



BLICKPUNKT ARBEIT UND WIRTSCHAFT

10/2006

10. Juli 2006

Arbeitsmarktpolitik: Innovative Evaluierungs- und Steuerungsansätze mit TrEffeR



Dr. Bruno Kaltenborn
Kaltenborn@wipol.de
Tel. 030/400 43 58-8



Kathrin Göggel
Goeggel@wipol.de
Tel. 030/400 43 58-5

Mitarbeiterin
im Team Dr. Kaltenborn

Fotos: Silke Rudolph
u. Fotostudio Thomas

Einleitung

Im Auftrag der Bundesregierung wird die Umsetzung des Ersten, Zweiten und Dritten Gesetzes für moderne Dienstleistungen am Arbeitsmarkt (sog. Hartz-Reformen) eingehend evaluiert. Acht Forschungskonsortien mit insgesamt etwa 20 namhaften Forschungseinrichtungen wurden beauftragt und vom Team Dr. Kaltenborn koordiniert. Parallel dazu wird derzeit in der Bundesagentur für Arbeit ein System quantitativer Wirkungsanalysen arbeitsmarktpolitischer Maßnahmen unter der Bezeichnung „Treatment Effects and Prediction“ (TrEffeR) aufgebaut. In beiden Projekten werden zahlreiche arbeitsmarktpolitische Maßnahmen hinsichtlich ihrer Arbeitsmarktwirkungen untersucht.

Nunmehr liegen aus beiden Projekten erste Befunde vor (vgl. Tabelle 1). Die zentralen Zwischenergebnisse der Hartz-Evaluierung wurden in dieser Reihe bereits dargestellt. Diese Ausgabe berichtet vor diesem Hintergrund über Vorhaben und erste retrospek-

tive Evaluierungsergebnisse des TrEffeR-Projekts, teilweise im Vergleich zur Hartz-Evaluierung.

Untersuchungsansatz

Beide Projekte evaluieren verschiedene arbeitsmarktpolitische Instrumente retrospektiv. Dabei konzentriert sich die Hartz-Evaluierung auf solche Instrumente, die von den Hartz-Reformen betroffen waren. TrEffeR dagegen berücksichtigt grundsätzlich alle arbeitsmarktpolitischen Instrumente der Bundesagentur für Arbeit. Dies liegt nahe, da TrEffeR überdies künftig anhand von Prognosen im Einzelfall eine Orientierungshilfe für Vermittler/innen der Bundesagentur für Arbeit für die Zuweisung in Maßnahmen bereitstellen soll. Während die Hartz-Evaluierung fast abgeschlossen ist