*“ Methode sei, um kontrafaktische

[26] Eine ausführliche Diskussion findet sich bei Deheija (2004) und Smith/Todd (2004).

Ergebnisse zu schätzen und nur unter bestimmten Umständen zu validen Schätzungen kausaler Effekte führt.

III.3.3 Differenz-von-Differenz Methode

III.3.3.1 Einführung

Der Differenz-von-Differenz (DvD) Schätzer und auf ihn basierende Schätzer gewannen in den 70er Jahren mit dem Aufkommen von Paneldaten große Popularität (Fitzenberger/Prey 1998: 49) und entwickelten sich zu einer häufig praktizierten Evaluationsmethode (Fitzenberger/Speckesser 2002: 11).

Ansätze zur Ermittlung kausaler Effekte durch Differenzenbildung gehen von einer Selektionsverzerrung aus, die auf individuelle, unbeobachtbare und zeitinvariate Komponenten zurückzuführen ist. Es wird angenommen, dass der Fehlerterm der Ergebnisgleichung (8)

$$Y_{it} = X_{it}'\beta + D_i\theta_t + U_{it}, \tag{8}$$

folgende Form hat:

$$U_{it} = \alpha_i + \upsilon_{it}, \tag{33}$$

wobei α_i und υ_{it} einen Erwartungswert von Null haben und υ_{it} unabhängig und identisch über t und alle Individuen verteilt ist (Prey 1999: 116). α_i stellt die konstante unbeobachtbare Heterogenität dar, *fixed effect*, und ist für die Abhängigkeit von U_{it} und dem Fehlerterm V_i der Gleichung 9 verantwortlich. Wird ein fixed effect angenommen, dann ist der bedingte Erwartungswert der Differenz von U_{it} vor und nach der Maßnahme unabhängig von D_i:

$$E\left(U_{it_1} - U_{it_0} \mid D_i, X_{it}\right) = 0 \qquad \forall \quad t_1, t_0,\ t_1 > k > t_0. \tag{34}$$

Ein einfache Vorher-Nacher-Vergleich der Ergebnisvariablen aller Personen zum Zeitpunkt t_1 ergibt unter der Annahme eines fixed effects:

$$\left(Y_{it_1} - Y_{it_0}\right) = \left(X_{it_1} - X_{it_0}\right)'\beta + D_i\theta_{t_1} + \left(\upsilon_{it_1} - \upsilon_{it_0}\right). \tag{35}$$

Fixed-effect-Schätzer nutzen (35) als Regressionsgleichung und ermitteln θ_{t1} durch eine OLS-Schätzung. Diese Vorgehensweise hat den Vorteil, dass keine Kontrollgruppe zur Ermittlung von Weiterbildungseffekten benötigt wird (Prey 1999: 117).

Jedoch mussten Untersuchungen über die Güte dieses Schätzers im Vergleich zu experimentellen Ergebnissen den fixed-effect-Schätzers oft verwerfen (Heckman/Hotz 1989: 871).

III.3.3.2 DvD-Schätzer

Der Differenz-von-Differenz Ansatz ist eine Erweiterung des Vorher-Nachher-Vergleichs und berücksichtigt nicht nur die zeitliche Veränderung der Ergebnisvariablen der Teilnehmer, sondern stellt ihr auch die entsprechende Veränderung für Nichtteilnehmer gegenüber. Somit werden exogene Veränderungen der Ergebnisvariablen nicht mehr als Ergebnis der Fortbildung interpretiert, da sie durch die Differenzbildung eliminiert werden.

Es wird angenommen, dass Gleichung (34) analog für die Individuen der Teilnehmergruppe, wie auch für Kontrollpersonen erfüllt ist. Den ATET erhält man durch die Differenz der durchschnittlichen Veränderung der Ergebnisvariablen von Teilnehmern und Nichtteil-nehmern:

$$\theta_t^{DvD} = E\left(Y_{t_1}^1 - Y_{t_0}^0\right) - E\left(Y_{t_1}^0 - Y_{t_0}^0\right). \tag{36}$$

Der DvD-Schätzer führt zu inkonsistenten Ergebnissen, wenn sich die geförderten Personen und die der Kontrollgruppe wesentlich unterscheiden und exogene Veränderungen unterschiedlich auf die Ergebnisvariablen der beiden Gruppen wirken. Denkbar wäre ein unterschiedlich starker Anstieg der Nachfrage auf den Märkten für hoch und gering qualifizierte Arbeitskräfte bei gleichzeitiger Selektion einer Qualifikationsgruppe in die Maßnahmen.

III.3.3.3 Konditionaler DvD-Ansatz

Der konditionale DvD-Schätzer begegnet diesem Problem, indem die Kontrollgruppe durch Matchingverfahren bestimmt wird. Dieser Ansatz vergleicht, unter Berücksichtigung der Kovariablen, die mittleren Veränderungen der Ergebnisvariablen der Teilnehmer mit denen der Kontrollgruppe:

$$\theta_M^{DvD} = E\left(Y_{t_1}^1 - Y_{t_0}^0 \mid D=1, P(x)\right) - E\left(Y_{t_1}^0 - Y_{t_0}^0 \mid D=0, P(x)\right). \tag{37}$$

Formal wird Gleichung 37 aus der Gegenüberstellung der tatsächlichen und kontrafaktischen Differenzen der Ergebnisvariablen

$$E\left(Y_{t_1}^1 - Y_{t_0}^0 \mid D=1\right) - E\left(Y_{t_1}^0 - Y_{t_0}^0 \mid D=1\right) \tag{38}$$

und einer schwächeren Form der bedingten Unabhängigkeitsannahme

$$(Y_{t_1}^1 - Y_{t_0}^0, Y_{t_1}^0 - Y_{t_0}^0) \perp\!\!\!\perp D \mid P(x)$$

abgeleitet. Unter dieser Annahme ist

$$E\left(Y_{t_1}^0 - Y_{t_0}^0 \mid D=1, P(x)\right) = E\left(Y_{t_1}^0 - Y_{t_0}^0 \mid D=0, P(x)\right) \tag{39}$$

und anstelle der kontrafaktischen Differenz von Y der geförderten Personen, kann die Veränderung der Ergebnisvariable der Nichtteilnehmer zur Evaluierung herangezogen werden.

Eine Spezifikation von Gleichung 37 ist der *local linear difference-in-difference estimator* (Smith/Todd 2004: 12):

$$\theta_{ll}^{DvD} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ \left(Y_{i,t_1}^1 - Y_{i,t_0}^0 \right) - \sum_{j \in I_0 \cap S_p} W(i,j) \left(Y_{j,t_1}^0 - Y_{j,t_0}^0 \right) \right\}, \tag{40}$$

wobei die Gewichtung wie beim Kernel-Matching erfolgen kann.

III.3.3.4 Kritik

Durch Zusammenführung der DvD-Methode und dem Matchingverfahren können die jeweiligen Schwächen des anderen Verfahrens größtenteils eliminiert werden. Die Achillesferse des konditionalen DvD-Schätzers sind zeitinkonsistente, unbeobachtbare Variablen. Hängt zum Beispiel die Partizipations- sowie die Beschäftigungswahrscheinlichkeit von individuellen Merkmale ab, die nicht im Datensatz enthalten sind, und unterscheiden sich die Ausprägung dieser Merkmale in t_0 und t_1, dann wird der Maßnahmeeffekt durch einen DvD-Schätzer nur unzureichend bestimmt.

Ein weiteres Problem ergibt sich durch die Wahl von t_0. Als Ashenfelter (1978: 55) die Einkommenswirkungen von Trainingsmaßnahmen analysierte, stellte er fest, dass:

> „...all of the trainee groups suffered unpredicted earnings declines in the year prior to training. […] This suggests that simple before after comperisons of trainee earnings may be seriously misleading evidence on the effect of training on earnings even when a non-random comparison group is available to account for economy-wide earnings changes."

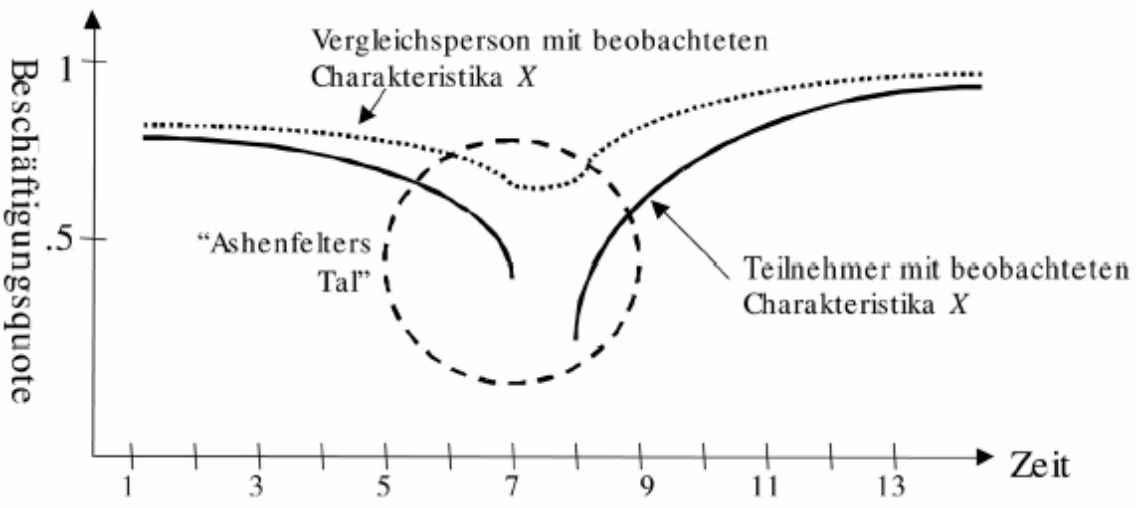
Dieses Phänomen, das in der Literatur als Ashenfelters Tal (Ashenfelter´s Dip) bezeichnet wird, wurde später auch bei der Ermittlung von Beschäftigungseffekten (Heckman/La Londe/Smith 1999) festgestellt und ist als empirische Regularität in Evaluationsstudien der AAMP anzusehen (Prey 1999: 118).

Dieser Effekt wird (bezogen auf die Beschäftigungsquote) auf eine Reduzierung oder sogar Einstellung der Suchintensität von Arbeitslosen schon vor Maßnahmebeginn zurückgeführt, wenn diese eine Maßnahmeteilnahme antizipieren (vgl. Bergemann et al.

2000: 201). Außerdem trägt in Deutschland die Gesetzgebung wesentlich zu diesem Effekt bei, da nur Personen zu der Teilnahme von Maßnahmen der FbW berechtigt sind, die unmittelbar vor der Maßnahme arbeitslos sind.

Aufgrund Ashenfelters Tal sind Teilnehmer und Nichtteilnehmer schon vor der Maßnahme nicht vergleichbar. Ein Reverenzlevel des DvD-Schätzers kurz vor der Maßnahme würde den Wirkungseffekt überschätzen. Abbildung 4 verdeutlicht graphisch das Problem von Ashenfelters Tal.

Abbildung 4: Ashenfelters Tal



Quelle: Fitzenberger und Speckesser (2002: 12)

Die Kurven stellen die durchschnittlichen Beschäftigungsquoten beider Gruppen dar. Die Maßnahme findet zwischen Zeitpunkt 7 und 8 statt. In der gleichen Zeit ist ein allgemeiner Rückgang der Beschäftigungsquoten zu verzeichnen. Ashenfelter Tal tritt in den Perioden 5 bis 7 deutlich auf. Hier fallen die Beschäftigungsquoten von Teilnehmer wesentlich stärker als die der Nichtteilnehmer. Um den Effekt zu berücksichtigen schlagen Fitzenberger und Prey (2000: 509) oder Bergemann et al. (2001: 12) einen DvD-Ansatz vor, bei dem t_0 hinreichend lange vor dem Maßnahmebeginn liegenden Zeitraum gewählt wird. Im Schaubild 4 würde man bei der Differenzenbildung zum Beispiel Werte der Ergebnisvariablen aus den Perioden 1 bis 3 heranziehen.

Der konditionale DvD-Schätzer erfährt in der jüngsten Evaluationsforschung eine große Popularität[27]. Kann dem Ashenfelters Tal noch relativ einfach Rechnung getragen

[27] Nachdem Fitzenberger Ende der 90er ausschließlich rein ökonometrische Verfahren verwendete greift er in jüngeren Studien auf den konditionalen DvD-Ansatz zurück (Bergemann/Fitzenberger/Speckesser 2001, 2004 und Bergemann et al. 2000). Andere bekannte Evaluationsforscher wie Eichler und Lechner wählten zuletzt auch diesen Ansatz (Eichler/Lechner 2000, 2002).

werden, sollte aber eines nicht außer Acht gelassen werden: Nur unter der Annahme der Abwesenheit zeitinkonsistenter, unbeobachtbarer Variablen können die Schätzergebnisse die tatsächlichen Effekte der Partizipation an Maßnahmen der FbW adäquat wiedergeben.

III.3.4 Hazardraten-Analyse

III.3.4.1 Einführung in die Verweildaueranalyse

Die Verweildaueranalyse nutzt die Arbeitslosigkeitsdauer nach Ende der Maßnahme als Indikator, um den Maßnahmeeffekt auf geförderte Personen festzustellen[28]. Ein einfacher Vergleich der durchschnittlichen Arbeitslosigkeitsdauer der Teilnehmergruppe und einer gematchten Kontrollgruppe ist unzulässig, da die Indikatorvariable rechtszensiert ist (am Ende des Betrachtungszeitraums sind nicht alle Personen der Population aus der Arbeitslosigkeit ausgeschieden). Ein anderer Grund ist in der Änderung der Zusammensetzung beider Gruppen im Zeitablauf durch den Austritt von Personen aus der Arbeitslosigkeit zu sehen. Die durch das Match erzielte Vergleichbarkeit der Gruppen ist zeitinkonsistent. Die Verweildaueranalyse gilt als valider ökonometrischer Ansatz, indem sie als Zielvariable nicht die Arbeitslosigkeitsdauer, sondern die Hazardrate definiert (Christensen 2001: 13). Die Hazardrate, h(t), ist die Wahrscheinlichkeit, in einem infinitiv kleinen Zeitraum einen beobachteten Zustand zu verlassen, unter der Bedingung, dass der Zustand bis zu diesem Zeitpunkt „angehalten“ hat:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \le T \le t + \Delta t \middle| T \ge t\right)}{\Delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}, \tag{41}$$

wobei T das Merkmal Verweildauer beschreibt, f(t) die Dichtefunktion und F(t) die Verteilungsfunktion von T ist. S(t) wird in der Literatur Überlebensfunktion (*Survivorfunktion*) genannt. Sie gibt die Wahrscheinlichkeit an, dass bis zum Zeitpunkt t kein Übergang stattgefunden hat.

Ermittelt man für die Gruppe der Teilnehmer und der gematchten Kontrollgruppe die Survivorfunktionen, können anhand einer graphischen Gegenüberstellung Unterschiede zwischen beiden Gruppen in der Arbeitslosigkeitsdauer festgestellt werden.

Anhand Abbildung 5 soll dieses kurz erläutert werden.

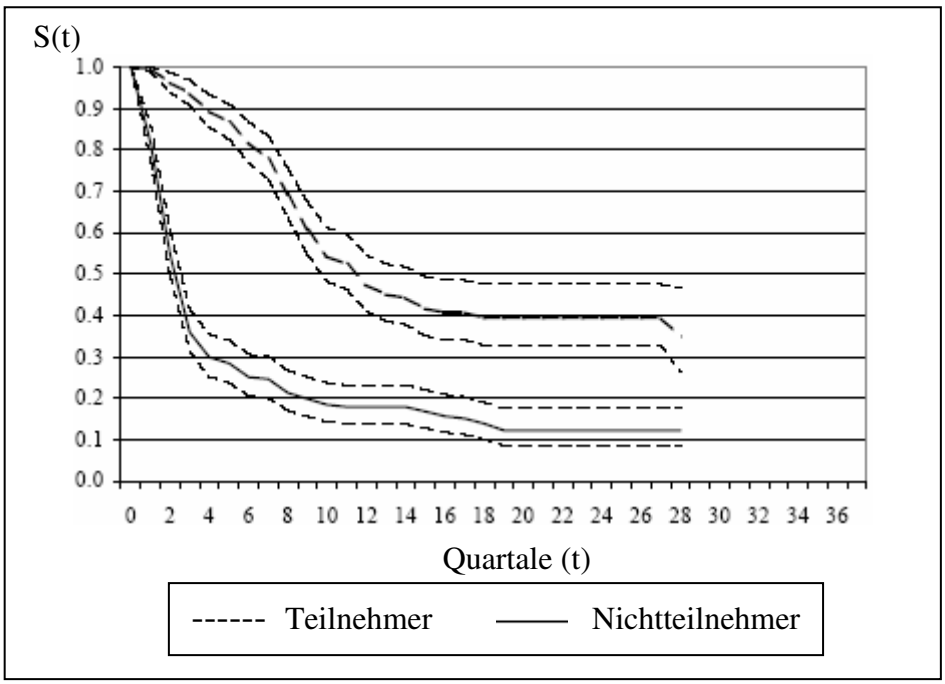
Die beiden Survivorfunktionen haben einen typischen Verlauf: Direkt nach Ende der Maßnahme (t = 0) ist die Wahrscheinlichkeit für beide Gruppen, dass noch kein Übergang in Arbeit stattgefunden hat, $S(0)$, 100 Prozent. Über die Zeit hinweg sinkt diese Wahrscheinlichkeit und bereits nach drei Quartalen ist ein signifikanter Unterschied

[28] Vergleiche für eine Einführung in die Hazardrate-Analyse zum Beispiel Kalbfleisch und Prentice (1980) oder Blossfeld, Hamerle und Mayer (1986).

zwischen den beiden Funktionen zu erkennen. Obwohl die Survivorfunktion danach etwas konvergieren, bleibt ein langfristiger Unterschied der Werte der Funktionen um 20 Prozent.

Die Survivorfunktion der geförderten Personen verläuft stets oberhalb der der Nichtteil-

Abbildung 5: Survivorfunktionen



Quelle: In Anlehnung an Reinowski, Schultz und Wiemers (2004: 20).

nehmer. Letztere beenden im Durchschnitt also früher die Arbeitslosigkeit als Maßnahmeteilnehmer. So haben zum Beispiel 50 Prozent der Kontrollgruppe bereits nach etwa 10 Monaten eine Arbeit aufgenommen. Bei Teilnahme an einer Maßnahme ist das gleiche Resultat erst nach cirka 44 Monaten zu beobachten. Geht man von einer hohen Qualität des Matches aus und unterstellt, dass der einzige Unterschied der beiden Gruppen in der Maßnahmeteilnahme liegt, ist in diesem Beispiel ein negativer Effekt der beruflichen Weiterbildung festzustellen.

III.3.4.2 Das Modell der proportionalen Hazardraten von Cox: Partial-Likelihood

Mittels ereignisanalytischer Methoden können die Effekte individueller Merkmale auf die Arbeitslosigkeitsdauer analysiert und die Ergebnisse deskriptiver Untersuchungen statistisch abgesichert werden. Das *proportional Hazards Model* von Cox (1972, 1975) ist

eines der populärsten Modelle der Verweildaueranalyse und wird ebenfalls in Evaluationsstudien zur AAMP angewendet[29].

Das Modell basiert auf der Annahme, dass die Hazardraten zweier Individuen sich proportional zueinander entwickeln. Es wird angenommen, dass sie sich aus einer multiplikativen Verknüpfung einer Basisübergangsrate (Baseline-Hazardrate) und einem log-linearen zeitinvariaten Kovariablenvektor[30] zusammensetzen:

$$h(t \mid X_i) = h_0(t) \exp\{X_i' \beta\}, \tag{42}$$

wobei $h_0(t) = h_0(t \mid 0)$ ist.

Die funktionale Form der Baseline-Hazardrate wird nicht näher spezifiziert und somit müssen, im Gegensatz zu vollständig parametrischen Modellen, keine ex ante Annahmen über den Verlauf der Hazardraten getroffen werden (Christensen 2001: 14). Ist die Basisübergangrate gleich eins, sind die individuellen Hazardraten konstant, und aus Gleichung 41 wird leicht ersichtlich, dass die Verweildauer demnach exponential verteilt sein muss.

Die Proportionalität der Hazardraten wird anhand der Betrachtung des Quotienten der Hazardraten zweier Individuen i und j deutlich:

$$\frac{h(t \mid X_i)}{h(t \mid X_j)} = \exp\{(X_i - X_j)'\beta\}, \qquad i \neq j. \tag{43}$$

Das Verhältnis der Hazardraten zweier Individuen ist somit über den gesamten Beobachtungszeitraum konstant.

Darüber hinaus lässt sich anhand des Quotienten die Wirkung einer Kovariablen besonders anschaulich interpretieren, wenn bei Konstanthaltung der jeweils anderen Variablen der Wert, z, der Variablen k um eine Einheit steigt:

$$\frac{h(t \mid X_i)}{h(t \mid X_j)} = \exp\{(z - z + 1)\beta_k\} = \exp\{\beta_k\}. \tag{44}$$

Der Koeffizient β_k ist der natürliche Logarithmus der Relation der Hazardraten, wenn z um eine Einheit steigt und alle übrigen Variablen unverändert bleiben. Die Hazardrate verändert sich somit um $(\exp\{\beta_k\}\text{-}1) \times 100\%$.

Ist zum Beispiel der geschätzte Koeffizient der Dummy-Variable „Maßnahmeteilnahme“ gleich 0,3, so ist die Hazardrate von geförderten Personen um knapp 35 % höher als die

[29] Die Studien von Hujer, Maurer und Wellner (1997a), Hujer und Wellner (2000a) sowie Reinowski, Schultz und Wiemers (2003, 2004) basieren zum Beispiel auf dieser Methode.

[30] Um die Notation einfach zu halten, geht das Merkmal der Maßnahmeteilnahme in den Kovariablenvektor mit ein.

von Nichtteilnehmern, wenn sie sich ansonsten nicht unterscheiden.
Ist das geschätzte Verhältnis der Hazardraten einer kontinuierlichen Variable (Alter in Jahren) zum Beispiel 1,15 bedeutet dieses, dass mit jedem zusätzlichen Jahr die geschätzte Hazardrate 1,15 mal soviel ist, wie für Personen, die ein Jahr jünger sind. Die Hazardrate einer 25-Jährigen wäre in diesem Falt $1{,}15^5$ soviel wie die einer 20-Jährigen.

Das proportionale Hazardratenmodell wird mit der Maximum-Likelihood-Methode geschätzt. Unter der Annahme proportionaler Hazardraten und ihrer Spezifikation gemäß Gleichung 42 nimmt die Likelihoodfunktion folgende Form an[31]:

$$L=\prod_{i=0}^{n} h_0(t)\exp\left\{X_i'\beta\right\}^{\delta_i}\exp\left\{-\int_0^t h_0(\upsilon)\exp\left\{X_i'\beta\right\}d\upsilon\right\}, \tag{45}$$

wobei δ_i den Wert Eins annimmt, wenn eine Beobachtung nicht zensiert ist und Null, wenn eine Person zum Ende des Befragungszeitraums noch immer arbeitslos ist.
Da die Basisübergangsrate nicht bekannt ist, kann diese Funktion nicht zur Schätzung der Parameter herangezogen werden. Cox schlug als Schätzansatz ein Verfahren vor, welches auf der sogenannten Partial-Likelihood beruht. Durch Faktorisierung der vollständigen Likelihood-Funktion kann man zwei partielle Likelihood-Funktionen erhalten, von denen eine nur von β und die andere von β und h_0 abhängt: $L(\beta, h_0)=L^*(\beta)L^*(\beta,h_0)$.

Zur Schätzung der Parameter wird nur $L^*(\beta)$ herangezogen. Da die Parameter ebenfalls im unberücksichtigten Teil der gesamten Likelihood-Funktion enthalten sind, erscheint dieses Vorgehen zunächst fragwürdig. Es konnte jedoch gezeigt werden, dass die Partial-Likelihood-Schätzung konsistent und asymptotisch normalverteilt ist (Amemiya 1985: 450). Außerdem ist sie asymptotisch äquivalent zu der Maximum-Likelihood-Schätzung der gesamten Likelihood-Funktion (Bailey 1984).
Die Partial-Likelihood-Funktion hat folgende Form:

$$L^*(\beta)=\prod_{f=1}^{F}\frac{\exp\left\{X_i'\beta\right\}}{\sum_{j\in R(t_f)}\exp\left\{X_j'\beta\right\}}. \tag{46}$$

Um diese Funktion interpretieren zu können, lohnt sich eine nähere Betrachtung der Herleitung der individuellen Beiträge zur partiellen Likelihood Funktion.
Zunächst müssen die Übergangszeitpunkte aller Personen in einer zeitlichen Reihenfolge geordnet werden: $t_1<t_2<...<t_F$. Weiterhin werden die „Risikogruppen“, $R(t_f)$, definiert,

[31] Die Herleitung der Likelihood-Funktion ist im Appendix im Anschluss dieses Abschnittes beschrieben.

die jeweils alle Personen umfassen, deren Verweildauer unmittelbar vor t_f noch nicht beendet ist, $R(t_f)=\left\{i\middle|t_i \geq t_f\right\}$. Diese Menge nimmt also über die Zeit ab.

Die bedingte Wahrscheinlichkeit, dass eine Person i aus der Risikogruppe zum Zeitpunkt t_f ihren Arbeitslosenstatus verlässt, ergibt sich somit aus der individuellen Hazardrate zum Zeitpunkt t_f, dividiert durch die kumulierten Hazardraten der Individuen, die zu der Risikogruppe gehören:

$$\frac{h(t_f \mid X_i)}{\sum_{j\in R(t_f)} h(t_f \mid X_j)} = \frac{h_0(t_f)\exp\left\{X_i'\beta\right\}}{\sum_{j\in R(t_f)} h_0(t)\exp\left\{X_j'\beta\right\}} = \frac{\exp\left\{X_i'\beta\right\}}{\sum_{j\in R(t_f)} \exp\left\{X_j'\beta\right\}}. \tag{47}$$

Das Produkt der bedingten Hazardraten über alle Übergangszeitpunkte ergibt die Partial-Likelihood-Funktion (46).

III.3.4.3 Partial-Likelihood-Schätzmethode nach Breslow und Efron

Ein Problem bei der Anwendung der erläuterten Methode ergibt sich daraus, dass mikroökonomische Datensätze oft nur für Zeitintervalle verfügbar sind und Übergangszeitpunkte mehrerer Personen in ein Intervall fallen können. Das Modell von Cox berücksichtigt dieses nicht, da durch die stetige Modellierung der Zeit der Austritt zweier Individuen nie zum gleichen Zeitpunkt stattfindet. Diesem als „Ties“ bezeichnetem Problem könnte man mit der Anwendung eines diskreten Hazardraten-Modells begegnen. Allison (1984: 22) argumentierte hingegen, dass in einem Datensatz ohne zeitvariierenden Kovariablen die durch Verwendung eines stetigen Hazardratenmodells auftretende Verzerrung der Schätzergebnis vernachlässigbar klein ist.

Breslow (1974) formulierte eine Abwandlung des Partial-Likelihood-Modells von Cox, um das Problem der Ties zu berücksichtigen. Anstatt die Übergangszeiten jeder einzelnen Person zu betrachten, werden Personen, die im gleichen Intervall eine Arbeit aufnehmen, zu Gruppen zusammengefasst und die bedingten Wahrscheinlichkeiten, dass eine Gruppe g_f von d Personen im Intervall t_f in Beschäftigung übergeht, formuliert:

$$\frac{h_{g_f}(t_f \mid s_f)}{\sum_{j\in R(t_f)} h(t_f \mid X_j)} = \frac{\prod_{i\in g_f} h_0(t_f)\exp\left\{X_i'\beta\right\}}{\left[\sum_{j\in R(t_f)} h_0(t_f)\exp\left\{X_j'\beta\right\}\right]^{d_f}} = \frac{\exp\left\{s_f'\beta\right\}}{\left[\sum_{j\in R(t_f)} \exp\left\{X_j'\beta\right\}\right]^{d_f}}, \tag{48}$$

wobei für die Gruppe der übergehenden Personen $s_f = \sum_{i\in g_f} X_i$ gilt.

Die Partial-Likelihood-Funktion zur Schätzung von β wird wieder aus dem Produkt über alle Übergangszeiten gebildet:

$$L^{*}(\beta)=\prod_{f=1}^{F}\frac{\exp\{s_{f}'\beta\}}{\sum_{j\in R(t_f)}\exp\{X_j'\beta\}^{d_f}} \tag{49}$$

Um die Einflüsse der individuellen Merkmale zu schätzen, wird auf ein iteratives, numerisches Lösungsverfahren zurückgegriffen. Die übliche Vorgehensweise zur Bestimmung der Parameter (Bildung der partiellen Ableitungen der Likelihood-Funktion und Lösung des Gleichungssystems) ist aufgrund der Komplexität des Schätzproblems nicht möglich. Standardmäßig wird deswegen der Newton-Raphson-Algorithmus verwendet (Reinowski/Schultz/Wiemers 2003: 25).

Der Schwachpunkt der Breslow Methode ist in den getroffenen Annahmen zur Behandlung der Ties zu sehen. Die erfolgte Umformulierung der bedingten „Gruppen-Hazardraten“ in Gleichung 48 ist problematisch, da sie davon ausgeht, dass die Anzahl der Personen in der Risikogruppe innerhalb eines Intervalls konstant ist. Die Gruppen-Hazardrate setzt sich aus den individuellen Hazardraten zusammen, die jeweils durch die kumulierte Hazardrate der Risikogruppe geteilt wird. Für jedes Individuum der jeweils betrachteten Gruppe ist die Risikogruppe gleich groß. Bei der Ermittlung der Gruppen-Hazardrate müsste jedoch die Risikogruppe für das zweite und alle nachfolgenden Individuen um die Anzahl der bereits ausgeschiedenen Gruppenmitglieder abnehmen. Besteht die Risikogruppe zu Beginn eines Intervalls zum Beispiel aus fünf Personen, e_1, e_2, e_3, e_4 sowie e_5, und scheiden Individuum e_1 und e_2 innerhalb des Intervalls aus, können zwei Szenarien zugrunde liegen. Zum einem könnte e_1 aus der Risikogruppe $e_1+e_2+e_3+e_4+e_5$ und anschließend e_2 aus der Risikogruppe $e_2+e_3+e_4+e_5$ ausgeschieden sein oder umgekehrt zunächst e_2 aus der Risikogruppe $e_1+e_2+e_3+e_4+e_5$ und dann e_1 aus $e_1+e_3+e_4+e_5$. Weil nicht beobachtet werden kann, in welcher Reihenfolge die Individuen innerhalb einer Gruppe ausscheiden, ignoriert die Approximation von Breslow dieses Problem und es wird einfach angenommen, dass beide Individuen aus der Risikogruppe $e_1+e_2+e_3+e_4+e_5$ scheiden.

Die Efron-Methode berücksichtig das Problem, indem sie davon ausgeht, dass die erste Risikogruppe $e_1+e_2+e_3+e_4+e_5$ sein muss und die zweite entweder $e_2+e_3+e_4+e_5$ oder $e_1+e_3+e_4+e_5$. Somit sind e_1 und e_2 mit einer Wahrscheinlichkeit von 0.5 in der zweiten Risikogruppe enthalten und das zweite Indivividuum scheidet aus der Risikogruppe $1/2(e_1+e_2)+e_3+e_4+e_5$ aus.

Die, durch die Annahmen hervorgerufenen Ungenauigkeiten der Breslow Methode fallen nicht ins Gewicht, wenn die Anzahl der betrachteten Austritte aus einer Risikogruppe relativ gering zu ihrer Größe ist. Liegt der Schätzung ein solcher Datensatz zugrunde, erzielt die Approximation nach Breslow durchaus gute Ergebnisse. Ist ein Datensatz jedoch durch einen kurzen Beobachtungszeitraum und einen großen Stichprobenumfang gekennzeichnet, wird die Methode von Efron (1977) als dominierende Approximation empfohlen (Cleves/Gould/Gutierrez 2002: 132).

III.3.4.4 Kritik ereignisanalytischer Modelle

Von einer verfahrensspezifischen Methodenkritik des PH-Modells wird hier abgesehen, zumal die Schwachpunkte des Modells bereits diskutiert wurden. Sollten die zugrundeliegenden Annahmen verletzt sein, bietet die Verweildaueranalyse auch andere Modelle, die herangezogen werden können (zum Beispiel diskrete Hazardratenmodelle oder Modelle mit zeitabhängigen Variablen).

Die Verweildaueranalyse an sich ist nicht für die Korrektur der Selektionsverzerrung verantwortlich und schätzt nur den Maßnahmeeffekt innerhalb des Samples. Die Eliminierung der Selektionsverzerrung erfolgt durch die Bildung einer adäquaten Kontrollgruppe durch Matchingverfahren und daher hängt die Güte der Ergebnisse von der Qualität des Matches ab. Ist die Selektionsverzerrung auf unbeobachtbare Variablen zurückzuführen, können die kausalen Effekte einer Maßnahmenpartizipation nicht valide geschätzt werden.

III.3.4.5 Appendix: Herleitung der Likelihood-Funktion im Proportionalen-Hazardraten-Modell von Cox

Bei der Anwendung der Maximum-Likelihood-Schätzung muss für jede Beobachtung eine Realisation des in Frage stehenden zufälligen Merkmals (T_i) vorliegen. Wenn die Daten rechtszensiert sind, lässt sich die genaue Ausprägung der Zufallsvariable nicht angeben. Sei L_i der Zeitpunkt, an dem der Beobachtungszeitraum des Individuums i endet, ergeben sich die Verweildauern in der Stichprobe durch $t_i = \min(T_i, L_i)$. Die Maximum-Likelihood-Methode bietet eine Möglichkeit, rechtszensierte Daten explizit im Schätzvorgang zu berücksichtigen. Zu diem Zweck wird der Zensierungsindikator

$$\delta_i = \begin{cases} 1 \text{ falls } T_i \le L_i \\ 0 \text{ falls } T_i > L_i \end{cases} \qquad \text{(A.1)}$$

eingeführt. Um die individuellen Beiträge zur Likelihoodfunktion zu ermitteln, müssen die gemeinsamen Wahrscheinlichkeiten von (T_i, δ_i) bestimmt werden.

Ist die Beobachtung einer Person zensiert ($\delta_i = 0$), ist $t_i = L_i$ und die Wahrscheinlichkeit, dass eine Person i zum Zeitpunkt L noch arbeitslos ist, gerade der Wert der Survivorfunktion an der Stelle L.

Für $\delta_i = 1$ ist $T_i < L_i$ und die gemeinsame Wahrscheinlichkeit

$$\begin{aligned} f(t_i, \delta_i = 1) &= f(t_i \,|\, \delta_i = 1)\, P(\delta_i = 1) \\ &= f(t_i \,|\, T_i < L_i)\,\big(1 - S_i(L_i)\big) \\ &= \frac{f(t_i)}{1 - S_i(t)}\big(1 - S_i(t)\big) = f(t_i). \end{aligned}$$

Somit lautet die Likelihood-Funktion:

$$L = \prod_{i=0}^{n} f_i(t_i)^{\delta_i}\, S_i(L_i)^{(1-\delta_i)} \tag{A.2}$$

Ferner werden zwei Beziehungen genutzt, um die Likelihoodfunktion zu modifizieren. Einerseits ist die Dichtfunktion der Verweildauer durch

$$f(t) = h(t)S(t) \tag{A.3}$$

gegeben (vgl. Gleichung 41). Andererseits kann unter Beachtung der Beziehung

$$\int_0^t h(\upsilon)d\upsilon = \int_0^t \frac{f(\upsilon)}{1 - F(\upsilon)}d\upsilon = \Big[-\ln\big(S(\upsilon)\big)\Big]_0^t = -\ln\big(S(t)\big) \tag{A.4}$$

die Survivorfunktion durch

$$S(t) = \exp\left\{-\int_0^t h(\upsilon)d\upsilon\right\} \tag{A.5}$$

dargestellt werden. Unter Berücksichtigt der Gleichungen A.3 und A.5 kann A.2 wie folgt formuliert werden:

$$L = \prod_{i=0}^{n} h(t_i)^{\delta_i} \exp\left\{-\int_0^t h(\upsilon)d\upsilon\right\}. \tag{A.6}$$

Unter den Annahmen des proportionalen Hazardratenmodells

$$h(t \,|\, X_i) = h_0(t)\exp\{X_i'\beta\},$$

folgt schließlich:

$$L = \prod_{i=0}^{n} h_0(t)\exp\{X_i'\beta\}^{\delta_i} \exp\left\{-\int_0^t h_0(\upsilon)\exp\{X_i'\beta\}d\upsilon\right\}. \tag{A.7}$$

IV. Stand und Perspektiven der Evaluationsforschung in Deutschland

IV.1 Überblick

Trotz der ernormen Anstrengungen der Evaluationsforschung in den letzten 15 Jahren bezeichnet Lechner (2002c: 21) die Lehre, die man aus diesen Studien ziehen kann, als eher „ernüchternd“[32]. Die Studien kommen, je nach Datenlage und methodischer Vorgehensweise, zu widersprüchlichen Ergebnissen.

Aber auch wenn die Studien ein homogenes Bild abgeben würden, wäre die Formulierung wirtschafts- und sozialpolitischer Handlungsempfehlungen nicht gerechtfertigt. Dieses liegt nicht etwa an mangelnden Kompetenzen der Evaluatoren, sondern an der schlechten Datenlage in Deutschland (Lechner 2002c: 21), die die heterogenen Maßnahmearten nicht ausreichend unterscheiden und nicht ermöglichen, die Selektion in die Maßnahme hinreichend zu berücksichtigen[33]. Dieses schmälert schon von vornherein die Validität der Forschungsergebnisse (Sachverständigenrat 2002: 86).

Darüber hinaus wäre eine differenzierte Evaluation nach Regionen und Personenkreisen ebenfalls wünschenswert, da auch hier angenommen werden kann, dass sich Maßnahmeeffekte unterscheiden[34]. Bei Kenntnis der unterschiedlichen Wirkungen sollten in den jeweiligen Regionen und für bestimmte Teilpopulationen vorwiegend jene Maßnahmen angeboten werden, von denen eine hohe Programmwirkung zu erwarten ist. Sicherlich ist es aus Gerechtigkeitsgründen bedenklich, das Leistungsangebot der AAMP rein nach Effizienzkriterien auszurichten. Trotzdem sollte der Einsatz von AAMP-Maßnahmen durch fundierte Evaluationserkenntnisse nach bestimmten Verhältnismäßigkeitserwägungen gesteuert werden.

Trotz der schwierigen Datenlage unternehmen einige Evaluationsstudien den Versuch, die Heterogenität der Maßnahmen zu berücksichtigen. So unterscheiden zum Beispiel Hujer, Maurer und Wellner (1997a) sowie Hujer und Wellner (2000a, 2000b) zwei beziehungsweise drei Maßnahmearten nach ihrer Fristigkeit. Andere Studien wiederum lassen unterschiedliche Maßnahmeeffekte für verschiedene Teilpopulationen zu. Häufig werden die Effekte der Fortbildung für Frauen und Männer getrennt evaluiert

[32] Zu diesem Ergebnis kommen auch Fitzenberger und Hujer (2002: 155), Fitzenberger und Speckesser (2002: 22) oder der Sachverständigenrat (2000: 86).

[33] Tabelle A.1 im Anhang 1 gibt einen kurzen Überblick über die Probleme bei der Anwendung des AMO, AMSA oder des GSOEPs im Rahmen der Evaluationsforschung AAMP-Maßnahmen.

[34] Die Studie von Reinowski, Schultz und Wiemers (2004) evaluiert den Effekt der FbW für verschiedene Subgruppen in Ostdeutschland. Der durchschnittliche Beschäftigungseffekt aller Maßnahmen auf alle Teilnehmer ist negativ. Darüber hinaus zeigen die Autoren, dass die Förderung sich besonders nachteilig auf Frauen und ältere Personen sowie Teilnehmer aus den Regionen Chemnitz und Leipzig (verglichen mit der Region Dresden) auswirkt.

(vgl. Fitzenberger/Prey (2000); Kraus/Puhani/Steiner (1999); Prey (1997, 1999) sowie Reinowski/Schultz/Wiemers (2004)). Die Studien von Staat (1997) und Reinowski, Schultz und Wiemers (2004) berücksichtigen zusätzlich unterschiedliche Alters- und Qualifikationsgruppen. Die letztgenannte Arbeit ist auch die einzige Studie, die über die übliche Unterscheidung der Maßnahmeeffekte für Ost- und Westdeutschland hinaus, eine feinere regionale Differenzierung vornimmt.

Es ist jedoch davon auszugehen, dass diese Ansätze die heterogenen Effekte nicht ausreichend berücksichtigen. Eine Unterscheidung der Maßnahmen nach ihrer Dauer ist unzureichend. Dieses verdeutlichen die Verteilungen der Laufzeiten der verschiedenen Maßnahmen in Abbildung 3, die sich zum Teil stark überlappen. Die Differenzierung nach Teilpopulationen ist zwar wünschenswert, aber auch diese Ergebnisse sind wenig aussagekräftig, wenn nicht nach Maßnahmearten unterschieden wird.

Obwohl Evaluationsstudien auf Basis von Daten des AMOs (AMSAs) oder dem GSOEPs nicht zur Herleitung von wirtschaftspolitischen Handlungsempfehlungen geeignet sind, waren die unternommenen Anstrengungen nicht vergebens. Ihr wissenschaftlich wertvoller Beitrag ist in der breiten Methodendiskussion zu sehen. In den letzten 15 Jahren kam es zu bemerkenswerten methodischen Fortschritten, die es ohne die breite Diskussion nicht gegeben hätte (Fitzenberger/Speckesser 2002: 23; Lechner 2002c: 22).

Der Schwerpunkt der aktuellen Evaluationsanstrengung liegt auf der Aufarbeitung der Datenbasis (Lechner 2002c: 19)[35]. Die neuen Datensätze entstehen im Wesentlichen durch die Verknüpfung bereits existierender Datensätze. Dabei leistet die Bundesagentur für Arbeit durch die Aufarbeitung ihrer Daten, um sie der wissenschaftlichen Forschung bereitzustellen, einen wichtigen Beitrag.

Im Herbst 2000 hat das Institut für Arbeitsmarkt- und Berufsforschung (IAB) in Kooperation mit Lehrstühlen der Universität Mannheim (Professor Fitzenberger) und St. Gallen (Professor Lechner) ein Projekt zur Evaluation der FbW mittels prozessproduzierter Daten begonnen.

In einer ersten Phase wurde ein neuer Datensatz, bestehend aus der IAB-Beschäftigungsstichprobe (IABS) [36] von 1980-1997, der BA-Statistik über die Teilnahme

[35] Bender, Fitzenberger und Lechner (2002: 12) heben jedoch hervor, dass mittelfristig die Erstellung eines verwertbaren Datensatzes nicht durch Wissenschaftler von Universitäten erfolgen sollte. Sie sollten derartige Arbeiten nur beratend begleiten.

[36] Eine detaillierte Beschreibung des IABS findet sich bei Bender et al. (1996).

an FbW-Maßnahmen seit 1980 (training participant data: TPD)[37] und der Leistungsempfängerdatei (LED), generiert. Anhand dieses neu gewonnen Datensatzes kann die Heterogenität der Maßnahmen, Teilnehmer und Regionen angemessen berücksichtigt werden. Die Fusion der unterschiedlichen Datensätze war mit erheblichen Problemen behaftet und konnte erst im Jahr 2002 abgeschlossen werden (Bender/Fitzenberger/ Lechner 2002:12).
Der Schwerpunkt der zweiten Phase liegt auf der detaillierten Evaluation des Maßnahmeerfolgs und wird im Jahr 2005 abgeschlossen sein[38].

Zum Zeitpunkt dieser Arbeit lag bereits ein Zwischenbericht zur Evaluation langfristiger Effekte öffentlich finanzierter Weiterbildungsmaßnahmen in Westdeutschland vor[39] (Lechner/Miquel/Wunsch 2004a). Entgegen den bisherigen Evaluationsergebnissen, die tendenziell einen negativen Effekt der FbW feststellen (Lechner 2002c: 21), kommt diese Studie zu dem Schluss, dass die FbW durchaus positive Beschäftigungseffekte haben kann. Für Westdeutschland zeigen die Autoren, dass sich die verschiedenen Fortbildungsarten in ihren Beschäftigungseffekten deutlich unterscheiden. Die Heterogenität der geschätzten Effekte bekräftigt die Zweifel an der Aussagekraft früherer Evaluationsstudien.
Im folgenden Abschnitt wird die Arbeit von Lechner, Miquel und Wunsch (2004a) kurz vorgestellt. Aufgrund der bedingten Aussagekraft früherer Evaluationsstudien werden diese nicht ausführlich diskutiert. Es befindet sich aber im Anhang 3 ein Überblick über den Großteil der Studien mit Beschreibung des verwendeten Datensatzes und der Evaluationsmethode sowie den Ergebnissen.

IV.2 Evaluationsstudie von Lechner, Miquel und Wunsch (2004a)

Diese Studie evaluiert den langfristigen Effekt der Teilnahme einer staatlichen Fortbildung zwischen den Jahren 1993 und 1994 auf die Beschäftigung und das Einkommen. Der neu generierte Datensatz (Abschnitt IV.1) erlaubt eine nach Maßnahmeart differenzierte Evaluation. Es wird zwischen Umschulung, Kurz- und

[37] Für TPD siehe Miquel, Wunsch und Lechner (2002).
[38] Die Arbeit von Klose und Bender (2000) ist hinsichtlich der verwendeten Datenbasis mit dieser Studie vergleichbar. Als Datenbasis verwendeten sie ebenfalls die IABS, die um Zeiten der Teilnahme an Maßnahmen der FbW ergänzt worden ist. Eine Differenzierung nach Maßnahmearten wird jedoch nicht vorgenommen und somit wird lediglich der Effekt der Teilnahme an irgendeiner Fortbildungsmaßnahme untersucht.
[39] Ein weiterer Bericht von Bender, Bergemann, Fitzenberger, Lechner, Miquel, Speckesser und Wunsch („Die Wirksamkeit von FuU-Maßnahmen“) erscheint in Kürze in “*Beiträge zur Arbeitsmarkt- und Berufsforschung*“.

Langmaßnahmen sowie praxisorientierten Weiterbildungen unterschieden[40]. Neben der Wirksamkeitsanalyse einzelner Maßnahmen im Vergleich zur Nichtteilnahme, erfolgt ein Vergleich der Wirksamkeit der Maßnahmen untereinander.
Die Autoren verwenden zur Evaluation einen Matching-Schätzer. Die Bildung der Kontrollgruppe vollzieht sich durch eine, dem Caliper-Verfahren ähnliche Vorgehensweise. Nach Ermittlung der Kontrollpersonen und den entsprechenden kontrafaktischen Ergebnissen, werden diese den Ergebnissen der Teilnehmer gegenübergestellt und somit die Effekte evaluiert.

IV.2.1 Der Matchingalgorithmus

In einem ersten Schritt werden durch die Schätzung eines *multinomial probit model* die propensity-scores aller Individuen für jede der vier Maßnahmearten geschätzt. Die Anzahl der schlechten Matches wird durch die Konzentration auf den *common support* verringert. Alle Teilnehmer und Nichtteilnehmer, deren propensity-score nicht in diesem Bereich liegen, werden aus der Stichprobe entfernt. Im nächsten Schritt wird jedem Teilnehmer ein Nichtteilnehmer zugeordnet, für den die Mahalanobis Distanz minimal ist. Zugeordnete Vergleichspersonen werden nicht aus dem Sample gestrichen, um möglicherweise als Matches für andere Teilnehmer herangezogen werden zu können. Ist jedem Teilnehmer einer Maßnahme eine Vergleichsperson zugeordnet worden, wird das Match ermittelt, für welches die Mahalanobis Distanz maximal ist. Auf Basis dieser Distanz, multipliziert mit einem bandwidth Parameter R, werden Intervalle um die propensity- scores gebildet. Jede Vergleichsperson, die in diesem Caliper einer geförderten Person liegt, wird dieser zugeordnet. Die Autoren setzten R = 0.9, dass heißt es werden nur Vergleichspersonen für ein Match zugelassen, deren Distanz zum betrachteten Teilnehmer nicht größer als 90% der Distanz des schlechtesten Matchs des eins-zu-eins Matching ist.

IV.2.2 Ermittlung der kausalen Effekte

Der kontrafaktische Wert der Ergebnisvariable jeder geförderten Person ermittelt sich, durch die gewichtete Summe der Ergebnisvariablen, der für das Match herangezogenen Vergleichspersonen. Die Gewichtung wird derart bestimmt, dass Vergleichspersonen, die dem Teilnehmer „näher“ sind, stärker gewichtet werden und die Summe der Gewichte Eins ergibt. Die Aufsummierung dieser individuellen

[40] Die Definitionen der Maßnahmearten können im Abschnitt II.2.3 nachgelesen werden.

kontrafaktischen Werte und anschließende Division durch die Anzahl der Teilnehmer der betrachteten Maßnahmeart führt zum durchschnittlichen Wert der Ergebnisvariable der Vergleichsgruppe.

Bevor dieses Ergebnis zur Ermittlung des kausalen Effekts herangezogen wird, wird dieser Wert noch um die Höhe des verbliebenen Selektionsniveaus korrigiert. Zu diesem Zweck führen die Autoren in der Stichprobe der Kontrollgruppe eine gewichtete lineare Regression der Variablen, die zur Definition der Distanz verwendet wurden, auf die Ergebnisvariablen durch. Auf Basis der geschätzten Koeffizienten werden die potentiellen Ergebnisse aller Personen aus der Teilnehmer- wie Nichtteilnehmergruppe geschätzt. Das Selektionsniveau ergibt sich durch die Differenz des durchschnittlichen Wertes der geschätzten Ergebnisvariable der geförderten Gruppe und des geschätzten kontrafaktischen Wertes.

Abschließend wird der kausale Effekt durch die Differenz des tatsächlichen durchschnittlichen Wertes der Ergebnisvariable der Teilnehmergruppe und dem gewichteten, um das Selektionsniveau korrigierten, Wert der Ergebnisvariable der Kontrollgruppe ermittelt.

IV.2.3 Ergebnisse der Wirkungsanalyse[41]

IV.2.3.1 Effekte der verschiedenen Maßnahmearten

Der Datensatz umfasst für die Jahre 1993 und 1994 insgesamt 38.810 Beobachtungen. Die Autoren lassen jedoch etliche Beobachtungen fallen, die bestimmte Kriterien wie zum Beispiel Alter oder Bezug von Unterhaltsgeld nicht erfüllen. Anhand Tabelle 4 kann der Selektionsprozess und die resultierende Anzahl der Beobachtungen nachvollzogen werden. Nach Anwendung des common support Kriterium besteht die endgültige Stichprobe aus 8367 Personen, von denen 6964 an keiner Maßnahme, 247 an einer praxisorientierten Weiterbildung, 503 an einer Kurzmaßnahme, 267 an einer langen Maßnahme und 386 Personen an einer Umschulung teilgenommen haben.

Die Anwendung des common support Kriterium führt dazu, dass 6,5 bis 19 Prozent der Beobachtungen der verschiedenen Maßnahmearten in der letzten Selektionsstufe fallengelassen werden. Dieses sollte bei der Interpretation der Ergebnisse berücksichtigt

[41] Im Rahmen dieser Arbeit ist eine umfassende Erörterung der methodischen Ergebnisse nicht möglich, da alleine der Appendix zur Darstellung der analytischen Ergebnissen der angewendeten Verfahren 56 Seiten umfasst. Diese Ergebnisse, die nicht der Studie Lechner/Miquel/Wunsch 2004a angehängt sind, sind separat im Internet zugänglich (Lechner/Miquel/Wunsch 2004b).

Tabelle 4: Stichprobenselektion

	NONPARTI-CIPATION	PRACTICE FIRM	SHORT TRAINING	LONG TRAINING	RETRAINING
Persons entering unemployment between Jan.´93 and Dec.´94					
Observations:	36965	324	644	380	497
Simulated programme start after the entry in unemployment and before the end of the observation period					
Remaining observations:	26022	324	644	380	497
Eligibility: Only individuals receiving unemployment benefits or assistance in the month of and before the programme start					
Remaining observations:	13091	309	618	350	450
Personal characteristics: a) 19<age<56; b) no trainees or apprentices; c) at least one observation of employment; d) no home workers; e) no part-time worker less than half of a full-time work					
Remaining observations:	9197	273	572	329	413
Imposition of the common support criterion					
Final Sample:	6964	247	503	267	386
Note: All variables are measured before or in the same year as the start of the programme					

Quelle: Lechner, Miquel und Wunsch (2004a: 18 und 2004b: 5)

werden. Es ist zu überdenken, ob die gemessenen Effekte repräsentativ für die Gesamtgruppe sind.

In der Tabelle 5 sind zehn paarweise Vergleiche aufgeführt. Die jeweils betrachteten Gruppen werden mit m und l bezeichnet. In Spalte 3 und 4 sind die Stichprobengrößen der jeweils betrachteten Gruppen abgetragen. In der Spalte 5 sind die tatsächlichen durchschnittlichen Anteile der Beschäftigten der Gruppen m aufgeführt. Die Beschäftigungsanteile beziehen sich auf zwei beziehungsweise sieben Jahre nach Maßnahmebeginn. Die korrespondierenden kontrafaktischen Ergebnisse sind in Spalte 6 abgebildet. Aus diesen Werten ergeben sich unmittelbar die geschätzten mittleren Effekte der Teilnahme an der Maßnahme m verglichen zur jeweiligen Referenzgruppe l (Spalte 7). Es zeigt sich, dass der ATET für die meisten Programme im Vergleich zur Nichtteilnahme positiv ist (wenn auch nicht immer signifikant). Lediglich die Teilnahme an einer Umschulung hat nach zwei Jahren einen negativen Effekt, der aber nach fünf weiteren Jahren in einen signifikant positiven Effekt umschlägt. Der kurzfristige negative Effekt wird auf einen *lock-in* Effekt zurückgeführt. Demnach verringern Teilnehmer von Maßnahmen während einer Weiterbildung ihre Suchintensität, weil sie einerseits weniger Zeit haben und andererseits die Maßnahme beenden wollen, um etwa im Falle der Umschulung einen beruflichen Abschluss zu erwerben. Die paarweisen Vergleiche zwischen Umschulung und den drei übrigen Maßnahmearten zeigen, dass die Teilnahme an einer Umschulung langfristig praxisorientierte Weiterbildungen, Kurzmaßnahmen und lange Maßnahmen dominieren. Die Effekte sind jedoch für die zwei letztgenannten Gruppen lediglich auf einem 10 % Level signifikant.

Der direkte Vergleich von Kurzmaßnahmen und langen Maßnahmen führt zu dem Ergebnis, dass keine der beiden eine dominierende Weiterbildungsart ist. Der zusätzliche

Tabelle 5: Geschätzte Beschäftigungseffekte zwei und sieben Jahre nach Maßnahmebeginn

Outcome	Month after beginning (t)	Sample m	size l	$E(Y^m \mid D = m)$	$\hat{E}(Y^l \mid D = m)$	$\hat{\theta}_t^{m,l}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pracice Firm (m) compared to nonparticipation (l)						
Employed	24	246	6910	42.3	38.3	4.0
	84	242	6772	48.8	44.5	4.3
Pracice Firm (m) compared to short training (l)						
Employed	24	246	501	42.3	50.6	*-8.3*
	84	242	494	48.8	51.5	-2.7
Pracice Firm (m) compared to long training (l)						
Employed	24	246	267	42.3	44.0	-1.7
	84	242	263	48.8	45.8	3.0
Pracice Firm (m) compared to retraining (l)						
Employed	24	246	386	42.3	32.1	**10.2**
	84	242	381	48.8	62.5	**-13.7***
Short training (m) compared to nonparticipation (l)						
Employed	24	501	6910	54.1	36.5	**17.9***
	84	494	6772	54.7	46.8	**7.9**
Short training (m) compared to long training (l)						
Employed	24	501	267	54.1	50.2	3.9
	84	494	263	54.7	46.4	*8.3*
Short training (m) compared to retraining (l)						
Employed	24	501	386	54.1	32.5	**21.6***
	84	494	381	54.7	62.5	*-7.8*
Long training (m) compared to nonparticipation (l)						
Employed	24	267	6910	49.4	41.1	**8.3**
	84	263	6772	54.4	48.6	5.8
Long training (m) compared to retraining (l)						
Employed	24	267	386	49.4	29.8	**19.6***
	84	263	381	54.4	64.2	*-9.8*
Retraining (m) compared to nonparticipation (l)						
Employed	24	386	6910	35.0	43.2	**-8.2**
	84	381	6772	62.7	47.0	**15.7***

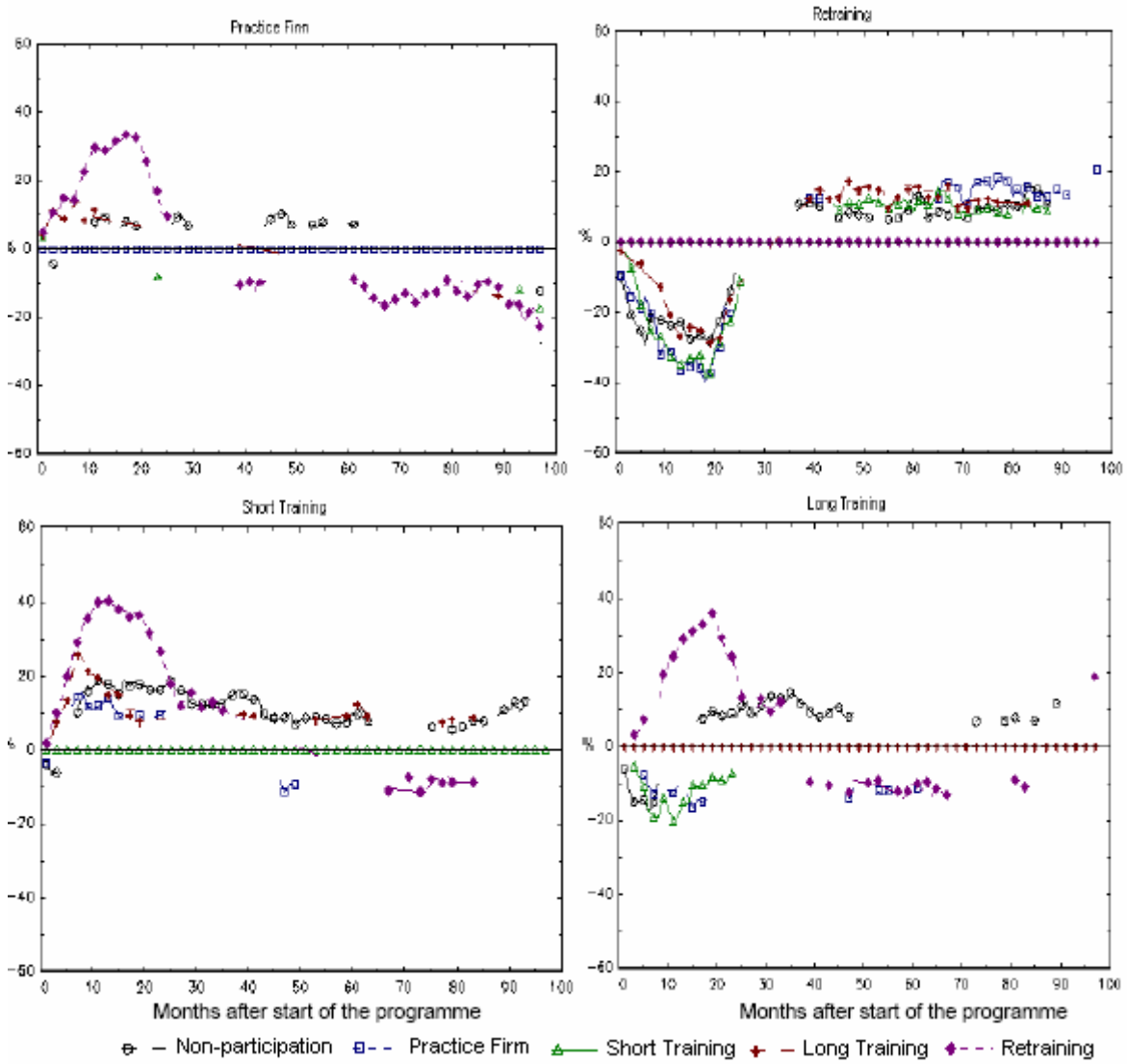
Note: **Bold** numbers indicate significance at the 5% level, numbers in *italics* relate to the 10% level and * to the 1% level.

Quelle: Lechner, Miquel und Wunsch (2004a: 32)

Mehrkosten- und Zeitaufwand für längere Maßnahmen scheint aus dieser Betrachtung nicht gerechtfertigt.

Abbildung 6 zeigt den zeitlichen Verlauf der geschätzten Effekte der verschiedenen Maßnahmearten auf die Beschäftigung der jeweiligen Teilnehmer, verglichen zur Referenzgruppe. Eine Linie über Null bedeutet, dass die als Referenzgruppe festgelegte Maßnahmeart diejenige, die mit der Linie dargestellt wird, dominiert. Die Autoren haben nur Effekte abgebildet, die auf einem 5 % Level signifikant sind.

Abbildung 6: Zeitliche Entwicklung des Effekts $\hat{\theta}_t^{m,l}$: Beschäftigungsunterschiede in %-Punkten



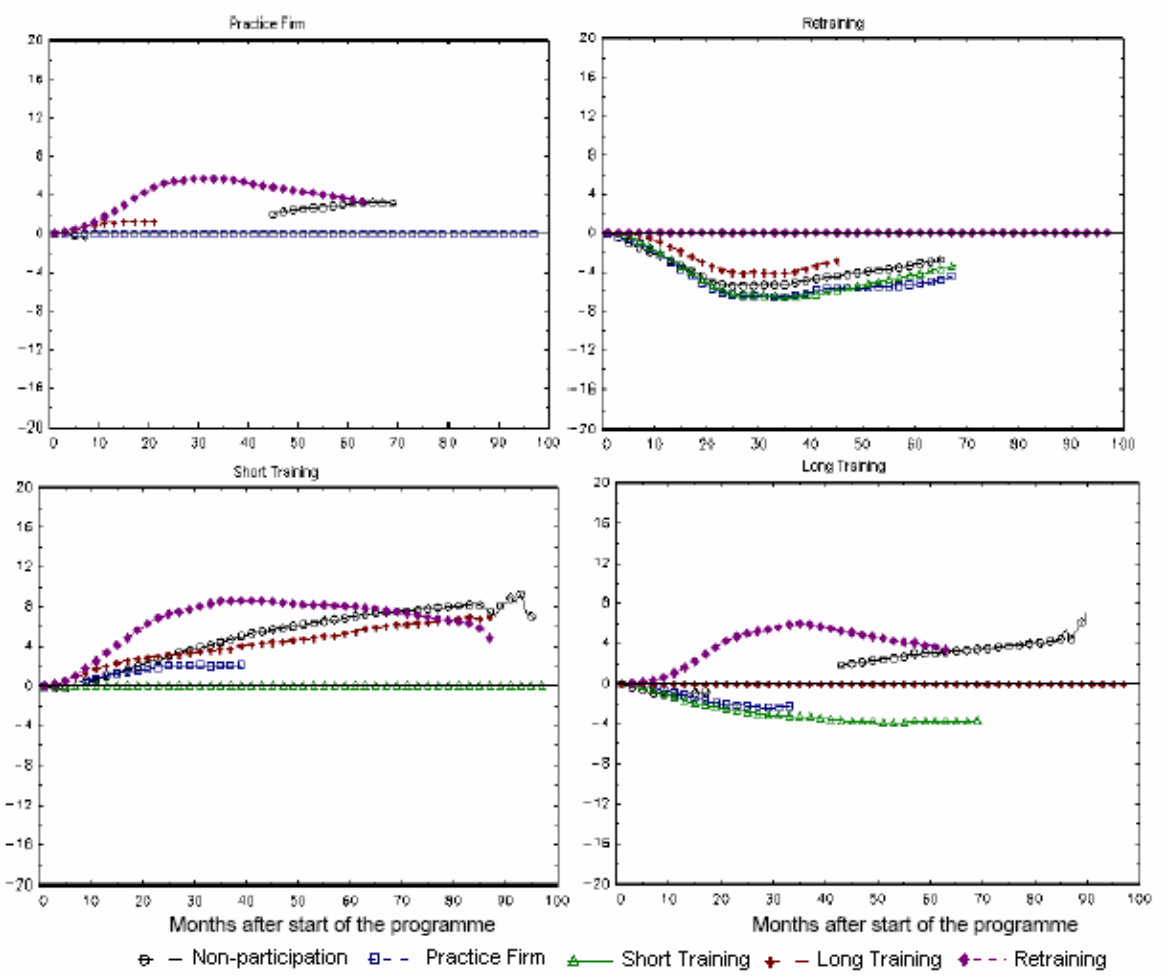
Quelle: Lechner, Miquel und Wunsch (2004a: 34). Es sind nur Effekte eingetragen, die auf einem 5% Level signifikant sind.

Für alle Maßnahmen zeigt sich, verglichen zur Gruppe der Nichtteilnehmer, ein unterschiedlich langer lock-in Effekt. Für Umschulungen beträgt dieser zwei bis drei Jahre, für lange Maßnahmen sechs bis zwölf Monate und für Kurzmaßnahmen und praxisorientierte Weiterbildungen drei bis sechs Monate. Langfristig ergeben sich für fast alle Maßnahmearten (gegenüber den Nichtteilnehmern) positive Effekte. Für Teilnehmer an einer Umschulung verbessern sich nach sieben Jahren die Beschäftigungswahrscheinlichkeiten um 10-15 %. Kurzmaßnahmen und lange Maßnahmen verzeichnen einen geringeren „Gewinn“ von 5 - 9 %. Die Teilnahme an einer praxisorientierten Weiterbildung hingegen hat keine beziehungsweise nur geringe positive Effekte.

Auch in dieser Darstellung zeigen sich langfristig keine Wirksamkeitsunterschiede zwischen lange Maßnahmen und Kurzmaßnahmen. Lediglich in den ersten zwei Jahren dominieren die Kurzmaßnahmen die langen Maßnahmen.
Umschulungen werden in den ersten zwei bis drei Jahren von allen anderen Weiterbildungsarten dominiert. Danach dominiert der Effekt einer Teilnahme an einer Umschulung alle übrigen Vergleichsgruppen.

Um die Wirksamkeit der Maßnahmearten endgültig beurteilen zu können ist es notwendig, die gegensätzlichen langfristigen Effekten den kurzfristigen Effekten gegenüberzustellen und somit den Nettoeffekt zu ermitteln. Hierzu dient Abbildung 7, die die kumulierten Beschäftigungseffekte der jeweiligen Maßnahmearten zur Referenzgruppe

Abbildung 7: Kumulierte Beschäftigungseffekte $\hat{\theta}_t^{m,l} = \sum_{\tau=1}^{t} \hat{\theta}_\tau^{m,l}$ in Monaten



Quelle: Lechner, Miquel und Wunsch (2004a: 37). Es sind nur Effekte eingetragen, die auf einem 5% Level signifikant sind. Eingetragenen Punkte sind wie folgt zu verstehen: „Im Durchschnitt verlängerte/verkürzte sich die Gesamtzeit in Beschäftigung für Teilnehmer m verglichen zu l zum Zeitpunkt t um $\hat{\theta}_\tau^{m,l}$ Monate.

abbildet. Der jeweilige Gesamteffekt kann anhand dieser Abbildung zu jedem Zeitpunkt abgelesen werden. Nach kurzfristigen negativen Gesamteffekten weisen alle Maßnahmearten verglichen mit der Gruppe der Nichtteilnehmer, mit Ausnahme der Umschulungen, langfristig positive Gesamteffekte auf. Die langfristig positiven Effekte der Umschulung reichen nicht aus, die lock-in Effekte zu kompensieren. Ein Vergleich der Umschulungen zu den drei anderen Fortbildungsarten führt zu dem gleichen Ergebnis.
Kurzmaßnahmen sind in dieser Darstellung die attraktivste Fortbildungsart. Diese Maßnahmen haben aufgrund ihrer kurzen Dauer nur geringe lock-in Effekte und die positiven Effekte akkumulieren sich über einen längeren Zeitraum. Sie dominieren in dieser Darstellung eindeutig die langen Maßnahmen. Somit festigen sich die Ergebnisse auf Basis der Tabelle 5, dass der Mehraufwand letztgenannter Fortbildungsart nicht gerechtfertigt ist.

IV.2.3.2 Heterogene Beschäftigungseffekte für verschiedene Teilpopulationen

Neben der Analyse der Beschäftigungseffekte der einzelnen Maßnahmearten untersuchen Lechner, Miquel und Wunsch die Wirkungen der verschiedenen Maßnahmearten auf verschiedene Teilpopulationen (2004a: 38 ff.). Verglichen zur Nichtteilnahme sind alle Fortbildungsarten wirkungsvoller in Regionen mit Arbeitslosenquoten unter 10 Prozent und für Personen, die weniger als ein Jahr arbeitslos sind. Geschlechtsspezifische Unterschiede ergeben sich nur für praxisorientierte Weiterbildungen, die für Männer ineffektiv und für Frauen äußerst wirkungsvoll sind. Dieses führen die Autoren aber darauf zurück, dass sich diese Fortbildungsart für Frauen und Männer unterscheiden und die Analyse eher heterogene Effekte zweier verschiedener Maßnahmen festgestellt hat. Hinsichtlich der Qualifikation und des gelernten Berufs der Fortbildungsteilnehmer ergeben sich keine signifikant unterschiedlichen Maßnahmewirkungen. Kurzmaßnahmen und lange Maßnahmen haben deutlich positive und Umschulungen negative Effekte in kleinen Städten (unter 100 Tausend Einwohner). In Großstädten verschlechtern alle Maßnahmen die Beschäftigungswahrscheinlichkeiten. Im Fall von Umschulungen ist dieser Effekt deutlich stärker als in Kleinstädten.

Es ist zu bedenken, dass die beschriebenen Effekte nur für einen bestimmten Zeitpunkt nach der Maßnahme geschätzt worden sind und somit fehlt die Kenntnis der kumulierten Effekte. Daher reicht diese Analyse nicht aus, abschließende Empfehlungen zu formulieren.

Trotzdem wird aus dieser Untersuchung die Notwendigkeit ersichtlicht, die Wirkung der FbW nicht nur hinsichtlich verschiedener Maßnahmearten sondern auch für verschiedene Populationen zu evaluieren. Aggregierte Wirkungsanalysen für Personengruppen, die sich in sozio-ökonomischen Charakteristika unterscheiden, reichen nicht aus, um die tatsächlichen Effekte zu evaluieren.

IV.2.4 Kritik

Die Evaluationsergebnisse sind unbestritten aussagekräftiger als die Ergebnisse früherer Studien. Sie erlauben ein genaueres Verständnis der Wirkung der FbW. Wurde bisher die Effektivität der staatlichen geförderten beruflichen Weiterbildung stark angezweifelt, ist nun offensichtlich, dass eine pauschale Evaluierung der FbW zu falschen Schlussfolgerungen führt. Diese geschätzten Ergebnisse gelten nur für Westdeutschland und es ist abzuwarten, zu welchem Schluss die Autoren im Abschluss des beschriebenen Projekts für Ostdeutschland kommen werden.

Bei der Interpretation der Ergebnisse ist jedoch zu bedenken, dass aufgrund des Selektionsprozesses, beschrieben im Abschnitt IV.2.3.1, bis zu 20 Prozent der Fortbildungsteilnehmer fallengelassen wurden. Ob die geschätzten Effekte gegenüber einer Evaluierung auf Basis der gesamten Stichprobe robust sind, bleibt offen.
Die Konzentration auf den common support hat jedoch den Vorteil, dass die verfahrensspezifische Selektionsverzerrung (B_1 und B_2) eliminiert wird. In der weiteren Analyse versuchen die Autoren die Verzerrung aufgrund unbeobachtbare Variable zu korrigieren. Ob die angewendete Methode (Abschnitt IV.2.2) dazu ausreicht, ist unklar.

V. Einfluss von Motivation auf Dauer der Arbeitslosigkeit

V.1 Intention der Untersuchung

Aus den in Abschnitt III beschriebenen Methoden haben sich in der neueren Evaluationsforschung überwiegend Matchingschätzer oder Methoden durchgesetzt, die Matching als Basis der weiteren Untersuchung verwenden, wie der konditionale DvD-Schätzer oder die Verweildaueranalyse.
Wie bereits in Abschnitt III.3.2.7 beschrieben, produzieren Matchingschätzer und die Verweildaueranalyse auf Basis einer gematchten Kontrollgruppe inkonsistente Schätzergebnisse, wenn eine Selektionsverzerrung hinsichtlich unbeobachtbarer Variablen vorliegt. Obwohl der konditionale DvD-Schätzer die Selektion hinsichtlich zeitkonsistenter Variablen eliminiert, sind auch dieser Methode Grenzen gesetzt, wenn das

Endogenitätsproblem auf zeitinkonsistenten, unbeobachtbaren Variablen zurückzuführen ist.
Die Qualität von Evaluationsergebnissen ist danach zu beurteilen, inwieweit es gelungen ist, die Selektionsverzerrung zu berücksichtigen. Da in Deutschland keine sozialen Experimente durchgeführt werden, fehlt die Grundlage dieses zu beurteilen.
Es ist zu vermuten, dass große Datensätze wie das GSOEP oder auch der neue Datensatz des IABs nicht alle Variablen beinhalteen, die auf die Zielvariablen und gleichzeitig auf die Partizipationswahrscheinlichkeit wirken. Solche Variablen könnten zum Beispiel das Vermögen einer Person, die Erwerbstätigkeit des Partners oder die eigene Motivation sein.

Die folgende Untersuchung greift dieses Problem auf und analysiert den Einfluss von Motivation auf die Beschäftigungswahrscheinlichkeit. Es ist keine Evaluationsstudie, die den Effekt der Teilnahme an einer Fortbildung zu bestimmen versucht, da der verwendete Datensatz nur Personen umfasst, die an einer Umschulung teilgenommen haben. Dennoch stellt diese Studie einen Beitrag zur Evaluationsforschung dar. Kann im Rahmen der Untersuchung gezeigt werden, dass Motivation einen Einfluss auf die Beschäftigungswahrscheinlichkeit hat, liegt die Vermutung nahe, dass Motivation auch die Partizipationswahrscheinlichkeit beeinflusst. In diesem Fall müsste ein Matching zur Bildung einer Kontrollgruppe Motivation als entscheidungsrelevantes Merkmal berücksichtigen. Andernfalls würde der Selektionsverzerrung nicht ausreichend Rechnung getragen werden[42].

Bevor der Datensatz, der mikroökonometrische Analyserahmen und die Ergebnisse dargestellt werden, ist es notwendig, Motivation genau zu definieren und die angewandten Messmethoden vorzustellen, mit denen diese erfasst wurde. Daher werden im folgenden Abschnitt zwei verschiedene Konstrukte der Motivation aus der Psychologie beschrieben und mögliche Messansätze vorgestellt.

[42] Die Studien von Hujer, Maurer und Wellner (1997a) sowie Hujer und Wellner (2000) berücksichtigen als erklärende Variable unter anderem einen Zufriedenheitsindex, der Werte zwischen Null (absolut unzufrieden) und Zehn (absolut zufrieden) annehmen kann. Die Autoren sehen darin eine Indikatorvariable für die Motivation. Ohne weitere Untersuchungen erscheint es jedoch fraglich, ob dieser Zufriedenheitsindex ein geeignetes Instrument zur Erfassung der Motivation ist.

V.2 Instrumentalisierung des Merkmals Motivation

V.2.1 Persönlichkeitseigenschaften als Quelle der Motivation

Neben Bedürfnissen, Interessen und Handlungsüberzeugungen werden in der Psychologie Motive als elementare Persönlichkeitseigenschaften gesehen, aus denen sich Motivation bildet (vgl. Asendorpf 2004: 211).

Aus dem Bedürfniskonzept lässt sich eine motivierende Wirkung durch physiologische Ungleichgewichte, dem Abweichen von einem aktuellen Ist-Zustand vom Sollwert, ableiten. Die Bedürfnis-Pyramide von Maslow (1954) war der letzte Versuch, alle Aspekte der Motivation auf Grundbedürfnisse wie physiologische (Hunger, Durst), Sicherheits-, soziale Bindungs- und Selbstverwirklichungsbedürfnisse zurückzuführen (Asendorpf 2004: 211).

Mittlerweile wird dem Modell von Maslow nur noch ein mäßiger Erklärungsgehalt zugesprochen und die Motivationsforschung konzentrierte sich vermehrt auf Interessen, resultierend aus der Bewertung von Handlungen, Motiven, die sich auf die Bewertung von Handlungsfolgen beziehen und Handlungsüberzeugungen. Unter letztgenannten werden Persönlichkeitsmerkmale wie Souveränität (zögerliches versus zielgerichtetes Handeln) oder Optimismus beziehungsweise Pessimismus bei der Bewertung der Erfolgserwartung von Handlungen verstanden.

Alle vier Persönlichkeitseigenschaften tragen unterschiedlich stark zur Motivation während der Arbeitssuche bei. Während das Arbeitslosengeld oder die Arbeitslosenhilfe noch die rudimentären physiologischen Bedürfnisse befriedigen, müssen die Betroffenen (sofern keine andere Einkommensquellen existieren) Einschnitte bei der Erfüllung von Bedürfnissen sozialer Bindungen oder Selbstverwirklichungsbedürfnissen hinnehmen. Dieses ist nicht nur finanziell sondern auch im Status der Arbeitslosigkeit begründet. So wird der Arbeit in Organisationen zwei Merkmale zugeordnet. Zum einem nötigen existenzielle Gründe Individuen zur Arbeitsverrichtung und zur Akzeptanz der damit verbundenen Mühen und Lasten. Auf der anderen Seite kann nach Lewin (1920: 11) Arbeit aber auch Züge der Schaffensfreude aufweisen und dem Leben Sinn und Gewicht geben. Somit werden Individuen, die in der Aufgabenerfüllung ihres Berufsfeldes eine interessante Tätigkeit sehen oder sich gar mit einem Beruf identifizieren, motivierter um eine neue Stelle bemühen. Gehen die Betroffenen zudem optimistisch und souverän an die Stellensuche, dürfte die Wahrscheinlichkeit steigen, einen Arbeitsplatz zu finden.

Dieses sind nur einige Komponenten der Wirkungszusammenhänge von persönlichen Eigenschaften und der Motivation im Hinblick auf die Arbeitssuche.

Eine Fragebogenerhebung, die allen genannten Aspekten des globalen Konzeptes der Motivation Rechnung trägt, wäre für den Rahmen einer Diplomarbeit zu aufwendig. Daher soll das Untersuchungsfeld auf einen Aspekt des Motivationskonzepts eingeschränkt werden, das hinsichtlich des Analyserahmens als geeignet erscheint.
Als geeignetes Motivationskonzept kommt das Konstrukt der Arbeitsmotivation oder der Leistungsmotivation in Frage.

V.2.2 Das Konstrukt der Arbeitsmotivation

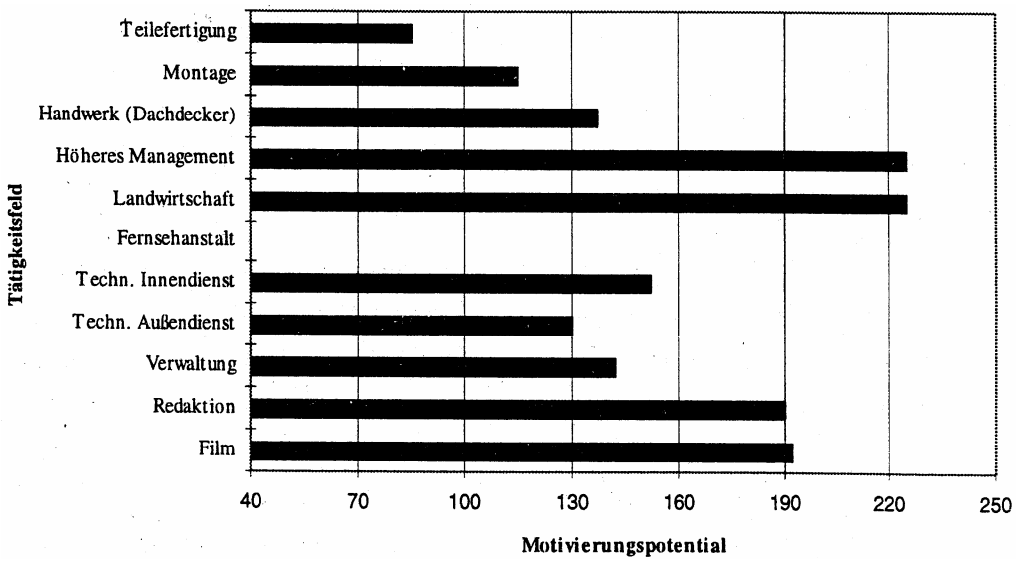
„Unter Arbeitsmotivation wird [...] die Bereitschaft verstanden, Fähigkeiten und Fertigkeiten zum Zweck produktiver und zielorientierter Arbeit einzusetzen“ (Frieling/Sonntag 1999: 150). Sie ist neben den Fähigkeiten und Fertigkeiten eines Arbeiters das zweite Element der Leistungsvoraussetzung und trägt wesentlich zur Produktivität bei (vgl. Kleinbeck 1996: 14). Die Arbeitsmotivation wird zum einem durch den Anregungsgehalt der Arbeitssituation und zum anderen durch persönliche Motive bestimmt. Zur Bestimmung des Anregungsgehalt der Arbeitssituation wird in der Literatur der „Motivationspotential der Arbeitssituation“-Index[43] (MPA) benutzt (Schmidt et al. 1985). Demnach ist eine Tätigkeit mit hohem Motivationspotential dadurch charakterisiert, dass sie dem Arbeiter Freiräume bei der Erfüllung seiner Aufgaben lässt und der Arbeiter eigenverantwortlich wichtige Tätigkeiten mit angemessenen Schwierigkeitsgrad verrichten kann.
In Abbildung 8 sind die MPA-Indices für neun unterschiedliche Tätigkeitsfelder aufgeführt, die Schmidt und Kleinbeck (1985) durch Befragung von 640 Personen ermittelten. Demnach weisen die aufgeführten Handwerks- und Fertigungssektoren niedrigere Motivationspotentiale auf als verantwortungsvollere und kreative Tätigkeiten (höheres Management, Redaktion, Film).
Es ist jedoch zu beachten, dass der MPA-Index lediglich leistungsthematische Motivationspotentiale berücksichtigt. Neben den leistungsthematischen Motivationspotentialen werden auch noch anschlussthematische (diese beziehen sich auf soziale Kontakte zu den Mitarbeitern) oder machthematische (Mitbestimmungsgehalt) Motivationspotentiale (Kleinbeck 1996: 33-37) diskutiert. Würde man diese Faktoren zusätzlich

[43] $$MPA = \left[\frac{\text{Anforderungs-vielfalt} + \text{Identifikation} + \text{Wichtigkeit der Aufgabe}}{3} \right] + \text{Autonomie des Aufgabenhandelns} + \text{Rückmeldung}$$

Zur ausführlichen Definition siehe Kleinbeck et al. (1980).

berücksichtigen, könnte ein modifizierter MPA-Index für einen kontaktfreudigen Fließbandarbeiter höher als der ursprüngliche MPA-Index ausfallen, wenn in seiner Arbeitsgruppe ein gutes Mitarbeiterklima herrscht.

Abbildung 8: Motivationspotentiale verschiedener Tätigkeitsfelder



Quelle: Kleinbeck (1996: 34)

Um die Arbeitsmotivation zu bestimmen, ist neben der Ermittlung des subjektiv wahrgenommenen MPAs auch die Kenntnis über persönliche Motive nötig. Bislang gibt es kein Testverfahren, welches beide Faktoren erhebt und miteinander vereinigt. Hätte man im Rahmen dieser Arbeit noch die Motivationspotentiale der Arbeitssituationen[44] bestimmen können, würde es doch an wissenschaftlichen Instrumenten zur persönlichen Motivmessung fehlen.

V.2.3 Leistungsmotivation

V.2.3.1 Das Konstrukt der Leistungsmotivation

Ein weiteres Motivationskonstrukt ist das Leistungsmotivationskonzept. Eingangs wurden Motive als Persönlichkeitseigenschaft charakterisiert.[45] Leistungsmotivation ist neben den kognitiven Fähigkeiten wichtiger Bestandteil des berufsbezogenen Leistungserfolgs (Schulter/Prochaska 2000: 61).

[44] Auch wenn die Befragten die Ausbildung in einer außerbetrieblichen Einrichtung absolviert haben, entsprachen die praktischen Inhalte der Umschulung den Anforderungen einer betrieblichen Ausbildung. Anhand dieser praktischen Erfahrungen könnten Teilnehmer von Umschulungen das Motivationspotential der Arbeitssituation durchaus bestimmen.

[45] Die Literatur nimmt häufig Bezug auf die Motivtheorie von McClelland, die Motive in Macht-, Anschluss- und Leistungsmotiv zergliedert.

Die Konstrukte der Arbeits- und Leistungsmotivation sind sehr verwandt, nehmen sie doch den gleichen Stellenwert bei der Determinierung von Produktivität ein. Setzt sich Arbeitsmotivation aus dem Motivierungspotential der Arbeit und persönlichen Motiven zusammen, so formulierte Atkinson ein einflussreiches Modell, nach dem Leistungsmotivation sich aus einem Erfolgs- bezeiehungsweise Misserfolgsmotiv als persönliche (situationsunabhängige) Eigenschaft sowie einer situationsabhängigen Erfolgserwartung zusammensetzt (Asendorpf 2004: 213). Überwiegen Erfolgsmotive als Charaktereigenschaft, spricht man von erfolgsmotivierten Personen. Die Leistungsmotivation ist umso höher, je mehr das individuelle Erfolgsmotiv das Misserfolgsmotiv übersteigt und je näher die Erfolgswahrscheinlichkeit (eine Aufgabe bewältigen zu können) am maximal motivierenden Wert 0,5 liegt (Atkinson 1957: 365) [46].

V.2.3.2 Methoden zur Messung der Leistungsmotivation

Das Leistungsmotiv als Bedürfnis, Leistung zu steigern oder zumindest hoch zu halten, wurde lange Zeit ausschließlich mit dem thematischen Apperzeptionstest (TAT) von Heckhausen (1963) gemessen. Es handelt sich hierbei um ein projektives Verfahren, bei dem den Probanden mehrdeutige Bilder vorgelegt werden. Die Bildbeschreibungen der Versuchspersonen werden hinsichtlich des Vorkommens bestimmter Themen überprüft. Von diesen Themen soll auf Persönlichkeitseigenschaften geschlossen werden.
Da dieses Verfahren sehr aufwendig ist und außerdem Zweifel an der Validität projektiver Verfahren insgesamt bestehen (Asendorpf 2004: 214ff.), besann man sich auf die Konzeption von Fragebögen. 1968 gelang es Hermans einen Fragebogen auszuarbeiten, der den Anforderungen an Validität, Objektivität und Reliabilität genügte. Auf eine neuere Fassung des Leistungsmotivationstest (Hermans/Petermann/Zielinski 1978) begründet sich auch der Fragebogen, der innerhalb der vorliegenden Arbeit verwendet wurde[47].
Der Leistungsmotivationstest (LMT) beinhaltet unter anderem zwei Dimensionen zur Messung von Ausdauer/Stress sowie Leistungsstreben. Der Fragebogen enthält 15 Fragen zur Erfassung des Leistungsstrebens und 13 Fragen zur Dimension Ausdauer. Die Auswertung des Tests erfolgt durch Schablonen, auf denen die Antwortmöglichkeiten, die auf hohes Leistungsstreben und hohe Belastbarkeit hindeuten, gekennzeichnet sind.

[46] Atkinson beruft sich auf empirische Befunde, die darauf hinweisen, dass die Höhe der Motivation als eine Funktion der Erfolgswahrscheinlichkeit dargestellt werden kann. Die Funktion hat einen U-förmigen Verlauf mit anfänglich positiver Steigung und einem Schwellenwert bei 0,5.
[47] Neben dem Leistungsmotivationstest existiert zwar noch ein neuerer Test aus dem Jahr 2000 von Schuler und Prochaska, das Leistungsmotivationsinventar (LMI), dieser wird zur Zeit aber nicht im Testarchiv der psychologischen Fakultät der Universität Konstanz geführt und ist nur kommerziell erwerblich.

Nachdem der Proband den Fragebogen bearbeitet hat, werden die Anzahl der Fragen addiert, deren Antworten gemäß des Schablonenvergleichs auf eine hohe Ausprägung der jeweiligen Dimension hinweisen.
Der Test wurde anhand einer Stichprobe von 587 Schülern der Berufs-, Berufsfach- und Realschule sowie Gymnasien geeicht[48]. Diese Normierung der Rohwerte ist für die weitere Untersuchung nicht dienlich, da sich die Eichstichprobe auf eine abweichende Population bezieht. Hinsichtlich des Ziels dieser Arbeit – die Ermittlung einer Kennzahl für Motivation – interessiert allerdings auch weniger der Vergleich zwischen verschiedenen Gruppen sondern vielmehr der Vergleich innerhalb der Gruppe selbst.

V.2.3.3 Modifikation des Leistungsmotivationstests

Im Rahmen der vorliegenden Untersuchung ergab sich das Problem, dass der LMT sehr umfangreich ist (28 Fragen zur Ermittlung einer Kennziffer für Motivation). Da der zeitliche Rahmen für diese Arbeit beschränkt war, galt es sich zwischen einer ausführlichen Umfrage mit einer kleinen Stichprobe und einer kurzen Umfrage mit großer Stichprobe zu entscheiden. Aufgrund der hohen Anzahl der Kovariablen, die in die Schätzung miteinfließen, ist ein großer Stichprobenumfang wünschenswert, um interpretierbare Ergebnisse zu erzielen. Daher galt es, den LMT zu verkürzen. Dieses ist nicht unproblematisch, da ein niedriger Halbierungskoeffizient[49] von 0,67 für die Dimension Leistungsstreben und 0,62 für die Demission Ausdauer und Fleiß auf eine Heterogenität der Fragen hinweist (vgl. Hermans/Petermann/Zielinski 1978: 16). Es ist also fraglich, ob alle Fragen konsistent in Bezug auf das zu erfassende Merkmal und somit austauschbar sind. Da der Halbierungskoeffizient nicht geeignet ist, die Reliabilität zu messen, wenn ein heterogener Test vorliegt, führten Hermans, Petermann und Zielinski (1978: 16) einen Retest[50] an einer Stichprobe von 23 Psychologie-Studenten[51] durch. Für diese Gruppe ergab sich ein Retestkoeffizient[52] von 0,8 für Leistungsstreben und 0,74 für Ausdauer und Fleiß. Ausgehend von diesem hohen Niveau der Reliabilität, scheint eine

[48] Eichungen werden vorgenommen, um Gruppen-Normen zu bestimmen. So ist der Vergleich individueller Testergebnisse mit einem Referenzwert der entsprechenden Gruppe möglich (Lienert 1989: 315).
[49] Der Halbierungskoeffizient ist ein Maß zur Messung der Reliabilität. Eine umfassende Erörterung von Konsistentkoeffizienten findet sich bei Lienert (1989: 225 ff.).
[50] Um die Zuverlässigkeit/Messgenauigkeit eines Test einzuschätzen, wiederholen die Versuchspersonen nach einer gewissen Zeit denselben Test. Anhand des Vergleichs der Testergebnisse wird die Konsistenz ermittelt. Wichtig Voraussetzung für diese Methode der Messung der Reliabilität, ist die zeitliche Stabilität der gemessenen Merkmale bei den Probanden.
[51] Psychologie Studenten und Teilnehmer an staatlichen Qualifizierungsmaßnahmen sind äußerst heterogene Gruppen. Gemäß Rost (2004: 356) werden die Maße der Reliabilität und der Validität von Tests meist an Stichprobengruppen erhoben, denen die getesteten Personen gar nicht zugehören.
[52] Korrelationskoeffizient von Spearman (Lienert 1989: Formel 53).

Verkürzung des Tests möglich. Dennoch besteht ein Trade-off zwischen einer Verkürzung des Fragebogens und der Reliabilität.
Die Selektion der Fragen erfolgte aufgrund einer inhaltlichen Strukturierung des LMTs. Insgesamt waren vier verschiedene Fragefelder zu unterscheiden, wie zum Beispiel Fragen zu zukunftsorientiertem Denken oder Fleiß in der Ausbildung. Der verwendete Fragebogen ist so aufgebaut, dass, entsprechend der ursprünglichen Relation, Fragen aus jedem Bereich berücksichtigt sind. Der Fragebogen ist im Anhang 4 beigefügt.
Um die Zuverlässigkeit dieses Tests zu messen und die Ergebnisse somit wissenschaftlich abzusichern, wäre es nötig, einen Retest durchzuführen. Dieses ist aber mangels Zeit und verfügbarer Ressourcen im Rahmen dieser Arbeit nicht möglich.

V.3 Empirische Untersuchung

V.3.1 Beschreibung des Datensatzes

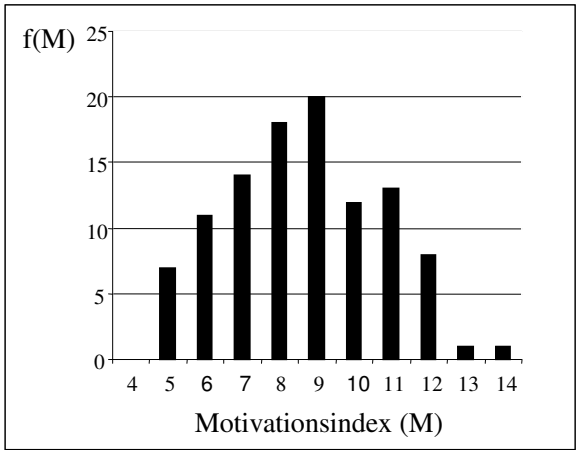
Für die empirische Untersuchung wurden Daten aus der Teilnehmdatenbank des privaten Bildungsträgers „Internationaler Bund“ verwendet. Die Datenbank umfasst neben demografischen Merkmalen der Teilnehmer Informationen über die Art der Fortbildung, an der sie teilgenommen haben, in welchem Zeitraum diese stattgefunden hat und durch wenn diese finanziert wurde.

Zur Analyse wurden nur Personen herangezogen, die im Raum Stuttgart an einer staatlich finanzierten Umschulung teilgenommen haben und diese zwischen November und Januar der Jahre 2003 und 2004 beenden konnten. Dadurch wird eine Homogenität der Teilnehmer hinsichtlich regionaler und konjunktureller Einflüsse auf die Ergebnisvariable erzielt. Hinsichtlich der unterschiedlichen Abschlussjahre wird angenommen, dass die Arbeitsmarktlage in den Wintermonaten beider Jahre nicht signifikant verschieden war und dadurch keine zusätzliche Heterogenität erzeugt wird. Von den verbleibenden 219 Personen waren 105 Personen telefonisch zu erreichen und bereit, den Fragen des beschriebenen Instruments zur Ermittlung der Motivation zu beantworten.
Anhand Abbildung 9 wird deutlich, dass der Motivationsindex des überwiegenden Teils der Befragten einen Wert um 9 hat. Das Minimum liegt bei 5 und ein Individuum weist den höchstmöglichen Wert des Motivationskoeffizienten von 14 auf.
Weiterhin wurde erhoben, ob die geförderten Personen innerhalb der letzten zehn Monate nach Ende der jeweiligen Maßnahme eine Arbeit aufnehmen konnten und wenn ja, nach wie vielen Monaten dieses geschah. Die empirische Untersuchung basiert auf einem

Hazardraten-Modell und daher ist die Kenntnis der Dauer der Arbeitslosigkeit nach Ende der Umschulung erforderlich.

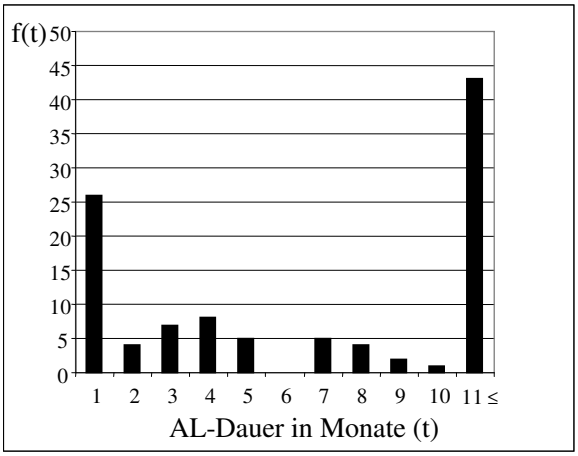
In Grafik 10 ist die Häufigkeitsverteilung der Arbeitslosigkeitsdauer abgebildet. Ein Großteil der Personen konnte innerhalb des ersten Monats nach Ende der Umschulung eine Arbeit aufnehmen. Dieses kann mit darauf zurückgeführt werden, dass die Maßnahmeteilnehmer in den letzten Monaten der Umschulung ein betriebliches Praktikum absolvieren und einige von ihnen übernommen werden. Zeitlich betrachtet, nehmen die Arbeitsaufnahmen nach diesem anfänglichen Mitnahmeeffekt rapide ab und in den folgen-

Abbildung 9: Häufigkeitsverteilung des Motivationsindexes



Quelle: Eigene Berechnungen

Abbildung 10: Häufigkeitsverteilung der AL-Dauer nach Maßnahmeende



Quelle: Eigene Berechnungen

den neun Monaten können durchschnittlich nur vier Personen pro Monat in das Erwerbsleben eintreten. Von den 105 betrachteten Personen sind am Ende des Betrachtungszeitraums, zehn Monate nach Maßnahmeende, noch immer 43 Personen arbeitslos. Abschließend wurden die ehemaligen Teilnehmer befragt, wie viele Monate sie insgesamt in den letzten fünf Jahren vor Beginn der Umschulung arbeitslos waren und wie oft sie in dem gleichen Zeitraum den Arbeitgeber gewechselt haben.

V.3.2 Erklärende Variablen der Untersuchung

Im folgendem soll der Einfluss persönlicher Merkmale, insbesondere der Motivation, auf die Dauer der Arbeitslosigkeit nach Ende der Umschulung ermittelt werden. In die Untersuchung werden folgende Variablen mit aufgenommen:

- **Sozio-demografische und weitere persönliche Merkmale:** Geschlecht, Nationalität (Deutsch vs. Ausländer), Alter am Ende der Maßnahme (drei Kategorien: bis 30, 31 – 37 und über 37 Jahre), Heiratsstatus (verheiratet vs. nicht verheiratet), Motivation (auf

einer Skala von 1 (niedrig) bis 14 (hoch)), Schulabschluss (drei Kategorien: Abitur und Fachhochschulreife, mittlere Reife sowie Hauptschulabschluss) und die Note der Abschlussprüfung der Umschulung (auf einer Skala von 1 (sehr gut) bis 5 (mangelhaft)).

- **Merkmale der Erwerbsvorgeschichte:** Kumulierte Arbeitslosigkeitsdauer innerhalb der letzten fünf Jahre vor Maßnahmebeginn (in Monaten) und Anzahl der Arbeitsplatzwechsel innerhalb des gleichen Zeitraums (ein oder kein vs. mindestens zwei Arbeitsplatzwechsel).
- **Sektorale Ausrichtung der Fortbildung:** Ausbildungsbereich der Umschulung (fünf Kategorien: kaufmännischer Bereich, Gastronomie, Elektroniker und Mechaniker, technischer Zeichner, IT-Systemelektroniker und KFZ-Mechaniker)

Bei den sozio-demografischen Variablen wird erwartet, dass jüngere, deutsche, verheiratete und motivierte Männer am schnellsten in ein neues Beschäftigungsverhältnis einmünden. In Hinsicht auf die Abschlussnote der Umschulung sollte eine gute Note auf ein ausgeprägtes Engagement des Teilnehmers und gute fachlichen Kenntnisse schließen lassen. Daher wird hier ein positiver Effekt auf die Wahrscheinlichkeit, die Arbeitslosigkeit zu verlassen, erwartet. Außerdem sollten Personen mit einem höheren Schulabschluss schneller eine Beschäftigung aufnehmen können als andere. In Bezug auf die Erwerbsgeschichte wird von einem negativen Einfluss der kumulierten Arbeitslosigkeitsdauer ausgegangen. Hinsichtlich der Variable Arbeitsplatzwechsel sind zwei gegensätzliche Wirkungen denkbar. Einerseits zeugt eine hohe Fluktuation für Flexibilität des Arbeitnehmers. Andererseits könnte man aber auch vermuten, dass der Betroffene sprunghaft ist oder unzureichende Arbeitsleistungen erbracht hat, welches zur Aufhebung der verschiedenen Arbeitsverhältnisse führte.

V.3.3 Deskriptive Auswertung

Bevor mittels ereignisanalytischer Modelle die oben genannte Fragestellung untersucht wird, wird zunächst eine deskriptive Auswertung präsentiert. In den Abbildungen 11.1 bis 11.4 sind die Survivorfunktionen nach Geschlecht, Nationalität, Familienstatus und Motivation einander gegenübergestellt.

Ein deutlicher Unterschied in der Dauer der Arbeitslosigkeit besteht zwischen Deutschen und Ausländern (Abbildung 11.1). Nach vier Monaten konnte über die Hälfte der deutschen Umschüler eine Arbeit aufnehmen. In der Gruppe der Ausländer wurde dieses Ergebnis auch am Ende des Beobachtungszeitraums nicht erreicht.

Anhand der Abbildung 11.2 ist zu erkennen, dass die Survivorfunktion der verheirateten Personen stets unterhalb der Vergleichsgruppe verläuft. Die Wahrscheinlichkeiten der beiden Gruppen, noch arbeitslos zu sein, unterscheiden sich jedoch nicht so deutlich wie

Abbildung 11.1: Survivorfunktionen nach Nationalität getrennt

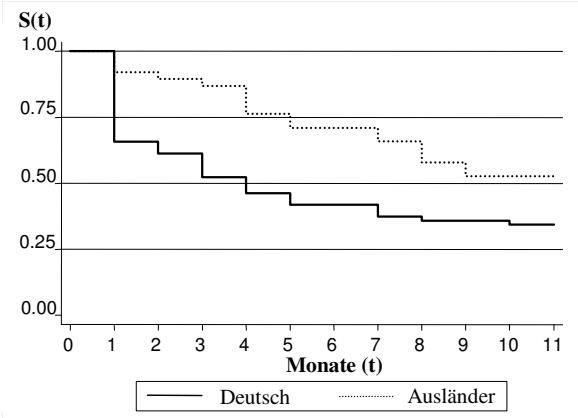


Abbildung 11.2: Survivorfunktionen nach Heiratsstatus getrennt

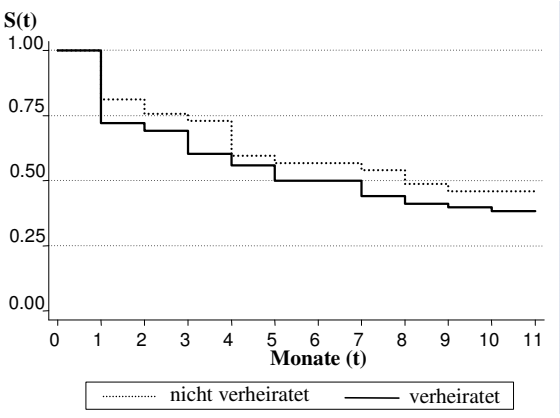


Abbildung 11.3: Survivorfunktionen nach Geschlecht getrennt

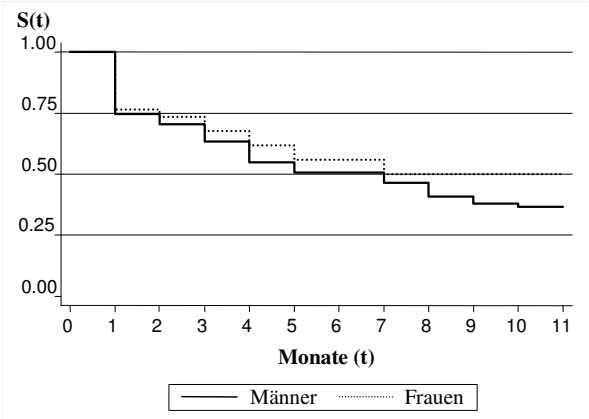
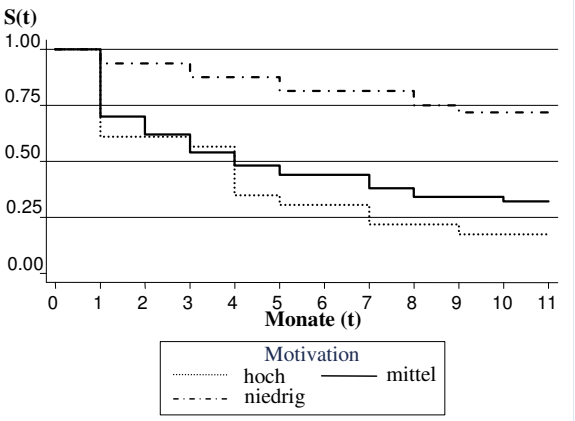


Abbildung 11.4: Survivorfunktionen nach Motivation getrennt



Quelle: Eigene Berechnung

bei Ausländern und Deutschen. Nach einem Monat konnten gegenüber 20 Prozent der Nicht-Verheirateten, cirka 30 Prozent der Verheirateten eine Arbeit aufnehmen. Nach zehn Monaten lag dieser Anteil der letztgenannten Gruppe bei etwa 60,5 Prozent und bei ungefähr 55,0 Prozent in der Gruppe der nicht verheirateten Personen.

Die Vermutung, dass Frauen schlechtere Aussichten auf dem Arbeitsmarkt haben als Männer, kann nur bedingt bestätigt werden (Abbildung 11.3). Bis zum achten Monat nach Maßnahmeende haben die Survivorfunktionen einen nahezu identischen Verlauf. Erst danach kristallisiert sich eine höhere Wiederbeschäftigungswahrscheinlichkeit für Männer heraus. Während am Ende des Betrachtungszeitraums etwa 64 Prozent der Männer einer regulären Beschäftigung nachgingen, vermochte dieses nur die Hälfte aller Frauen.

Um eine deskriptive Auswertung des Merkmals Motivation durchführen zu können, muss eine Gruppierung vorgenommen werden, da der Datensatz nicht ausreicht, für jede der 14 Ausprägungen des Indexes eine hinreichend große Gruppe zu bilden. Eine Unterteilung in drei Gruppen (wenig motiviert = 0 bis 7, mittlere Motivation = 8 bis 10, hoch motiviert = 11 bis 14) scheint sinnvoll und ausreichend, um die Ergebnisse inhaltlich zu interpretieren. Das Resultat ist in Abbildung 11.4 dargestellt. Die Vermutung, dass hoch motivierte Personen schneller Arbeit finden als niedrig motivierte Personen wird deutlich bestätigt. Der Unterschied zwischen den Gruppen der hoch und mittel motivierten Personen fällt zwar weniger stark aus, beträgt nach 10 Monaten aber immerhin noch etwa 13 Prozent. Von den gering motivierten Personen waren am Ende des Betrachtungszeitraums immer noch 72 Prozent arbeitslos (gegenüber 18 Prozent bei den hoch motivierten Personen). Auffällig ist, dass die Mitnahmeeffekte in der Gruppe der gering Motivierten nicht anfallen. Bei den hoch und mittel motivierten Personen sowie in den Gruppen der vorangegangenen Analyse waren diese Effekte, bis auf die Gruppe der Ausländer, stets sichtbar. Dieses deutet darauf hin, dass die weniger motivierten Umschüler die Möglichkeit, sich im Praktikumsbetrieb zu beweisen, nicht nutzen und einen mäßigen Eindruck beim Arbeitgeber hinterließen.

V.3.4 Präsentation der Schätzergebnisse

Die deskriptive Auswertung der Daten hat einen ersten Einblick in die Wirkungszusammenhänge der einzelnen persönlichen Merkmale und der Dauer der Arbeitslosigkeit gegeben. Diese sollen nun mittels der Verweildaueranalyse statistisch abgesichert werden.
Das stetige PH-Modell von Cox ist für die Untersuchung nicht geeignet, da die Zielvariable Arbeitslosigkeitsdauer in Form von Monatsintervallen vorliegt und mehrere Personen innerhalb eines Monats ausgeschieden sind. Weiterhin deckt der Datensatz nur einen kurzen Beobachtungszeitraum ab und somit ist die Anzahl der ausgeschiedenen Personen innerhalb eines Intervalls im Vergleich zur Risikogruppe relativ hoch. Daher wird für die Schätzung die Approximation des PH-Modells nach Efron verwendet.

In das Basismodell fließen als erklärende Variablen zunächst sozio-demografische und persönliche Merkmale ein. In drei weiteren Schätzungen werden sukzessiv die Kovariablen um die oben beschriebenen Variablen erweitert. In Tabelle 6 sind die Ergebnisse der Schätzung aufgeführt. Es ist zu beachten, dass nicht die geschätzten Para-

Tabelle 6: Geschätzte Hazardratenverhältnisse des PH-Modells

Modell	**1**	**2**	**3**	**4**
Alter [Über 37 Jahre]				
bis 30 Jahre	2.324***	2.3199***	2.1098**	2.2776**
	(0.7577)	(0.7542)	(0.7030)	(0.8036)
31 bis 37 Jahre	1.1786	1.2526	1.2523	1.3989
	(0.4094)	(0.4515)	(0.4678)	(0.5504)
Geschlecht – weiblich	0.8121	0.8175	0.9719	0.6806
	(0.2380)	(.3012)	(0.2985)	(0.5504)
Nationalität - nicht deutsch	0.5182*	0.4929**	0.5112*	0.4754**
	(0.1536)	(0.1464)	(0.1541)	(0.1491)
Verheiratet	1.8113*****	1.8145*	**1.7251**	1.8978*
	(0.5254)	(0.5256)	(0.5022)	(0.5833)
Abschlussnote der Umschulung	0.6572*	**0.6666**	**0.6965**	0.7208
	(0.1299)	(0.1389)	(0.1394)	(0.1503)
Motivation	1.2679***	1.2718***	1.2845***	1.3374***
	(0.0916)	(0.0923)	(0.0957)	(0.1090)
Schulbildung [Hauptschulabschluss]				
Abitur/ Fachhochschulreife		1.0870	1.1514	1.1427
		(0.4535)	(0.4860)	(0.4892)
mittlere Reife		0.6796	0.7437	0.7999
		(0.2201)	(0.2468)	(0.2746)
Merkmale zur Erwerbsvorgeschichte				
kum. AL-Dauer in Monaten			0.9801	**0.9745**
			(0.1224)	(0.0134)
mehr als 2 Arbeitgeberwechsel			1.4718	1.3926
			(0.4531)	(1.3926)
Ausbildungsbereich [KFZ]				
kaufmännischer Bereich				1.6751
				(1.0273)
Gastronomie				1.6384
				(1.0601)
Mechaniker/Elektroniker				1.4860
				(0.7195)
Technischer Zeichner				4.8521*
				(3.4785)
IT-Systemelektroniker				0.9462
				(.0.5861)
Liklihood-Ratio- χ^2	40.18	42.02	47.24	53.55
Prob > χ^2	0.0000	0.0000	0.0000	0.0000

Anmerkung: Standardabweichung in runden Klammern. Referenzgruppe in eckigen Klammern. **Fett** unterlegte Zahlen deuten Signifikanz auf 10 % Level an, * deuten 5% Level, ** 2,5% Level und *** 1% Level an.

Quelle: Eigene Berechnung.

meter selbst, sondern die Verhältnisse der Hazardraten aufgeführt sind, die sich gemäß Gleichung 44 unmittelbar aus den Koeffizienten errechnen. Ein Wert über 1 deutet somit einen positiven Einfluss der betrachteten Variablen auf die Hazardrate an und Werte unter 1 einen negativen Einfluss (siehe Erläuterung im Abschnitt III.3.4.2).

Die Schätzung des Basismodells (Modell 1) zeigt die erwarteten Zusammenhänge der Merkmale auf die Arbeitslosigkeitsdauer. Als wichtigstes Ergebnis ist festzuhalten, dass - unter Kontrolle aller anderen Merkmale - der Effekt der Motivation auf einem 1%-Level signifikant ist und den erwarteten Einfluss hat. Hoch motivierte Personen haben gegenüber vergleichbaren Personen, die wenig motiviert sind, höhere Hazardraten und können somit

die Arbeitslosigkeit in der Regel schneller beenden[53].

Die Hazardraten von Personen unter 30 Jahren betragen das 2,3-fache der Hazardraten der über 37-Jährigen. Dieser Zusammenhang ist auf einem 1% Level signifikant. Während zwischen Personen mittleren Alters (31 – 37 Jahre) und der Referenzgruppe sowie zwischen Männern und Frauen keine statistisch abgesicherte Unterschiede in der Dauer der Arbeitslosigkeit festzustellen sind, kann die Vermutung im Hinblick auf die Nationalität auf einem 5 % Signifikanzlevel bestätigt werden. Die Wahrscheinlichkeit, eine Arbeit aufzunehmen, ist zu jedem Zeitpunkt für Ausländer um etwa 48 Prozent geringer als für Deutsche. Verheiratete und Personen mit guten Abschlussnoten kehren signifikant schneller in die Beschäftigung zurück. Bei letzterer Variable ist zu beachten, dass die Ausprägung Eins für sehr gute Leistung und Fünf für mangelhafte Leistung steht. Hat eine Person mit der Note Vier abgeschlossen, ist ihre Hazardrate um etwa 35 Prozent niedriger als die jener Personen, die mit der Note Drei abgeschlossen haben, und um 57 Prozent niedriger als die jener Personen, die eine Zwei erzielten.

In das Modell 2 wird zu den bisher betrachteten Variablen die Schulbildung mit aufgenommen. Wie erwartet verlassen Teilnehmer der Umschulung mit Abitur oder Fachhochschulreife die Arbeitslosigkeit eher als Personen mit einem Hauptschulabschluss. Dieser Effekt ist jedoch nicht signifikant.

Auffallend ist die schlechtere Ausgangssituation für Teilnehmer mit mittlerer Reife. Da dieser Effekt nicht signifikant von Null verschieden ist und sich durch die Hinzunahme weiterer Variablen in den Modellen 3 und 4 abschwächt, wird an dieser Stelle auf einen spekulativen Erklärungsansatz verzichtet.

---

[53] Das Modell interpretiert die Ausprägungen der Motivation als kardinalskaliert. Demnach ist der Unterschied in den Hazardraten zwischen Personen mit einem Motivationsindex von 6 und 9 genauso hoch wie der Unterschied zwischen Personen mit den Ausprägungen 10 und 13.
Von einer Interpretation der Koeffizienten wird an dieser Stelle jedoch abgesehen. Die Art der Motivationsmessung gibt keinen Aufschluss über die Skalierung der Werte und es ist fraglich, ob anzunehmen ist, dass der Unterschied in der Motivation zwischen Personen mit Werten von 6 und 9 genauso hoch ist wie für Personen mit einem Motivationsindex von 10 und 13. Da eine objektive Messung der Unterschiede in der Motivation mit großen Schwierigkeiten behaftet, wenn nicht sogar unmöglich ist, kann nicht per se angenommen werden, dass das Merkmal kardinalskaliert ist. Eine ordinale Skalierung scheint hier naheliegender. Daher kann man nur Aussagen darüber treffen, ob eine Person mehr oder weniger motiviert ist als eine andere. Von quantitativen Aussagen ist daher abzusehen.

Im Modell 3 wird zusätzlich die Anzahl der Arbeitgeber und die kumulierte Dauer der Arbeitslosigkeit berücksichtigt. Personen, die in den letzten fünf Jahren vor Umschulungsbeginn lange arbeitslos waren, haben größere Schwierigkeiten, wiederbeschäftigt zu werden als andere.
Personen, die in der Vergangenheit häufiger den Arbeitgeber gewechselt haben, können schneller eine Arbeit aufnehmen. Wie bereits erwähnt, könnte dieses daran liegen, dass Arbeitgeber dieses als Indikator der Flexibilität der potentiellen Arbeitnehmer interpretieren. Der Effekt erweist sich jedoch als nicht signifikant, so dass ein solcher Erklärungsansatz nur mit Vorsicht zu genießen ist.

Im Modell 4 werden schließlich die Ausbildungsbereiche der Umschulung berücksichtigt. Eine Umschulung zum technischen Zeichner erhöht im Vergleich zu Umschülern im KFZ-Bereich beträchtlich die Hazardrate. Angesichts des geringen Stichprobenumfangs und der Unterscheidung der Ausbildungsbereich-Variable in sechs Kategorien ist es jedoch durchaus möglich, dass dieses Ergebnis durch Zufall entstanden ist. Der Maximum-Likelihood-Schätzer hat zwar die wünschenswerten Eigenschaften der Konsistenz und Erwartungstreue, doch gelten diese nur asymptotisch.

Aus der Betrachtung der vier verschiedenen Modelle ist ersichtlich, dass die qualitativen Ergebnisse der geschätzten Effekte der sozio-demografischen und persönlichen Merkmale gegenüber der Modellspezifikation robust sind. Demnach ist es statistisch abgesichert, dass höher motivierte, deutsche, verheiratete Personen unter 30 Jahren schneller wiederbeschäftigt werden als andere. Der Schulabschluss und das Geschlecht sowie die Anzahl der Arbeitsplatzwechsel haben keinen signifikanten Einfluss auf die Ergebnisvariable. Der signifikant positive Effekt einer guten Abschlussnote im Modell 1 verliert sich durch die Hinzunahme weiterer Variablen. Werden alle Variablen berücksichtigt, kann man hier nicht mehr von einem signifikanten Einfluss der Abschlussnote sprechen.

Die Güte der Modelle wird durch einen Likelihood-Ratio-Test ermittelt. Anhand dieses Tests wir die Hypothese überprüft, dass die Koeffizienten in der Grundgesamtheit gleich Null sind[54]. Die Ergebnisse des Tests können den beiden letzten Spalten entnommen werden. Für jedes Modell wird die Null-Hypothese verworfen. Wie zu

[54] Die Prüfgröße des Likelihood-Ratio-Test ergibt sich aus der Differenz der logarithmierten Wahrscheinlichkeit (die Likelihood), dass alle Koeffizienten gleich Null sind und der logarithmierten Wahrscheinlichkeit, dass die Koeffizienten die geschätzten Werte annehmen. Multipliziert man die Differenz mit (-2) erhält man die Likelihood-Ratio-Statistik, die χ^2-verteilt ist.

erwarten ist, steigt der Likelihood-Ratio Wert und somit die Güte mit der Hinzunahme von Variablen und ist im Modell 4 am höchsten.
Anders als in einem Regressionsmodell kann anhand der Schätzung aber nicht ermittelt werden, wie hoch der Erklärungsanteil des Modells ist. Abgeleitet werden kann lediglich eine Verbesserung des Modells mit steigendem Likelihood-Ratio Wert.

V.3.5 Sensibilitätsanalyse

V.3.5.1 Test der PH-Annahme

Trotz der großen Einsatzbreite ist das PH-Modell von Cox an die Annahme der proportionalen Hazardraten gebunden. Das Verhältnis der Hazardraten zweier Individuen ist im PH-Modell gemäß Gleichung 43:

$$\frac{h(t|X_i)}{h(t|X_j)} = \exp\{(X_i - X_j)'\beta\}, \qquad i \neq j. \tag{43}$$

Eine Verletzung der Annahme würde implizieren, dass der Koeffizient einer Variablen X zeitabhängig ist:

$$\beta_x(t) = \beta_x + q_i g(t), \tag{50}$$

wobei β_x der konstante Teil des Parameters ist, q_i ein Koeffizient und g(t) irgendeine Funktion der Zeit ist. Unter den Ahnnahmen des Modells von Cox müsste $q_i = 0$ sein.

Zur Überprüfung der PH-Annahme schlagen Grambsch und Therneau (1994) vor, Schönfeldresiduen (r) zu betrachten (Schönfeld 1982). Diese Residuen ergeben sich aus der Differenz der Ausprägung einer Kovariablen eines Individuums i zum Ausscheidungszeitpunkt und dem erwarteten Wert (der Durchschnitt der jeweiligen Risikogruppe) der Variablen zu diesem Zeitpunkt. Somit ergeben sich für jedes Merkmal und für jeden Zeitpunkt so viele Residuen wie Personen ausscheiden. Grambsch und Therneau (1994) führten eine Methode der Skalierung der Schönfeld-Residuen ein, so dass

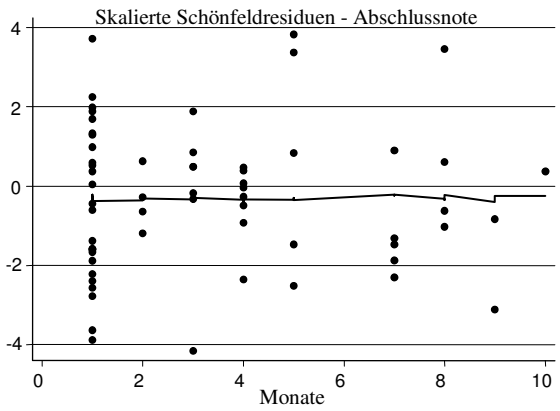
$$E(r^*_{x,i} + \beta_x) = \beta_x(t) \tag{51}$$

gilt. Folglich haben die skalierten Residuen eines Merkmals unter der PH-Annahme die Eigenschaft, dass sie zu jedem Zeitpunkt den gleichen Erwartungswert haben. Das heißt, dass für jede Kovariable des Modells ein Scatterplot der Residuen zu den jeweiligen Austrittszeitpunkten keinen Trend aufweist.

Neben der grafischen Analyse kann die Untersuchung auf $H_0: q_i = 0$ statistisch abgesichert werden. Unter der Nullhypothese der PH-Annahme sollte ein Graf, der die durchschnittlichen Ausprägungen der Residuen verbindet, keine Steigung aufweisen.

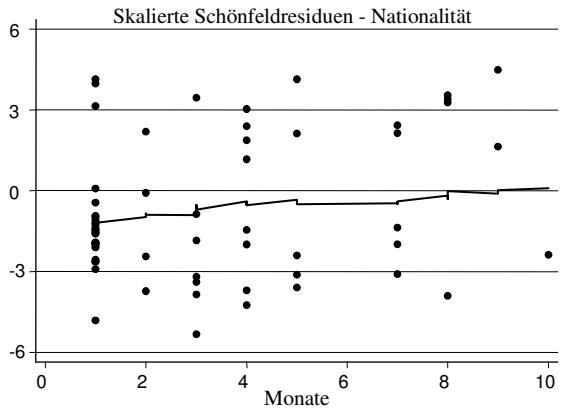
Abbildung (12.1) und (12.2) präsentieren die grafische Analyse der PH-Annahme des Modells 4 in Bezug auf die Merkmale „Note der Abschlussprüfung“ und „Nationalität“. Die Abbildungen der übrigen Merkmale sind im Anhang 6 aufgeführt.
Die skalierten Schönfeldresiduen sind in Abbildung 12.1 zufällig um Null verteilt. Der eingezeichnete Graf weist keine Steigung auf und verdeutlicht daher, dass es keinen zeit-

Abb. 12.1: Test der PH-Annahme für das Merkmal Abschlussnote



Quelle: Eigene Berechnung

Abb.12.2: Test der PH-Annahme für das Merkmal Nationalität



Quelle: Eigene Berechnung

abhängigen Trend gibt. Abbildung 12.2 lässt jedoch vermuten, dass q in Bezug auf die Nationalität von Null verschieden ist und somit die Nullhypothese verworfen werden muss. Das geschätzte Verhältnis der Hazardraten zwischen Ausländern und Deutschen, das zu Beginn des Prozesses den Wert 0,4754 hat, ist über die Verweildauer hinweg nicht konstant und verbessert sich aus Sicht der Ausländer.
Mit Ausnahme des Merkmals „kaufmännischer Ausbildungsbereich“ lassen die Abbildungen 5 im Anhang 5 vermuten, dass die Proportionalitätsannahme ansonsten nicht verletzt ist.

Diese Ergebnisse werden durch den statistischen Test auf $H_0 : q_i = 0$ bestätigt (Tabelle 7.1). Während der *globale Test* statistisch nicht signifikant ist, zeigen die Tests nach den einzelnen Merkmalen, dass die PH-Annahme für das Merkmal „Nationalität“ und „kaufmännischer Bereich“ auf einem Signifikanzlevel von 5 % verworfen werden muss.
Wenn die Annnahme über den proportionalen Verlauf der Hazardraten für ein Merkmal verletzt ist, kann diese nicht als Kovariable sondern muss als Schichtungsvariable in das Cox-Modell aufgenommen werden (Blossfeld/Hamerle/Mayer 1986: 58). In diesem Fall wird die Annahme der Existenz einer einzigen Baseline-Hazardrate für die gesamte

Tabelle 7.1: Test der PH-Annahme

	rho	χ^2	Df	Prob> χ^2
bis 30 Jahre	0.0443	0.14	1	0.7095
31 bis 37 Jahre	0.0202	0.03	1	0.8680
Geschlecht - weiblich	0.0426	0.11	1	0.7382
Nationalität - nicht deutsch	0.2694	4.99	1	0.0254
Verheiratet	-0.1468	1.57	1	0.2104
Abschlussnote der Umschulung	0.0104	0.01	1	0.9282
Motivation	0.0394	0.13	1	0.7171
Abitur/ Fachhochschulreife	0.1242	0.95	1	0.3307
Mittlere Reife	0.1730	2.14	1	0.1430
kum. AL-Dauer in Monaten	-0.0923	0.61	1	0.4356
mehr als 2 Arbeitgeberwechsel	0.0571	0.23	1	0.6304
kaufmännischer Bereich	-0.2725	4.25	1	0.0393
Gastronomie	-0.0047	0.00	1	0.9686
Mechaniker/Elektroniker	-0.1056	0.73	1	0.3919
Technischer Zeichner	-0.0821	0.49	1	0.4860
IT-Systemelektroniker	-0.0945	0.65	1	0.4210
global test		17.55	16	0.3508

Quelle: Eigene Berechung

Population gelockert und angenommen, dass es schichtspezifische Baseline-Hazardraten gibt. Gibt es r = 1,…, R Schichten kann Gleichung 42 zum Beispiel wie folgt erweitert werden:

$$h_r(t|X_i) = h_{0r}(t)\exp\{X_i'\beta\}, \tag{51}$$

wobei das Modell weiterhin davon ausgeht, dass die geschätzten Koeffizienten für alle Individuen identisch sind. Darüber hinaus verändern sich die individuellen Beiträge zurLikelihoodfunktion. Wurden diese in Gleichung 47 durch das Verhältnis der Hazardrate einer Person i und der kumulierten Hazardrate der Risikogruppe gebildet, so werden bei der Stratifikation nur noch jene Personen der Risikogruppe im Nenner berücksichtigt, die zur Schicht von i gehören (vgl. Singer/Willett 2003: 558)[55].

Im folgendem wird am PH-Modell von Cox festgehalten und die PH-Annahme anhand des Modells 4 unter Berücksichtigung der Schichten Nationalität und kaufmännischer Bereich geprüft.

Der statistische Test weist nun darauf hin, dass die Annahme nicht verworfen werden kann (Tabelle 7.2). Der *global test* weist eine deutlich höhere Fehlerwahrscheinlichkeit

[55] Die beschrieben Änderungen der individuellen Beiträge zur Likelihoodfunktion lassen sich auf die Gruppenbetrachtung übertragen.

auf als der erste Test. Während die Proportionalitäts-Annahme in Bezug auf das Merkmal „mittlere Reife“ eine Fehlerwahrscheinlichkeit auf einem 5 %-Level zeigt, besteht für die

Tabelle 7.2: Test der proportionalen Hazardannahme bei Schichtung der Merkmale Nationalität und kaufmännischer Bereich

	rho	χ^2	df	Prob> χ^2
bis 30 Jahre	0.0714	0.39	1	0.5339
31 bis 37 Jahre	0.0452	0.15	1	0.7033
Geschlecht - weiblich	0.0516	0.16	1	0.6849
Verheiratet	-0.1368	1.58	1	0.2087
Abschlussnote der Umschulung	0.0001	0.00	1	0.9997
Motivation	0.0588	0.30	1	0.5839
Abitur/ Fachhochschulreife	0.0476	0.14	1	0.7074
mittlere Reife	0.2100	3.10	1	0.0785
kum. AL-Dauer in Monaten	-0.1077	0.82	1	0.3640
mehr als 2 Arbeitgeberwechsel	0.0630	0.27	1	0.6053
Gastronomie	-0.0194	0.03	1	0.8660
Mechaniker/Elektroniker	-0.1082	0.80	1	0.3712
Technischer Zeichner	-0.0562	0.22	1	0.6358
IT-Systemelektroniker	-0.0992	0.72	1	0.3967
global test		7.71	14	0.9038

Quelle: Eigene Berechung

übrigen Merkmale kein Zweifel an der Richtigkeit der Annahme. Durch die Stratifizierung werden die Zusammenhänge zwischen den Hazardraten und den Kovariablen im PH-Modell besser getroffen.

Obschon sich der Erklärungsgehalt des Modells durch die Schichtung verringert - die Sichtungsvariablen gehen nicht mehr als erklärende Variablen in das Modell ein - ist die Stratifizierung unerlässlich, um nicht gegen die Annahmen des Modells zu verstoßen.

Daher wird nun das Modell 4 unter Berücksichtigung der Schichtungsvariablen „Nationalität“ und „kaufmännischer Bereich“ geschätzt. Anhand der Ergebnisse in Tabelle A.6.1 im Anhang 6 ist zu erkennen (Modell 5), dass sich die geschätzten Einflüsse gegenüber dem Modell 4 nicht wesentlich verändern.

Das Verhältnis der Hazardraten variiert nur geringfügig und mit Ausnahme des Merkmals „Abitur/Fachhochschulreife“ bleiben die qualitativen Einflüsse der Merkmale erhalten. Motivation hat weiterhin einen positiven Einfluss auf die Wahrscheinlichkeit der Wiederbeschäftigung, der statistisch auf einem 1 %-Niveau abgesichert ist. Lediglich das Signifikanzniveau des Einflusses der Variable „Heiratsstatus“ fällt auf ein 10 %-Level. Dafür steigt die statistische Sicherheit, dass der negative Einfluss der Dauer der Arbeitslosigkeit von Null verschieden ist.

Wie zu erwarten war, ist der Liklihood-Ratio Wert geringer als im Modell 4. Die Hypothese, dass alle Parameter nicht von Null verschieden sind, wird aber auch im Schichtungsmodell eindeutig verworfen.

V.3.5.2 Goodness-of-Fit

Der Likelihood-Ratio-Test aus Abschnitt V.3.4 reicht nicht aus, um den Erklärungsgehalt des Modells beurteilen zu können. Sobald einige Kovariablen signifikant sind, wird der Test, dass alle Koeffizienten gleich Null sind, verworfen. Wie gut die Hazardraten durch das Modell spezifiziert sind, ist jedoch nicht ersichtlich.

Grønnesby und Borgan führten 1996 einen Test auf Basis von martingale Residuen ein, mit dem die Güte von PH-Modellen eingeschätzt werden kann. Diese Residuen können als Differenz der tatsächlich beobachteten Austritte und der erwarteten Austritte zu jedem Zeitpunkt interpretiert werden. Die erwarteten Austritte ergeben sich aus den geschätzten Parametern des Modells und den Baseline-Hazardraten, die zusätzlich geschätzt werden müssen[56]. Der Grønnesby und Borgan Test sieht vor, dass die Beobachtung hinsichtlich ihrer „risk scores", $\hat{r} = X_i' \hat{\beta}$, in G Gruppen unterteilt werden[57]. Danach werden für jedes Individuum die martingale Residuen bestimmt und diese innerhalb der Gruppen aufsummiert. Wenn die Modellspezifikation die Hazardraten gut modelliert, sollten diese Summen gleich Null sein.

May und Hosmer (1998: 112) bezeichnen die Herleitung der Prüfgröße des Grønnesby und Borgan Tests als recht mühsam. Sie zeigten, dass der Test algebraisch identisch mit solchen ist, die dem PH-Modell Gruppenindikatorvariablen beifügen und mittels *score*-Tests die Hypothese testen, dass die Koeffizienten der Indikatorvariablen der Gruppen gleich Null sind[58]. Somit kann die Güte des Modells mit Softwareprogrammen wie STATA und SAS relativ einfach bestimmt werden. Darüber hinaus ist es möglich, anstatt

---

[56] Die Herleitung der erwarteten Austritte kann dem Appendix B der Arbeit von May und Hosmer (1998) entnommen werden.

[57] Hosmer und Lemeshow (1980, 1989) empfehlen die Bildung von 10 Gruppen, um die Güte von log-linearen Regressionen einzuschätzen. May und Hosmer (1998: 114) legen jedoch nahe, möglichst wenig Gruppen zu bilden. Ist eine asymptotische Annäherung der martingale Residuen innerhalb der Gruppen erwünscht, kann dieses bei kleinen Stichproben und einer großen Anzahl an Gruppen leicht ausbleiben. May und Hosmer (2004: 286f.) zeigten, dass auch für große Stichproben, die Höhe der GrØnnesby und Borgan Teststatistik inkorrekt spezifiziert wird, wenn die Anzahl der risk-score Gruppen zu hoch ist. Aus ihrer Untersuchung leiteten sie folgende Strategie zur Gruppenbildung ab (May/Hosmer 2004a: 288):
G = integer of (max(2, min(10, number of failures/40))).

[58] Werden zum Beispiel fünf Gruppen gebildet müssen, vier Dummy-Variablen generiert werden, die den Wert Eins annehmen, wenn der risk-score einer Person in dem Caliper der Gruppe liegt und Null, wenn diese Person nicht zu der Gruppe gehört. Eine Gruppe wird als Referenzgruppe festgelegt und daher geht diese Indikatorvariable nicht mit in das PH-Modell mit ein.

eines score Tests den asymptotisch äquivalenten Likelihood Ratio oder Wald Test zu verwenden.

Der Arbeit von May und Hosmer (2004b) folgend, wird der Erklärungsgehalt des vorliegenden Modells anhand eines Likelihood-Ratio-Tests überprüft. Als Modellspezifikation wird das Modell 5 (Anhang 6) verwendet, in das die Indikatorvariablen für die Gruppen mit eingehen. Als Referenzgruppe wird diejenige bestimmt, in der die Personen mit den niedrigsten risk-scores vertreten sind. Wird die

Tabelle 8: Goodness-of-Fit Test

i	2	3	5	8
LR χ^2 (i-1)	0.18	0.53	0.78	2.88
Prob > χ^2	0.6731	0.7672	0.9407	0.8957

Quelle: Eigene Berechnung

Strategie von May und Hosmer zur Gruppenbildung angewendet, ergibt sich bei 62 Austritten G = 2. Zusätzlich wird der Test auch für drei, fünf und acht Gruppen durchgeführt.

Die Ergebnisse sind in Tabelle 8 aufgeführt. Bei zwei Gruppen beträgt die Likelihood-Ratio-Statistik 0.18 und ein p-Wert von 0.67. Der Test kann die Hypothese, dass der Parameter der Indikatorvariablen gleich Null ist, nicht verwerfen. Zum gleichen und noch deutlicheren Ergebnis kommen die drei anderen Tests.

Das geschichtete PH-Modell kann die Hazardraten der vorliegenden Stichprobe sehr gut erklären. Die berücksichtigten Variablen und das Modell können als ausreichend erachtet werden, qualitative Unterschiede in der Arbeitslosigkeitsdauer zu erklären.

V.3.5.3 Analyse der Altersgruppenwahl

Ziel der Bildung unterschiedlicher Altersgruppen ist die Unterscheidung zwischen jungen und alten Fortbildungsteilnehmern sowie Personen mittleren Alters. Die genaue Abgrenzung basiert auf keinen ökonomischen Überlegungen und orientiert sich danach, möglichst gleichgroße Gruppen zu erhalten. Um die Schätzergebnisse abzusichern, werden zwei alternative Abgrenzungen der Altersgruppen vorgenommen und diese zur Schätzung des Modells 5 herangezogen.

Die Ergebnisse der Modelle 5´ und 5´´ (Tabelle A.6.1 im Anhang 6) zeigen, dass die zwei neuen Schätzungen im Vergleich zur Referenzschätzung keine qualitativen Unterschiede aufweisen. Die quantitativen Unterschiede der Koeffizienten der Altersgruppe sind

hingegen nicht vernachlässigbar. Demnach ist weiterhin von einem signifikant positiven Effekt der jüngsten Altersgruppe auszugehen, der jedoch nicht punktgenau zu verifizieren ist. Die Schätzergebnisse weisen darauf hin, dass die Hazardraten von jüngeren Personen im Vergleich zu Älteren mit ansonsten gleichen Merkmalsausprägungen um 150 bis 250 Prozent höher sind. Die Werte der übrigen signifikanten Koeffizienten reagieren weitaus weniger stark auf die verschiedenen Altersgruppen.
Die ökonomisch nicht fundierte Abgrenzung der Altersgruppen hat somit keinen Einfluss auf die entscheidenden Ergebnisse.

V.3.5.4 Motivation – Simultanitätsproblem in der Ereignisanalyse

In der bisherigen Analyse und in der Diskussion des methodischen Teils zur Ereignisanalyse wurde angenommen, dass die Kovariablen zeitkonsistent sind. Es ist fraglich, ob diese Annahme für das Merkmal Motivation erfüllt ist.
Aufgrund von Langzeitarbeitslosigkeit und dem ausbleibenden Erfolg der Fortbildung kann vermutet werden, dass die entsprechenden Personen enttäuscht sind und ehemals motivierte Arbeitssuchende könnten, angesichts der aussichtslosen Lage am Arbeitsmarkt, resignieren.
Wenn Motivation eine zeitabhängige Variable ist, könnte ein Simultanitätsproblem vorliegen und es wäre nicht zu bestimmen, ob Motivation kausal auf die Wahrscheinlichkeit der Wiederbeschäftigung wirkt oder umgekehrt, die Aufnahme einer Arbeit die Motivation erhöht.
Dieses Problem ist auf eine Schwäche der methodischen Vorgehensweise einer ex ante Befragung zurückzuführen, wenn zeitinkonsistente Merkmale in das Modell miteinfließen. So müsste eine Befragung zur Ermittlung persönlichen Merkmale unmittelbar nach Ende der Umschulung durchgeführt werden und diese Information später dazu genutzt werden, um den Einfluss der Merkmale auf die Arbeitslosigkeitsdauer zu analysieren. Schließlich sind die Ausprägungen der Merkmale zu Beginn der Arbeitssuche entscheidend für den Erfolg.
Eine OLS-Schätzung der Regression des Beschäftigungsstatus (Beschäftigt = 1, Arbeitslos = 0) auf den Motivationsindex zeigt, dass sich der durchschnittliche Motivationsindex beider Gruppen signifikant unterscheidet (Tabelle 9). Im Mittel haben arbeitslose Personen einen Motivationsindex von 7,58. Der Index der beschäftigten Personen ist um 1,82 Punkte höher.

Ist dieses Ergebnis darauf zurückzuführen, dass eine Wiederbeschäftigung motivierend wirkt oder haben vorwiegend motivierte Personen eine Arbeit aufnehmen können?
Zieht man die grafische Analyse in diese Überlegungen mit ein, ist zu vermuten, dass das

Tabelle 9: OLS-Schätzergebnisse

	Koeffizient	Std.Err.	t	P > t
Beschäftigungs-status	1.8218	0.3738	4.87	0.000
Konstante	7.5814	0.2872	26.40	0.000

Quelle: Eigene Berechnung

positive Ergebnis der Motivation nicht das alleinige Produkt der methodischen Vorgehensweise ist. In der Gruppe der hoch Motivierten konnten innerhalb eines Monats 40 Prozent der Personen eine Arbeit aufnehmen. Geht man davon aus, dass ein Großteil vom Praktikumsbetrieb übernommen wurde, nutzten die betroffenen Umschüler die Möglichkeit, ihre Fähigkeiten oder ihr Engagement während des Praktikums unter Beweis zu stellen. Die Mitnahmeeffekte wären nicht so stark ausgefallen, wenn sie sich in der Praktikumszeit nicht engagiert hätten. Daher ist davon auszugehen, dass diese Gruppe auch bereits am Ende der Maßnahme motiviert war und nicht nur zum Zeitpunkt der Befragung. Umgekehrt konnten die Personen, die zum Befragungszeitpunkt als gering motiviert eingestuft wurden, das Praktikum nicht dazu nutzen, den Praktikumsbetrieb von den eigenen Leistungen zu überzeugen. Lediglich 5 Prozent der gering Motivierten konnten innerhalb des ersten Monats nach Ende der Umschulung eine Arbeit aufnehmen. Dieses ist ein signifikanter Unterschied zu dem Ergebnis der hoch motivierten Personen und es ist daher anzunehmen, dass Motivation (zumindest in der vorliegenden Stichprobe) kausal auf die Beschäftigungswahrscheinlichkeit wirkt.

Sicherlich ist hier weiterer Forschung nötig, um die Argumentation zu stützen. Es müsste genau ermittelt werden, wie viele Teilnehmer vom Praktikumsbetrieb übernommen wurden und welche Faktoren dazu führten. Motivation ist nur EIN Merkmal, welches die Wahrscheinlichkeit der Beschäftigung beeinflusst.
Ein zeitaufwendigerer Ansatzpunkt besteht darin, zwei Befragungen durchzuführen. Direkt am Ende der Maßnahme müssten alle potentiell zeitabhängigen Merkmale erhoben und nach einigen Monaten ermittelt werden, wer wann eine Arbeit aufnehmen konnte. Außerdem würde eine erneute Erhebung des Merkmals Motivation Klarheit schaffen, ob Motivation unabhängig von der Zeit ist oder ob es einen Zusammenhang zwischen dieser und der individuellen Arbeitsmarksituation gibt.

VI. Schlussbemerkung

Die vorliegende Studie untersucht den Einfluss von Motivation auf die Dauer der Arbeitslosigkeit nach Ende einer Umschulung. Das Verständnis von Motivation basiert auf dem Konstrukt der Leistungsmotivation, das neben kognitiven Fähigkeiten eine wichtige Determinante des berufsbezogenen Leistungserfolgs repräsentiert. Zur Messung der Leistungsmotivation wurde der Leistungsmotivationstests von Hermans, Petermann und Zielinski (1978) herangezogen.

Im Rahmen eines proportionalen Hazardraten-Modell von Cox konnte gezeigt werden, dass es einen positiv signifikanten Zusammenhang zwischen Motivation und der Wiederbeschäftigungswahrscheinlichkeit gibt. Da das Merkmal Motivation am Ende des Betrachtungszeitraums erhoben wurde und Motivation zeitabhängig sein kann, ist jedoch fraglich, ob Motivation kausal auf die Dauer der Arbeitslosigkeit wirkt oder umgekehrt.

In der weiteren Untersuchung konnte gezeigt werden, dass Motivation in der zugrundegelegten Stichprobe relativ zeitkonsistent ist. Diese Schlussfolgerung ergibt sich aus der deskriptiven Auswertung der Daten. Es ist zu beobachten, dass innerhalb des ersten Monats nach Maßnahmeende 25 Prozent der Befragten eine Beschäftigung aufnehmen konnten. In den Folgemonaten vermochten dieses nur durchschnittlich 3,8 Prozent. Angesichts der Tatsache, dass am Ende einer Umschulung ein betriebliches Praktikum vorgesehen ist, kann vermutet werden, dass die deutlichen Unterschiede unter anderem darauf zurückführen sind, dass ein Teil der Untersuchungsteilnehmer von den Praktikumsbetrieben übernommen worden sind. Es ist davon auszugehen, dass hohes Engagement, fachliche Fähigkeiten und Ausdauer während eines Praktikums die Übernahmewahrscheinlichkeit deutlich erhöht und hoch motivierte Personen deutlich mehr, im Sinne einer anschließenden Festanstellung, von einem Praktikum profitieren können als andere. Tatsächlich nahmen gegenüber 5 Prozent der niedrig Motivierten, 40 Prozent der hoch motivierten Personen direkt nach der Umschulung eine Arbeit auf. Der obigen Argumentation folgend, waren somit die Personen, die zum Befragungszeitraum einen hohen Motivationsindex hatten auch schon zu Beginn ihrer Arbeitssuche hoch motiviert. Daher weist der signifikante Zusammenhang auf einen kausalen Effekt der Motivation auf die Wiederbeschäftigungswahrscheinlichkeit hin.

Dieses ist nur unter den genannten Annahmen gültig und es bedarf weiterer Untersuchungen, um den Zusammenhang zwischen Motivation und Arbeitslosigkeitsdauer besser verstehen zu können.

Von einer allgemeinen Zeitinkonsistenz des Merkmals Motivation auszugehen, erscheint höchst fraglich, widerspricht dies doch dem intuitiv nahe liegenden Sachverhalt, dass anhaltende Arbeitslosigkeit zu Resignation und einer Minderung der Motivation führen kann. Um diesem Problem zu begegnen, müssten ereignisanalytische Untersuchungen durchgeführt werden, die das Merkmal Motivation berücksichtigen, dessen Ausprägung am Ende der Maßnahme erhoben worden ist.

Trotzdem weisen die Ergebnisse der Untersuchung eindeutig darauf hin, dass die Nichtbeachtung des Charakteristikums Motivation in der Evaluationsforschung problematisch sein und eine Selektionsverzerrung verursachen kann. Matchingschätzer und DvD-Schätzer auf Basis der in Deutschland verfügbaren Datensätze führen ohne zusätzliche Selektionskorrekturen zu keinen validen Ergebnissen. Ob die angewendeten Verfahren zur Korrektur der Selektionsverzerrung ausreichend sind, ist nur schwer zu beurteilen, weil die kontrafaktischen Ergebnisse niemals beobachtet werden können. Daher ist es notwendig, alle für die Ergebnisvariable und die Partizipationsneigung entscheidungsrelevanten Merkmale zu berücksichtigen, um den Einfluss unbeobachtbarer Variablen zu minimieren. Angesichts des aufwendigen Verfahrens der Motivationsmessung ist die Realisierbarkeit, Motivation in breit angelegten Paneldatenerhebungen zu berücksichtigen, nicht gegeben. Könnte man jedoch geeignete Instrumente zur ökonomischen und effizienten Erfassung von Motivation finden (vielleicht Mobilitätsbereitschaft), wäre man dem Ziel einer validen Schätzung von Effekten der Teilnahme an einer Maßnahme der Förderung beruflicher Weiterbildung etwas näher gekommen.

Anhang 1

Tabelle A.1: Verwendete Datensätze in deutscher Evaluationsforschung

Deutsches sozio-ökonomisches Panel	Das **GSOEP** wird seit 1984 als jährliche Längsschnittsbefragung durchgeführt. Es beinhaltet umfangreiche Daten zu vielen ökonomischen und sozialen Aspekten. Der Nachteil besteht jedoch in dem geringen Stichprobenumfang zu speziellen Fragestellungen, wie zum Beispiel einer FbW-Teilnahme. Obwohl die detaillierten Daten eine ausreichende Differenzierung zwischen den verschiedenen Arten der FbW zulassen würden, reicht die Anzahl der Beobachtungen nicht aus, eine angemessene Analyse der heterogenen Effekte unterschiedlicher Maßnahmearten durchzuführen.
Arbeitsmarkt-monitor Ost	Der **AMO** wurde 1990 von der BA als kontinuierliche Längsschnittbefragung zur Beobachtung des Arbeitsmarktes in Auftrag gegeben und 1994 nach acht Umfragewellen eingestellt. Lediglich für das Land Sachsen-Anhalt gibt es eine Folgebefragung (AMSA). Der AMO weist relativ hohe Fallzahlen bei Personen mit Maßnahmeerfahrung auf. Eine hinreichende differenzierende Befragung der Teilnahme nach Maßnahmearten wird nicht vorgenommen und die relativ kurze Zeitspanne die mit dem AMO abgedeckt wird, verhindert eine Evaluation langfristiger Effekte.
Arbeitsmarkt-monitor Sachsen Anhalt	Der **AMSA** ist eine Fortsetzung des AMO im Land Sachsen-Anhalt. Dieser Datensatz hat den entscheidenden Nachteil, dass eine Differenzierung zwischen staatlich und privat geförderter Weiterbildung nicht möglich ist. Dieses geschieht im AMO und GOSEP durch Betrachtung des Bezugs von Unterhaltsgeld. Bei gleichzeitigem Bezug von Unterhaltsgeld und der Teilnahme an einer Maßnahme wird von einer staatlich geförderten Maßnahme ausgegangen.

Quelle: Eigene Zusammenfassung

Anhang 2

Matching-Algorithmus

Step 1	Divide the individuals into two separate groups called trainees and non-trainees according to whether they have participated in public vocational training during the given time span (trainee group) or not (non-trainee group).
Step 2	Randomly select a trainee (denoted by i) from the trainee group. If this trainee participated in more than one public vocational training course take the earliest one as being relevant for the following steps.
Step 3	Based on the estimated unbalanced panel probit model compute the (unbounded) propensity score $X_{it}'\hat{\beta}$ and its variance $Var(X_{it}'\hat{\beta})$ for the trainee i in wave t, where t refers to the date of the interview prior the beginning of public vocational course. Construct the interval (caliper) $X_{it}'\hat{\beta} \pm c\sqrt{Var(X_{it}'\hat{\beta})}$ for this trainee, and choose c such that one obtains a 90%-confidence interval around $X_{it}'\hat{\beta}$.
Step 4	Find observations in the non-trainee group (denoted by j), that obey $X_{jt}'\hat{\beta} \in \left(X_{it}'\hat{\beta} \pm c\sqrt{Var(X_{it}'\hat{\beta})} \right)$ in wave t.
Step 5 a)	If there is non non-trainee lying between the given limits of the confidence interval, trainee i will not be considered further and step 2 has to be repeated.
b)	If there are one or more observations in the confidence interval proceed as follows: Compute additional match variables related to monthly pre-training employment status and a subset of variables already included in the estimation of the propensity score. Denote these variables as a_{it} and a_{jt}. Evaluate the distance $d(j,i) = \left(X_{jt}'\hat{\beta}, a_{jt} \right)' - \left(X_{it}'\hat{\beta}, a_{it} \right)'$ between each non-trainee j and trainee i. Choose the non-trainee who is the "closest neighbor" of the trainee i in terms of the Mahalanobis distance, defined as: $md(j,i) = d(j,i)' L^{-1} d(j,i)$, where L is the estimated sample covariance matrix of $\left(X'\hat{\beta}, a \right)'$ in the group of non-trainees in wave t,
Step 6	Remove the trainee and non-trainee (now matched control) from their respective groups. If there are any observations left in the trainee group, start again with step 2.

Quelle: In Anlehnung an Hujer und Wellner (2000a: 43).

.

Anhang 3

Tabelle A.3.1: Übersicht über Evaluationsstudien von Qualifizierungsmaßnahmen in Westdeutschland

STUDIE	DATENSATZ/ BEOBACH-TUNGEN	ART DER QUALI-FIZIERUNG (QM)	ERFOLGS-KRITERIUM	EVALUATIONSMETHODE	WIRKUNG AUF ER-FOLGSKRITERIUM
Hujer/Maurer/ Wellner (1997a)	GSOEP West 1986 – 1993 n(KT)=1996 n(QM)=100	berufliche Weiter-bildung	Abgänge aus AL	Diskretes Hazardratenmo-dell mit unbeobachteter Heterogenität. IV-Metho-den zur Berücksichtigung von Selektionsverzerrun-gen und PPT.	• kurze Maßnahmen: + (weniger als 12 Monate) • lange Maßnahmen: 0/+ (mehr als 12 Monate)
Hujer/Maurer/ Wellner* (1997b)	GSOEP West 1984 – 1994 n(QM)=121[1]	berufliche Weiter-bildung	Abgänge aus AL	Diskretes Hazardratenmo-dell (matches-sampling An-satz)	kurzfr.: + langfr.: +
Hujer/Maurer/ Wellner (1998)	GSOEP West 1986 – 1993 n=1180 n(QM)=113[1]	berufliche Weiter-bildung	Abgänge aus AL	Diskretes Hazardratenmo-dell mit unbeobachteter Heterogenität. Matching auf Basis propensity scores und zeitvariater Informationen (Abwandlung des R&R-Matchingalgorithmus)	kurzfr.: + langfr.: 0
Hujer/Wellner (2000b)	GSOEP West 1985 – 1993 n=1511 n(KT)=177 n(QM)=89	Fortbildung und Umschulung mit Unterhaltsgeld	Dauer der AL	ML-Schätzung eines diskre-ten Hazardraten-Modell mit unbeobachteter Heterogeni-tät basierend auf einem ge-matchten Sampel. (Oversampling-Matching nach R&R-Algorithmus)	• kurze Maßnahmen: (bis zu 6 Monaten) kurzfr.: + ; langfr.: 0 • lange Maßnahmen: (mehr als 6 Monate) kurzfr.: 0/- ; langfr.: 0/-
Klose/Bender (2000)	IABS und Teilnehmer-daten der FuU n=1986 i) n(QM)=985[1] ii) n(KT)=846 n(QM)=838	staatlich geförderte Fortbildung und Umschulung	i) Dauer der AL ii) Beschäfti-gungsstabilität	Ereignisanalytisches Modell: Piecewise-constant expo-nential Modell. (Hierarchisches Matching auf Basis der Erwerbsgeschichte)	i) kurzfr.: +; langfr.: 0 ii) -
Lechner/Mi-quel/Wunsch (2004a)	IABS/TPD/ LED 1993 – 1994 n=8367 n(i)= 503 n(ii)= 267 n(iii)=386 n(iv)=247	Staatlich geförderte FbW: i) Kurzmaßnahmen ii) lange Maßnahmen iii) Umschulung iv) praxisorientierte Weiterbildung	• Beschäftigungs-effekte (BE) • Einkommen (EE)	Matchingschätzer (Oversampling-Matching in An-lehnung an R&R-Algorithmus)	BE EE i) + + ii) + + iii) - - iv) 0/+ 0

[1] : Zahlen beziehen sich auf die Anzahl der gefundenen Paare (Matches)

-	: signifikant positiver Effekt	KA	: außerbetrieblich	kurzfr.	: kurzfristig
0	: kein signifikanter Effekt	KB	: innerbetrieblich	langfr.	: langfristig
+	: signifikant positiver Effekt	KT	: Kontrollgruppe	PPT	: Preprogram-Test
QM	: Qualifizierungsmassnahme	oft	: off-the-job training	AL	: Arbeitslosigkeit
mUG	: mit Bezug von Unterhaltsgeld	ojt	: on-the-job training	BE	: Beschäftigungseffekt
oUG	: ohne Bezug von Unterhaltsgeld	n(.)	: Stichprobenumfang der jeweiligen Gruppe	EE	: Einkommenseffekt

Fortsetzung auf der nächsten Seite ...

Quelle: Eigene Zusammenstellung mit Ausnahme * Prey (1999: 129f.)

... Fortsetzung der vorherigen Seite

STUDIE	DATENSATZ/ BEOBACH-TUNGEN	ART DER QUAL-IFIZIERUNG (QM)	ERFOLGS-KRITERIUM	EVALUATIONSMETHODE	WIRKUNG AUF ER-FOLGSKRITERIUM
Pannenberg* (1995)	GSOEP West 1984 – 1991 i) n=1965 n(ojt)=308 ii) n=715 n(oft)=26	• on-the-job train-ing (ojt) • off-the-job-train-ing (oft)	i) Einkommen ii) Abgangsrate aus AL	- Diskretes Hazardraten-Modell - Probit-/Logit-Modell - lineares Paneldatenmodell mit festen Effekten	i) ojt: + ojt (mUG): 0 ii) oft: + oft x Dauer der QM: - oft (mUG): 0
Prey (1997)*	GSOEP West 1984 – 1993 n= 7552 n(QM)=134 (1985)	Fortbildung und Umschulung mit (QM(mUG)) und ohne (QM(oUG)) Unterhaltsgeldför-derung	Beschäftigungs-wahrscheinlichkeit	Simultanes, dynamisches Random-Effects Probit-Modell mit PPT	• QM(mUG): Männer: - ; Frauen: 0 • QM(oUG): Männer: 0 ; Frauen: -
Prey (1999)	GSOEP West 1985 – 1994	Fortbildung und Umschulung mit (QM(mUG)) und ohne (QM(oUG))	i) Beschäftigungs-effekt ii) Lohneffekte	Dynamisches Random-Effects Probit und Tobit Modelle mit simultaner Schätzung der Teilnahme und Ergebnisvariablen	<table><tr><th>QM</th><th></th><th>oUG</th><th>mUG</th></tr><tr><td>kurzfr.</td><td>i)</td><td>0 / +</td><td>+</td></tr><tr><td></td><td>ii)</td><td>0 / +</td><td>0 / +</td></tr><tr><td>langfr.</td><td>i)</td><td>-</td><td>0</td></tr><tr><td></td><td>ii)</td><td>0 / -</td><td>0 / -</td></tr></table>
Staat (1997)*	GSOEP West 1992 – 1994 i) n=1702 n(QM)=311 ii) n=1569 n(QM)=247	Fortbildung und Umschulung mit Unterhaltsgeld	i) Dauer der Arbeitssuche ii) Beschäfti-gungstabilität	Ordered-Probit-Regression mit Instrumentvariablen	i) 0, außer bei: •Personen zwischen 45-54 Jahren: - •Personen ohne Berufs-ausbildung: - •Frauen: - ii) 0, außer bei: Personen ohne Berufs-Ausbildung: +

¹ : Zahlen beziehen sich auf die Anzahl der gefundenen Paare (Matches)

-	: signifikant positiver Effekt	KA	: außerbetrieblich	kurzfr.	: kurzfristig
0	: kein signifikanter Effekt	KB	: innerbetrieblich	langfr.	: langfristig
+	: signifikant positiver Effekt	KT	: Kontrollgruppe	PPT	: Preprogram-Test
QM	: Qualifizierungsmassnahme	oft	: off-the-job training	AL	: Arbeitslosigkeit
mUG	: mit Bezug von Unterhaltsgeld	ojt	: on-the-job training		
oUG	: ohne Bezug von Unterhaltsgeld	n(.)	: Stichprobenumfang der jeweiligen Gruppe		

Quelle: Eigene Zusammenstellung mit Ausnahme * Prey (1999: 129f.)

Tabelle A.3.2: Übersicht über Evaluationsstudien von Qualifizierungsmaßnahmen in Ostdeutschland

STUDIE	DATENSATZ/ BEOBACHTUNGEN	ART DER QUALIFIZIERUNG (QM)	ERFOLGSKRITERIUM	EVALUATIONSMETHODE	WIRKUNG AUF ERFOLGSKRITERIUM
Bergemann/ Fitzenberger/ Speckesser (2001)	AMSA 1990 – 1999 n=5224 n(QM(1))= 1021 n(QM(2))= 147	Teilnahme an einer (QM(1)) oder zwei (QM(2)) Maßnahmen der Fortbildung oder Umschulung mit Unterhaltsgeldföredrung	• Beschäftigungswahrscheinlichkeit • Übergangsrate in Beschäftigung	• conditionaler DvD-Schätzer (Kernel- und nearest-neighbor Matching auf Basis geschätzter propensity scores)	• QM(1): 0 • QM(2): - Gesamteffekt: 0 - *incremental effect* der zweiten FbW: 0/+
Bergemann/ Fitzenberger/ Speckesser (2004)	AMSA 1990 – 1999 n=5165 n(QM(1))= 1021 n(QM(2))= 150	Teilnahme an einer (QM(1)) oder zwei (QM(2)) Maßnahmen der Fortbildung oder Umschulung mit Unterhaltsgeldföredrung	i) Beschäftigungswahrscheinlichkeit • Übergangsrate in Beschäftigung nach FbW die: ii) 1990 begann iii)1994 oder später begann	• semiparametrischer, conditionaler DvD-Schätzer (Kernel-Matching auf Basis von propensity-scores)	• QM(1): i) - ; ii) + ; iii) 0 / - • QM(2): - Gesamteffekt: 0 - *incremental effect* der zweiten FuU: 0 / +
Bergemann et al. (2000)	AMSA 1990 – 1998 n=4645 n(QM(1))= 915[1] n(QM(2))= 185[1]	Teilnahme an einer (QM(1)) oder zwei (QM(2)) Massnahmen der Fortbildung oder Umschulung mit Unterhaltsgeldföredrung	Beschäftigungsquote	• conitionaler DvD-Schätzer (nearest neighbor Matchingansatz auf Basis propensity scores)	• QM(1): - kurzfr.: - - langfr.: - • QM(2): 0
Eichler/ Lechner (2000)	AMSA 1992 – 1993 n=1153 n(QMmUG)= 125 bzw. n=1202 n(ojt)=222	• Fortbildung und Umschulung mit Unterhaltsgeldförderung (QMmUG) • on-the-job training (ojt)	i) Arbeitslosenquote ii) Löhne	conditionaler DvD-Schätzer	• QMmUG: i) kurzfr.: - ; langfr.: 0 ii) 0 • ojt: i) kurzfr.: 0 ; langfr.: 0 kurzfr.: ii) 0/+ ; langfr.: +
Fitzenberger/ Prey** (1997)	AMO 1990 – 1992 n=8681 n(KA)=307[2] n(KB)=270[2]	• Fortbildung im Betrieb (KB) • Fortbildung außerhalb des Betriebs (KA)	• Beschäftigungswahrscheinlichkeit	Simultane Schätzung der Teilnahme und Ergebnisvariable eines dynamischen Random-Effects Probit und Tobit Modells	• KA(langfr.): + • KB (langfr.): - / 0
Fitzenberger/ Prey (1998)*	AMO 1990 – 1994	Fortbildung und Umschulung mit Unterhaltsgeldförderung	Erwerbstätigkeit Löhne	• dynamische (i) und statische (ii) Random Effects-Schätzung eines simultanen Probit-Tobit-Modells für Teilnahme und Ergebnisvariablen , Ermittlung der Nettoeffekte durch DvD-Schätzer und PPT • Matchingansatz (iii), (in Anlehnung am R&R-Algorithmus)	kurzf. langfr. i) - 0/+ ii) - 0/+ iii) - 0

[1] : Zahlen beziehen sich auf die Anzahl der gefundenen Paare (Matches)

[2] : Zahlen beziehen sich auf die durchschnittliche Zahl der Personen pro Welle, für die ein ex-post QM-Effekt gemessen wird

-	: signifikant positiver Effekt	KA	: außerbetrieblich	kurzfr.	: kurzfristig
0	: kein signifikanter Effekt	KB	: innerbetrieblich	langfr.	: langfristig
+	: signifikant positiver Effekt	KT	: Kontrollgruppe	PPT	: Preprogram-Test
QM	: Qualifizierungsmassnahme	oft	: off-the-job training	AL	: Arbeitslosigkeit
mUG	: mit Bezug von Unterhaltsgeld	ojt	: on-the-job training	M	: Männer
oUG	: ohne Bezug von Unterhaltsgeld	n(.)	: Stichprobenumfang der jeweiligen Gruppe	F	: Frauen

Fortsetzung auf der nächsten Seite...

Quelle: Eigene Zusammenstellung mit Ausnahme * Fitzenberger und Speckesser (2002: 364) und ** Prey (1999: 147f.)

... Fortsetzung der vorherigen Seite

STUDIE	DATENSATZ/ BEOBACHTUNGEN	ART DER QUALIFIZIERUNG (QM)	ERFOLGSKRITERIUM	EVALUATIONSMETHODE	WIRKUNG AUF ERFOLGSKRITERIUM
Fitzenberger/ Prey (2000)	AMO 1990 – 1994 n=4823 QM(M)=325 QM(W)=146	Fortbildung und Umschulung mit Unterhaltsgeldförderung, Männer (QM(M)) und Frauen (QM(W)) getrennt	i) Beschäftigung ii) Einkommen	Simultane Random-Effects Schätzung eines Probit und Tobit Modells für Teilnahme und Ergebnisvariable. Ermittlung der Nettoeffekte durch DvD-Schätzer und PPT	QM(W): i) 0/+ ii) 0 QM(M): i) 0/+ ii) 0
Hübler (1994)*	AMO 1990		i) Wahrscheinlichkeit derArbeitssuche ii) Arbeitszeit	Simultane Probit Schätzung für Teilnahme und Ergebnisvaribale	i) + ii) -
Hübler* (1997)	AMO 1990 – 1994 n=2886	staatlich geförderte berufliche Weiterbildung	Beschäftigungswahrscheinlichkeit	i) Multinominales Logitmodell Random-Effects-Probit-Schätzung mit: ii) Kontrolle für beobachtbare Merkmalen iii) nach Matching (Iteratives Matching in Anlehnung am R&R Algorithmus und PPT-orientiert)	i) + ii) + iii) Männer: + Frauen: -
Hujer/Wellner (2000a)	GSOEP Ost 1990 – 1993 i) n = 1543 n(QM)=142¹ ii) n=1524 n(KT)=242 n(QM)=123	Fortbildung und Umschulung mit Unterhaltsgeldförderung	Effekte auf: i) AL-dauer ii) Beschäftigungsstabilität (von Maßnahmen mit einer Dauer von bis zu 3, 4-6 oder mehr als 7 Monaten)	ML-Schätzung eines diskreten Hazardraten-Modells mit unbeobachteter Heterogenität, basierend auf einem gematchten Sampel. (Matching nach R&R-Algorithmus)	Keine signifikanten Effekte auf Ergebnisvariablen. (Dreimonatige Maßnahmen scheinen jedoch wirksamer zu sein als Maßnahmen, die länger als 6 Monate dauern.)
Hujer/Wellner (2000b)	GSOEP Ost 1990 – 1993 n=1458 n(KT)=272 n(QM)=142	Fortbildung und Umschulung mit Unterhaltsgeldförderung	AL-dauer nach Teilnahme an Kursen i) bis zu 6 Monaten ii) länger als 6 Monate	ML-Schätzung eines diskreten Hazardraten-Modells mit unbeobachteter Heterogenität, basierend auf einem gematchten Sampel. (Oversampling-Matching nach R&R-Algorithmus)	i) kurzfr.: 0 ; langfr.: 0 ii) kurzfr.: 0/- ; langfr.: 0/-
Kraus/Puhani/ Steiner (2000)	AMO 1990 – 1992 1992 – 1994 n ≈ 6000	Fortbildung und Umschulung mit Unterhaltsgeldförderung	Übergangsrate in Beschäftigung	Diskretes Hazardraten- Modell mit Kontrolle für beobachtbarer Charakteristika und PPT	<table><tr><td>Ort der FbW</td><td colspan="2">KA</td><td colspan="2">KB</td></tr><tr><td>Geschlecht</td><td>M</td><td>F</td><td>M</td><td>F</td></tr><tr><td>1990 - 1992</td><td>-</td><td>-</td><td>-</td><td>-</td></tr><tr><td>1992 - 1992</td><td>+</td><td>+</td><td>0</td><td>+</td></tr></table>
Lechner** (1996a)	GSOEP Ost 1990 – 1994 n=1163 n(QM)=103¹	Fortbildung und Umschulung mit Unterhaltsgeldförderung	i) Arbeitslosigkeit ii) Beschäftigung iii) Einkommen	nicht-parametrisches Matching zur Ermittlung des durchschnittlichen Effekts	i) kurzfr.: + ; langfr.:0 ii) kurzfr.: - ; langfr.:0 iii) 0
Lechner** (1996b)	GSOEP Ost 1990 – 1994 n=1075 n(QM)=185¹	betriebliche Weiterbildung	i) Arbeitslosigkeit ii) Beschäftigung iii) Einkommen	nicht-parametrisches Matching zur Ermittlung des durchschnittlichen Effekts	i) 0 ii) 0 iii) +

¹ : Zahlen beziehen sich auf die Anzahl der gefundenen Paare (Matches)

-	: signifikant positiver Effekt	KA	: außerbetrieblich	kurzfr.	: kurzfristig
0	: kein signifikanter Effekt	KB	: innerbetrieblich	langfr.	: langfristig
+	: signifikant positiver Effekt	KT	: Kontrollgruppe	PPT	: Preprogram-Test
QM	: Qualifizierungsmassnahme	oft	: off-the-job training	AL	: Arbeitslosigkeit
mUG	: mit Bezug von Unterhaltsgeld	ojt	: on-the-job training	M	: Männer
oUG	: ohne Bezug von Unterhaltsgeld	n(.)	: Stichprobenumfang der jeweiligen Gruppe	F	: Frauen

Fortsetzung auf der nächsten Seite...

Quelle: Eigene Zusammenstellung mit Ausnahme * Fitzenberger und Speckesser (2002: 364) und ** Prey (1999: 147f.)

... Fortsetzung der vorherigen Seite

STUDIE	DATENSATZ/ BEOBACHTUNGEN	ART DER QUALIFIZIERUNG (QM)	ERFOLGSKRITERIUM	EVALUATIONSMETHODE	WIRKUNG AUF ERFOLGSKRITERIUM
Lechner** (1998a)	GSOEP Ost 1990 – 1994 n=1339 n(QM)=122[1]	außerbetriebliche Weiterbildung	i) Arbeitslosigkeit ii) Einkommen	nicht-parametrisches Matching zur Ermittlung des durchschnittlichen Effekts	i) kurzfr.: + ; langfr.:0 ii) 0
Lechner* (1998b)	GSOEP Ost 1990 – 1994		i) Beschäftigungsquote ii) AL-quote iii) Löhne	Matching mit propensity score und zeitvariaten Kovariablen	i) kurzfr.: - ; langfr.: 0 ii) kurzfr.: + ; langfr.: 0 iii) 0
Lechner (1999)	GSOEP Ost 1990 – 1994 n=1339 n(QM)=140	Förderung der beruflichen Weiterbildung (42 Prozent der Personen beziehen Unterhaltsgeld)	i) Beschäftigungswahrscheinlichkeit ii)Einkommen	Schätzung der Differenz der durchschnittlichen Ergebnisvariablen der AB und KT auf Basis drei verschiedener Matchingverfahren	i) 0 ii) + (R&R-Matching ist dominierender Zuweisungs-prozeß)
Lechner (2000)	GSOEP Ost 1990 – 1996 n=1411 n(QM)=161[1]	berufliche Weiter-Bildung mit Unterhaltsgeldförderung	i) Beschäftigungswahrscheinlichkeit ii)Einkommen	conditionaler DvD-Schätzung (Zuordnungsprozeß in Anlehnung an R&R-Algorithmus)	i) kurzfr.: - langfr.: 0 ii) 0
Pannenberg** (1995)	GSOEP Ost 1990 – 1992 n=2017 n(QM)=76	Fortbildung und Umschulung	i) Abgangsrate aus AL ii) Löhne	• Diskretes Hazardraten-Modell • Lineare Panelschätzung (fixed effects) für Löhne	i) FuU: 0 ; FuU x UG: - FuU x Dauer der QM: + FuU x Dauer der AL: - ii) FuU: + ; FuU x UG: 0 FuU x Dauer der Al.: 0
Prey (1999)	GSOEP Ost 1990 – 1994 n=8751	Fortbildung inner- oder außerhalb des Betriebs mit oder ohne Unterhaltgeldsförderung, für Männer und Frauen getrennt: KA(mUG),KA(oUG) KB(mUG),KB(oUG)	i) Beschäftigungseffekt ii) Löhne	Random-Effects Schätzung eines dynamischen Probit und Tobit Modells mit simultaner Modellierung der Teilnahme und der Ergebnisvariablen	Beschäftigungseffekt: KA mUG / KA oUG / KB mUG / KB oUG M kurzfr. - 0 0 + M langfr. + 0 - 0 W kurzfr. - + 0/- + W langfr. + 0 0 0/- Lohneffekt: KA mUG / KA oUG / KB mUG / KB oUG M kurzfr. 0 - 0 0 M langfr. 0 + 0 + W kurzfr. - 0 - 0 W langfr. 0/+ + 0 +
Reinowski/ Schultz/ Wiemers (2003)	AMSA 2000 – 2001 n=1287 n(QM)=146[1]	Förderung der beruflichem Weiterbild-für Langzeitarbeitslose	Abgangsrate aus Arbeitslosigkeit	• Matching (durch iterative Zuordnung auf Basis propensity scores und Erwerbsvorgeschichte) • Cox-Proportional-Hazard-rate-Modell nach Breslow	kurzfr.: - ; langfr.: -
Reinowski/ Schultz/ Wiemers (2004)	AMSA 1990 – 2001 n(QM)=850[1]	staatlich geförderte Fortbildung und Umschulung	Abgangsrate aus Arbeitslosigkeit	•Matching: hungarian Algorithmus auf Basis propensity score und Erwerbsvorgeschichte •Cox-Proportional-Hazard-rate-Modell nach Breslow	kurzfr.: - ; langfr.: - Besonders nachteilig wirkt sich FbW auf: Frauen, Ältere und (verglichen zu Teilnehmern aus Dresden) auf Förderpersonen aus Raum Leipzig und Chemnitz aus.

[1] : Zahlen beziehen sich auf die Anzahl der gefundenen Paare (Matches)

-	: signifikant positiver Effekt	KA	: außerbetrieblich	kurzfr.	: kurzfristig
0	: kein signifikanter Effekt	KB	: innerbetrieblich	langfr.	: langfristig
+	: signifikant positiver Effekt	KT	: Kontrollgruppe	PPT	: Preprogram-Test
QM	: Qualifizierungsmassnahme	oft	: off-the-job training	AL	: Arbeitslosigkeit
mUG	: mit Bezug von Unterhaltsgeld	ojt	: on-the-job training	M	: Männer
oUG	: ohne Bezug von Unterhaltsgeld	n(.)	: Stichprobenumfang der jeweiligen Gruppe	F	: Frauen

Quelle: Eigene Zusammenstellung mit Ausnahme * Fitzenberger und Speckesser (2002: 364) und ** Prey (1999: 147f.)

Anhang 4

Tabelle A.4: Modifizierter LMT-Test nach Hermans, Petermann und Zielinski (1978)

	Frage	Antwortmöglichkeit	
1	Wenn Sie sich mit ihrer Zukunft beschäftigen, denke Sie dann meistens…	- sehr weit voraus.....	■
		- weit voraus.....	■
		- ziemlich weit voraus.....	□
		- nicht so weit voraus.....	□
2	Denke Sie, dass Leute in ihrem Alter...	- zu hart arbeiten.....	□
		- hart genug arbeiten.....	□
		- härter arbeiten sollten.....	■
3	Beim Arbeiten sind die Anforderungen, die Sie sich selber stellen…	- sehr hoch.....	■
		- hoch.....	■
		- ziemlich hoch.....	□
		- nicht so hoch.....	□
4	Mehr zu leisten als andere, halten Sie für…	- sehr wichtig.....	■
		- wichtig.....	■
		- ziemlich wichtig.....	■
		- nicht so wichtig.....	□
5	An einer Sache lange zu arbeiten, ohne zu ermüden…	- fällt Ihnen schwer.....	□
		- gelingt Ihnen nicht so gut.....	□
		- fällt Ihnen leicht.....	■
6	In der Umschulung waren Sie…	- sehr ehrgeizig.....	■
		- ziemlich ehrgeizig.....	■
		- wenig ehrgeizig.....	□
7	Für Menschen, die es im Leben weit gebracht haben empfinde Sie…	- sehr große Bewunderung.....	■
		- große Bewunderung.....	■
		- ziemlich große Bewunderung.....	□
		- wenig Bewunderung.....	□
8	Wie haben andere Sie während der Umschulung vermutlich eingeschätzt? Als…	- fleißig.....	■
		- nicht immer gleichbleibend fleißig.....	□
		- ziemlich bequem.....	□
9	Wenn Sie beim Lernen oder Arbeiten unterbrochen werde, dann ist das für Sie…	- störend.....	■
		- nicht so schlimm.....	□
		- eine angenehme Abwechslung.....	□
10	Sich lange auf eine wichtig Aufgabe vorzubereiten…	- ist häufig voreilig.....	□
		- kann manchmal nützlich sein.....	□
		- zeugt von Realitätssinn.....	■
11	Es im Leben zu etwas zu bringen,…	- wird in seiner Bedeutung überschätzt.....	□
		- nimmt in unserer Gesellschaft einen wichtigen Platz ein.....	□
		- stellt ein erstrebenswertes Ziel dar.....	■

Fortsetzung auf der nächsten Seite...

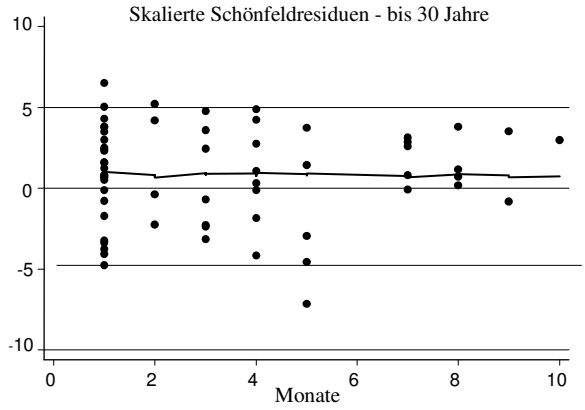
... Fortsetzung der vorherigen Seite

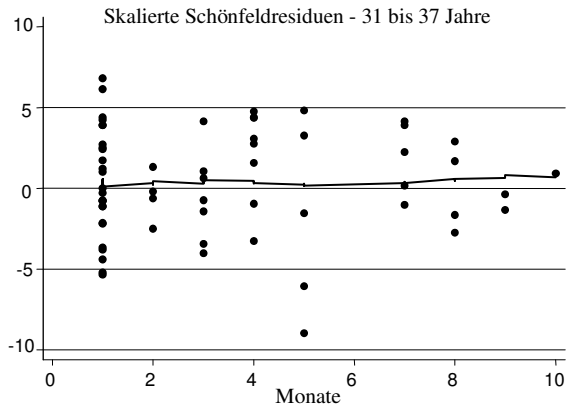
	Frage	Antwortmöglichkeit	
12	Wenn Sie mit einer schwierigen Sache beschäftigt sind…	- gebe Sie manchmal auf.....	☐
		- schieben Sie sie auf und mache später weiter.....	☐
		- bleiben Sie meistens dabei.....	■
13	Während des Unterrichts in der Umschulung…	- passten Sie meistens gut auf.....	■
		- hatten Sie manchmal Schwierigkeiten, bei der Sache zu bleiben.....	☐
		- schweiften Ihre Gedanken oft zu ganz anderen Dingen ab.....	☐
14	Was andere über Ihre Leistung denken, ist Ihnen…	- wichtig.....	■
		- ziemlich wichtig.....	☐
		- nicht so wichtig.....	☐

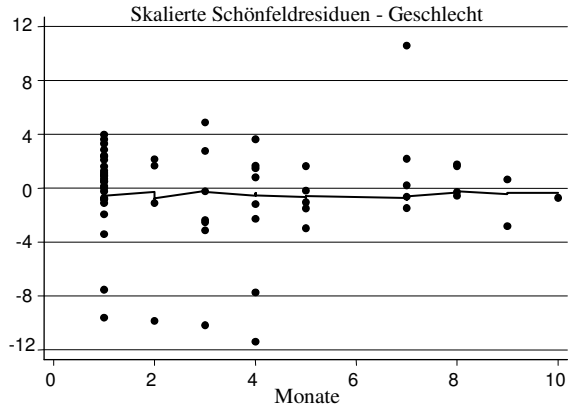
Quelle: Siehe Abschnitt V.2.3.2 und V.2.3.3. Grau unterlegte Antwortfelder deuten auf hohe Motivation des Befragten hin.

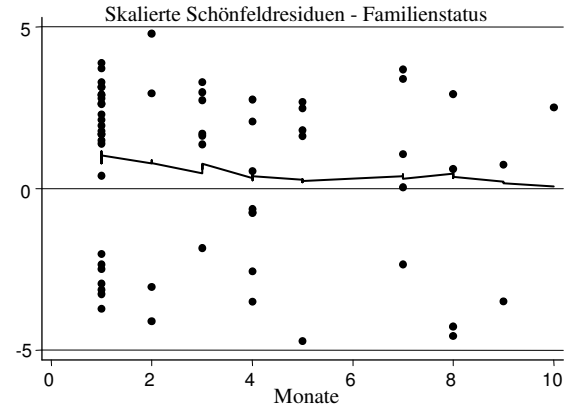
Anhang 5

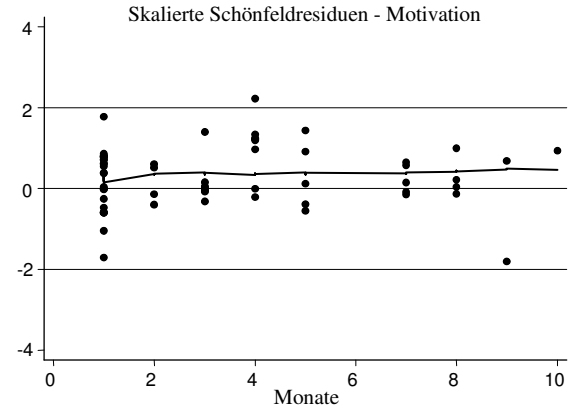
Abbildung A.5: Test der PH-Annahme

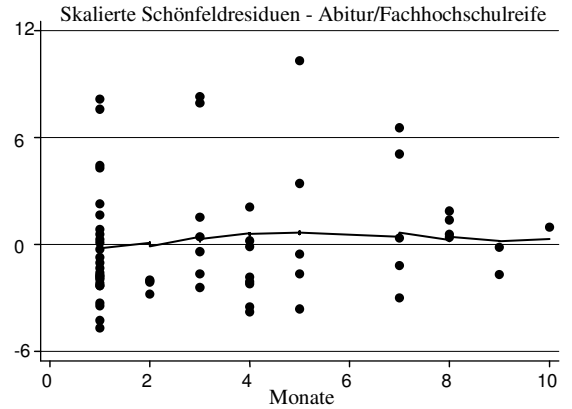


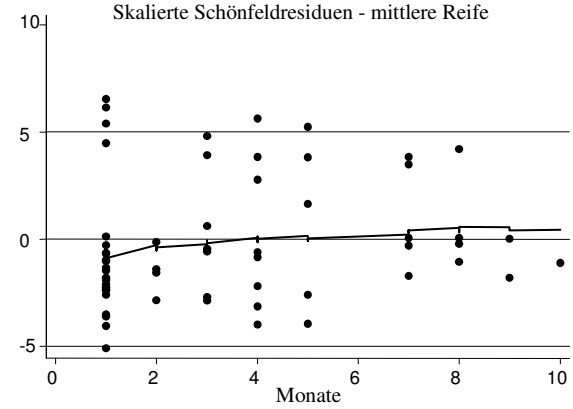












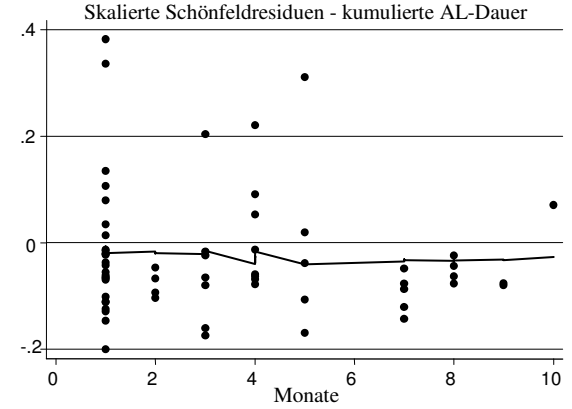
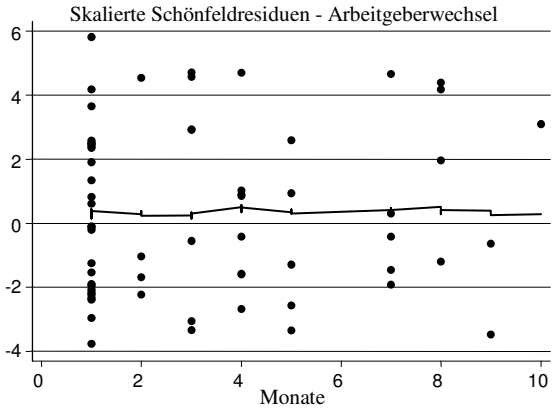
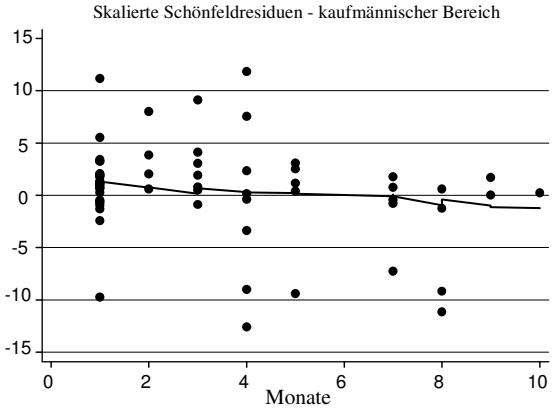
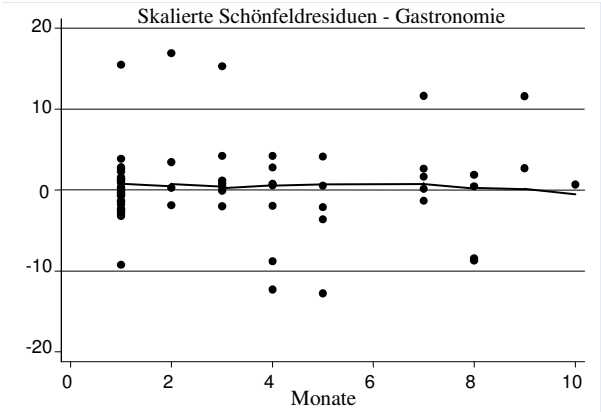


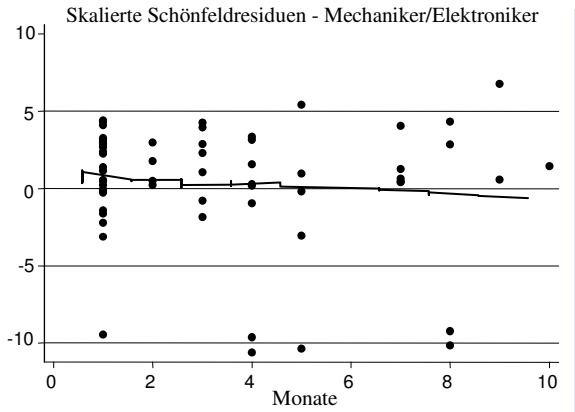
Abbildung wird fortgesetzt…

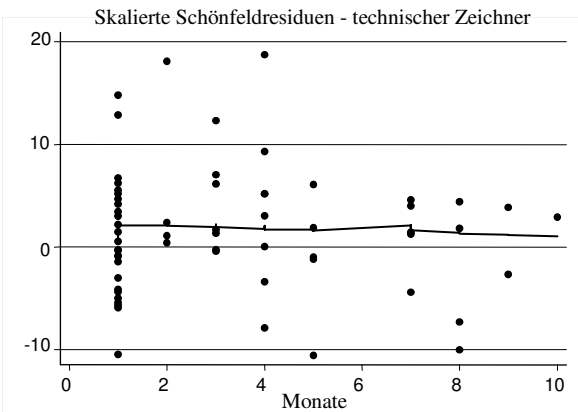
…Fortsetzung der Abbildung A.5

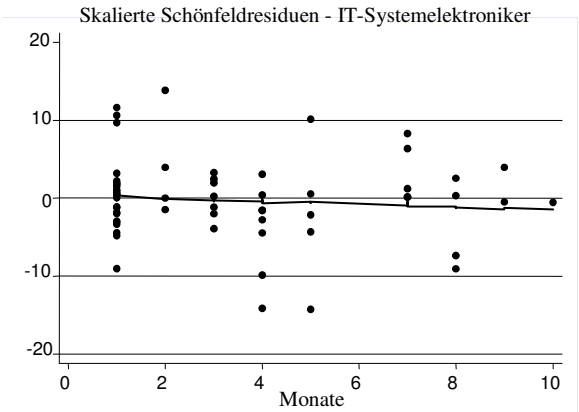












Quelle: Eigene Berechnung

Anhang 6

Tabelle A.6.1: Schätzergebnisse des geschichteten PH-Modells nach Cox

Modell	**5**	**5´**	**5´´**
Alter [AG 3°]			
AG 1°	2.6057**	2.9480***	3.1166***
	(0.9774)	(1.2126)	(1.3212)
AG 2°	1.5695	1.6200	1.6972
	(0.6500)	(0.5949)	(0.6294)
Geschlecht - weiblich	0.7137	0.8001	0.8304
	(0.3406)	(0.3858)	(0.3987)
Verheiratet	**1.7601**	**1.8561**	**1.8467**
	(0.5750)	(0.6112)	(0.6082)
Abschlussnote der Umschulung	0.7174	**0.6902**	**0.6865**
	(0.1490)	(0.1436)	(0.1432)
Motivation	1.3532***	1.3622***	1.3763***
	(0.1153)	(0.1130)	(0.1134)
Schulbildung [Hauptschulabschluss]			
Abitur/ Fachhochschulreife	0.9526	0.9452	0.9338
	(0.4353)	(0.4218)	(0.4153)
mittlere Reife	0.90236	0.8467	0.8622
	(0.3142)	(0.2928)	(0.2957)
kum. AL-Dauer in Monaten	0.9715*	0.9696*	0.9679*
	(0.0138)	(0.0144)	(0.0142)
mehr als 2 Arbeitgeberwechsel	1.3766	1.2948	1.3323
	(0.4223)	(0.4064)	(0.4143)
Ausbildungsbereich [KFZ]			
Gastronomie	1.4766	1.4314	1.4599
	(0.9844)	(0.9470)	(0.9663)
Mechaniker/Elektroniker	1.3615	1.5168	1.4838
	(0.6797)	(0.7703)	(0.7561)
Technischer Zeichner	5.0006*	4.8423*	5.0157*
	(3.7248)	(3.5794)	(3.7262)
IT-Systemelektroniker	0.8699	0.9496	0.9591
	(0.5681)	(0.6341)	(0.6424)
Liklihood-Ratio- χ^2	45.17	45.46	45.80
Prob > χ^2	0.0000	0.0000	0.0000

Anmerkung: Siehe Anmerkung unter Tabelle 6. Schichtungsvariablen sind „Nationalität“ und „kaufmännischer Bereich“. AG = Altersgurppe. ° Abgrenzung der Altersgruppen gemäß Tabelle A.7.2.

Quelle: Eigene Berechung

Tabelle A.6.2: Abgrenzung der Altersgruppen nach Jahren

Modell	**5**	**5´**	**5´´**
AG 1	bis 30	bis 29	bis 29
AG 2	31 bis 37	30 bis 34	30 bis 37
AG 3	über 37	über 34	über 37

Quelle: Eigene Berechnung

Literaturverzeichnis

Allison, P.D. (1984). Event History Analysis. Regression for Longitudinal Event Data. *Sage University Paper Series on Quantitative Applications in the Social Sciences,*7-46. Beverly Hills.

Amemiya, T. (1986). *Advanced Econometrics*. Oxford: Blackwell.

Angrist, J.D., Imbens, G.W. und D.B. Rubin (1996). Identification of Causal Effcts Using Instrumental Variables. *Journal of American Statistical Association, 91*, 444-455.

Asendorpf, J.B. (2004). *Psychologie der Persönlichkeit* (3. Auflage). Berlin u.a.: Springer Verlag.

Ashenfelter, O. und D. Card (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics, 67*, 648-660.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on earnings. *Review of economics and Statistics, 60*, 47-57.

Atkinson, J.W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*, 359-372.

Bailey, K. (1984). Asymptotic Equivalence Between the Cox Estimator and the General Maximum Likelihood Estimator of Regression and Survival Parameters. *Annals of Statistics, 12*(2), 730–736.

Bender, S., Fitzenberger, B. und M. Lechner (2002). *Über die Wirksamkeit von FuU-Maßnahmen – ein Evaluationsversuch mit prozessproduzierten Daten aus dem IAB*. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), IAB-Werkstattbericht 2/2002.

Bender, S., Hilzendegen, J., Rohwer, G. H. und Rudolph (1996). *Die IAB - Beschäftigungsstichprobe 1975-1990*. Nürnberg: IAB, Beiträge zur Arbeitsmarkt- und Berufsforschung Nr. 197.

Bergemann, A., Fitzenberger, B., Schultz, B. und S. Speckesser (2000). Multiple Active Labor Market Policy Pariticipation in East Germany: An Assessment of Outcomes. *Beihefte der Konjunkturpolitik, 51*, 195–244.

Bergemann, A., Fitzenberger, B. und S. Speckesser (2001*). Evaluating the Employment Effects of Public Sector Sponsored Training in East Germany, Conditional Difference-in-Differences and Ashenfelters's Dip*. URL am 06.12.2004: http://www.vwl.uni-mannheim.de/ls_oek/anlagen/eval14.pdf

Bergemann, A., Fitzenberger, B. und S. Speckesser (2004): *Evaluating the Dynamic Employment Effects of Training Programs in East Germany Using Conditional Difference-in-Differences*. Mannheim: Zentrum für Europäische Wirtschaftsforschung (ZEW), ZEW-Diskussionspapiere 04/41.

Bielenski, H.C., Brinkmann und B. Kohler (1995). *Erwerbsverläufe und arbeitsmarktpolitische Maßnahmen in Ostdeutschland*. Nürnberg: IAB, IAB-Werkstattbericht Nr. 12.

Blossfeld, H.P., Hamerle, A. und K.U. Mayer (1986). *Ereignisanalyse. Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften.* Frankfurt/ Main u.a: Campus Verlag.

Bovensiepen, N. (2005). Fast die ganze Wahrheit. *Süddeutsche Zeitung*, 02.02.2005, S. 2.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89-99.

Bryson, A., Dorsett, R. und S. Purdon (2002). The use of propensity score matching in the evaluation of active labour market policies. *Department for Work and Pensions Working Paper No.4*. URL am 09.12.2004: http://www.dwp.gov.uk/asd/asd5/WP4.pdf

Bundesagentur für Arbeit (1991-2004). Arbeitsstatistik – Jahreszahlen. *Amtliche Nachrichten der Bundesagentur für Arbeit*, 39-51, Nürnberg.

Bundesagentur für Arbeit (2003). Arbeitsmarkt 2002. *Amtliche Nachrichten der Bundesagentur für Arbeit*, 51, Sondernummer Nürnberg, 18. Juni 2003.

Christensen, B. (2001). *Berufliche Weiterbildung und Arbeitsplatzrisiko: Ein Matching-Ansatz*. Kiel: Institut für Weltwirtschaft (IfW), Kieler Arbeitspapiere 1033.

Cleves, M., Gould, W.W. und R.G. Gutierrez (2002). *An Introduction to Survival Analysis using Stata*. Texas: Stata Press.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34*, 187-220.

Cox, D.R. (1975). Partial Likelihood. *Biometrica, 62*, 269-276.

Dehejia, R.H. (2004). Practical propensity score matching: a reply to Smith and Todd, *Journal of Econometrics*, forthcoming. URL am 09.12.2004: http://www.columbia. edu/ ~ rd247/papers/practical_pscore.pdf

Deheija, R.H. und S. Wahba (1999). Causal Effects in Nonexperimental studies. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Dehejia, R.H. und S. Wahba (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economic and Statistics, 84*(1), 151-161.

Durbin, J. (1954). Errors in Variables. *Review of the International Institute, 22*, 23-32.

Efron, B. (1977). The Efficiency of Cox's likelihood function for censored Data. *Journal of the American Statistical Association, 72*, 557-565.

Eichler, M. und M. Lechner (2000). *Some Econometric Evidence on the Effectiveness of Active Labour Market Programmes in East Germany*. William Davidson Institute Working Papers Series, Nr. 318.

Eichler, M. und M. Lechner (2002). An Evalution of Public Employment Programmes in East German State of Sachsen-Anhalt. *Labour Economics, 9*, 143-186.

Fitzenberger, B. und R. Hujer (2002). Stand und Perspektiven der Evaluation der Aktiven Arbeitsmarktpolitik in Deutschland. *Perspektiven der Wirtschaftspolitik*, 3(2),139-158.

Fitzenberger, B. und H. Prey (1997). Assessing the Impact of Training on Employment: The Case of East Germany. *Ifo-Studien, 43*(1), 69-114.

Fitzenberger, B. und H. Prey (1998). Beschäftigungs- und Verdienstwirkungen von Weiterbildungsmaßnahmen im ostdeutschen Transformationsprozess: Eine Methodenkritik. In F. Pfeiffer und W. Pohlmeier (Hrsg.), *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*, Baden-Baden: Nomos, 39-95.

Fitzenberger, B. und H. Prey (2000). Evaluating public sector sponsored training in East Germany. *Oxford Economic Papers, 52*, 497-520.

Fitzenberger B. und S. Speckesser (2000). Zur wissenschaftlichen Evaluation der Aktiven Arbeitsmarktpolitik in Deutschland: Ein Überblick. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 33*(3), 357-370.

Fitzenberger B. und S. Speckesser (2002). *Weiterbildungsmaßnahmen in Ostdeutschland: Ein Misserfolg der Arbeitsmarktpolitik?* Mannheim: ZEW, ZEW-Diskussionspapiere 02/16.

Frieling, E. und K. Sonntag (1999). *Arbeitspsychologie* (2. Auflage). Bern u.a.: Huber Verlag.

Gangl, M. und T. DiPrete (2004). *Kausalanalyse durch Matchingverfahren*. Deutsches Institut für Wirtschaftsforschung (DIW), DIW- Diskussionspapiere 401.

Grambsch, P. und T.M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*, 515-526.

Grønnesby, J.K. und Ø. Borgan (1996). A method for checking regression models in suvival analysis based on risk score. *Lifetime Data Analysis, 2*, 315-328.

Hanefeld, U. (1987). *Das Sozioökonomische Panel – Grundlagen und Konzeption*, Frankfurt/Main u.a.: Campus.

Heckman, J.J.(1997). Instrumental Variables: A Study of the Implicit Assumptions Underlying One Widely Used Estimatorr For Program Evaluations. *Journal of Human Ressources, 32*, 441-462.

Heckman, J.J. und J.V. Hotz (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association, 84*, 862-874.

Heckman, J.J., Ichimura, H., Smith, J. und P. Todd (1996). Sources of selection bias in evaluating social programs: An interpretations of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Academy sciences, 93*, 13416-13420.

Heckman, J.J., Ichimura, H., Smith, J. und P. Todd (1998). Characterizing Selection Bias using Experimental Data. *Econometrica, 66*(5), 1017-1098.

Heckman, J.J., Ichimura, H. und P. Todd (1997). Matching as an Econometric Evaluation Esti-mator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies, 64*, 605-654.

Heckman, J.J., Ichimura, H. und P. Todd (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies, 65*, 261-294.

Heckmann, J.J., LaLonde, R.J. und J.A. Smith (1999). The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter und D. Card (Hrsg.), *Handbook of Labor Economics*, Amsterdam: North Holland, 1865-2097.

Heckhausen, H. (1963). *Hoffnung und Furcht in der Leistungsmotivation*. Meisenheim: Anton Hain.

Hermans, H. (1968). *Prestatie Motivatie Test*. Amsterdam: Swets and Zeitlinger.

Hermans, H., Petermann, F. und W. Zielinski (1978). *Leistungs-Motivations-Test (LMT),* Amsterdam: Swets and Zeitlinger.

Hosmer, D.W. und S. Lemeshow (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics, 10*, 1043-1069.

Hosmer, D.W. und S. Lemeshow (1989). *Applied Logistic Regressions*. New York u.a.: Wiley.

Hübler, O. (1994): Weiterbildung, Arbeitsplatzsuche und individueller Beschäftigungsumfang – Eine ökonometrische Untersuchung für Ostdeutschland. *Zeitschrift für Wirtschafts- und Sozialwissenschaften, 114,* 419-447.

Hübler, O. (1997): Evaluation beschäftigungspolitischer Maßnahmen in Ostdeutschland. *Jahrbücher für Nationalökonomie und Statistik, 216*(1), 21-44.

Hujer, R., Bellmann, L. und C. Brinkmann (2000). Evaluation aktiver Arbeitsmarktpolitik - Probleme und Perspektiven. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 33*(3), 341-344.

Hujer, R., Maurer, K.O. und M. Wellner (1997a). Estimating the Effect of Training on Unemployment Duration in West Germany – A Discrete Hazard Rate Model with Instrumental Variables. *Johann Wolfgang Goethe University Frankfurt, Frankfurter Volkswirtschaftliche Diskussionsbeiträge, Arbeitspapier Nr. 73.*

Hujer, R., Maurer, K.O. und M. Wellner (1997b). The Impact of Training on Unemployment Duration in West Germany - Combining A Discrete Hazard-Rate Model With Matching Techniques. *Johann Wolfgang Goethe University Frankfurt, Frankfurter Volkswirtschaftliche Diskussionsbeiträge, Arbeitspapier Nr. 74.*

Hujer, R., Maurer, K.O. und M. Wellner (1998). Kurz- und langfristige Effekte von Weitebildungsmaßnahmen auf die Arbeitslosigkeitsdauer in Westdeutschland. In F. Pfeiffer und W. Pohlmeier (Hrsg.), *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*, Baden-Baden: Nomos, 197-221.

Hujer, R. und M. Wellner (2000a). The Effects of Public Sector Sponsored Training on Indi-vidual Employment Performance in East Germany. *Forschungsinstitut zur Zukunft der Arbeit (IZA), IZA-Diskussionspapiere Nr.141.*

Hujer, R. und M. Wellner (2000b). Berufliche Weiterbildung und individuelle Arbeits-losigkeitsdauer in West- und Ostdeutschland: Eine mikroökonometrische Analyse. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 33*(3), 405-420.

Imbens, G.W. und Angrist J.D. (1994). Identification and Estimation of Locl Average Treatment Effects. *Econometrica, 62*, 446-475.

Kalbfleisch, J.D. und R.L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York u.a.: Wiley.

Kleinbeck, U. (1996). *Arbeitsmotivation.* Weinheim u.a.: Juventa Verlag.

Kleinbeck, U., Schmidt, K.H., Ernst, G. und J. Rutenfranz (1980). Motivationale Aspekte der Arbeitszufriedenheit. *Zeitschrift für Arbeitswissenschaft, 34*, 200-206.

Klose, C. und S. Bender (2000). Berufliche Weiterbildung für Arbeitslose – ein Weg zurück in Beschäftigung? *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 33*(3), 421-444.

Koch, S. und U. Walwei (2003). *Arbeitsmarkt-Reformen: Per Paket aus der Krise?* Nürnberg: IAB, IAB-Materialien Nr. 4.

Kraus, F., Puhani, P. und V. Steiner (2000). Do Public Works Programes Work in Eastern Germany? *Research in Labor Economics, 19*, 275-314.

Kühnlein, G. und B. Klein (2003). *Bildungsgutscheine – mehr Eigenverantwortung, mehr Markt, mehr Effizienz?* Düsseldorf: Hans-Böckler-Stiftung, Arbeitspapier, Nr. 74.

LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review, 76*, 604-620.

Lechner, M (1996a). *An Evalution of Public Sector Sponsored Continuous Vocational Training Programs in East Germany.* Mannheim: Institut für Volkswirtschaft und Statistik, Beiträge zur angewandten Wirtschaftsforschung, Nr. 539-96.

Lechner, M. (1996b). The Effects of Enterprise-related Continuous Vocational Training in East Germany on Individuel Employment and Earnings. Mannheim: ZEW, ZEW-Diskussionspapiere Nr. 542-96.

Lechner, M. (1998a): *Training the East German Labour Force. Microeconometric Evalua-tions of Continuous Vocational Training after Unification*. Heidelberg: Physica-Verlag.

Lechner, M. (1999): Earnings and Employment Effects of Continuous Off-the-job Training in East Germany after Unification. *Journal of Business Economic Statistics, 17*, 74-90.

Lechner, M. (2000): An Evaluation of Public Sector Sponsored Continuous Vocational Train-ing Programs in East Germany. *The Journal of Human Resources, 35*(2), 347-375.

Lechner, M. (2002a). Eine wirkungsorientierte aktive Arbeitsmarktpolitik in Deutschland und der Schweiz: Eine Vision - zwei Realitäten. *Perspektiven der Wirtschaftspolitik, 3*(2), 159-174.

Lechner, M (2002b). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society - Series A, 165*(1), 59-82.

Lechner, M. (2002c). Mikroökonometrische Evaluation arbeitsmarktpolitischer Massnahmen. *Universität St. Gallen,* Diskussionspapiere Nr. 20/2002.

Lechner, M., Miquel, R. und C. Wunsch (2004a). Long-run Effects of Public Sector Sponsored Training in West Germany. *Universität St. Gallen*: URL am 28.12.2004: http://www.siaw.unisg.ch/org/siaw/web.nsf/SysWebRessources/fuu_lmw_041217/$FILE/LMW-fuu_W_Pap_041217.pdf

Lechner, M., Miquel, R. und C. Wunsch (2004b). Second Appendix to the paper: Long-run Effects of Public Sector Sponsored Training in West Germany. *Universität St. Gallen*: URL am 28.12.2004: http://www.siaw.unisg.ch/org/siaw/web.nsf/SysWebRessources/lmw_fuuw_ia/$FILE/fuu_lmw_internet_appendix_171204.pdf

Lewin, K. (1920). Die Sozialisierung des Taylorsystems. *Schriftenreihe Praktischer Sozialis-mus, 4*, 3-36.

Lienert, G.A. (1989). *Testaufbau und Testanalyse*. München u.a.: Psychologie Verlags Union.

Maslow, A.H. (1954). Motivation and Personality. New York: Harper and Row.

May, S. und D.W. Hosmer (1998). A Simplified Method of Calculating an Overall Goodness-of-Fit Test for the Cox Proportional Hazards Model. *Lifetime Data Analysis, 4*, 109-120.

May, S. und D.W. Hosmer (2004a). A Cautionary Note on the Use of the Grønnesby and Borgan Goodness-of-fit Test for the Cox Proportional Hazards Model. *Lifetime Data Analysis, 10*, 283-291.

May, S. und D.W. Hosmer (2004b). Hosmer and Lemeshow type of Goodness-of-fit Statistics for the Cox Proportional Hazards Model. In N. Balaskrishnan und C.R. Rao (Hrsg.), *Handbook of Statistics 23 – Advances in Survival Analysis*, Amsterdam: North Holland, 383-394.

Miquel, R., Wunsch C. und M. Lechner (2002). Die FuU-Teilnehmer-Datei 1976-1997, *Graues Papier des Instituts für Arbeitsmarkt- und Berufsforschung*, Nürnberg.

Pannenberg, M. (1995). Weiterbildungsaktivitäten und Erwerbsbiographie. Eine empirische Analyse für Deutschland. Frankfurt/Main: Campus Verlag.

Projektgruppe Sozio-ökonomisches Panel (1995). Das Sozio-ökonomische Panel (SOEP) im Jahr 1994. *DIW – Vierteljahreshefte zur Wirtschaftforschung, 64*(1), 5-15.

Prey, H. (1997). *Beschäftigungswirkungen von öffentlich geförderten Qualifizierungsmaßnahmen. Eine Paneluntersuchung für Westdeutschland*. Konstanz: Center for International Labor Economics (CILE), Diskussionspapier, 41/1997.

Prey, H. (1999). *Wirkungen staatlicher Qualifizierungsmaßnahmen.* Bern u.a.: Paul Haupt.

Reinowski, E., Schultz, B. und J. Wiemers (2003). *Evaluation von Maßnahmen der aktiven Arbeitsmarktpolitik mit Hilfe eines iterativen Matching-Algorithmus – Eine Fallstudie über langzeitarbeitslose Maßnahmeteilnehmer in Sachsen*. Halle: Institut für Wirtschaftsforschung Halle (IHW), IHW-Diskussionspapiere 173/2003.

Reinowski, E., Schultz, B. und J. Wiemers (2004). Evaluation of Further Programmes with an Optimal Matching Algorithm. Halle: IHW, IHW-Diskussionspapiere 188/ 2004.

Rosenbaum, P.R. und D.B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41-55.

Rosenbaum, P.R. und D.B. Rubin (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistican, 39*, 33-38.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Auflage). Bern u.a.: Huber Verlag.

Roy, A.D. (1951). Some Thoughts on the Distribution of Income. *Oxford Economic Papers, 2*, 135-146.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D.B. (1977). Assignment to a Treatment Group on the Basis of a Covariate: *Journal of Educational Statistics, 2*, 1-26.

Rubin, D.B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects And the Critical Role of the Assignment Mechanism. *Biometrica, 47*, 1213-1234.

Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2000). *Jahresgutachten 2000/01: Chancen auf einen höheren Wachstumspfad*. Stuttgart: Metzler-Poeschel.

Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2002). *Jahresgutachten 2002/03: Zwanzig Punkte für Beschäftigung und Wachstum*. Stuttgart: Metzler-Poeschel.

Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2004). *Jahresgutachten 2004/05: Erfolge im Ausland - Herausforderungen im Inland.* URL am 23.12.2004:
http://www.sachverstaendigenrat-wirtschaft.de

Schönfeld, D. (1982). Partial residuals for the proportional hazards regression Model. *Biometrika, 69*, 239-241.

Schmidt, K.H., Kleinbeck, U., Ottmann, W. und B. Seidel (1985). Ein Verfahren zur Diagnose von Arbeitsinhalten: Der Job Diagnostic Survey (JDS). *Psychologie und Praxis, 29*, 162-172.

Schuler, H. und M. Prochaska (2000). Entwicklung und Konstruktvalidierung eines berufsbe-zogenen Leistungsmotivationstest. *Diagnostica, 46*(2), 61-72.

Singer J.D. und J.B Willet (2003). *Applied Longitudinal Data Analysis*. Oxford u.a: Oxford University Press.

Spitznagel, E. und K. Vogler-Ludwig (2004). *Wachstumsschwäche: Stellenangebot und Personalmangel nehmen weiter ab.* Nürnberg: IAB, IAB-Kurzbericht Nr. 8.

Smith, J.A. und P.E. Todd (2004). Does Matching Overcome Lalonde´s Critiquw of Nonex-perimental Estimators? *Journal of Econometrics,* forthcoming. URL am 06.12.2004: http://papers.ssrn.com/abstract=28629

Staat, M. (1997). *Empirische Evaluation von Fortbildung und Umschulung*. Baden-Baden: Nomos Verlagsgesellschaft.

Wooldridge, J.M. (2000). *Introductory Econometrics: A Modern Approach*. Cincinnati u.a.: South-Western College Publishing.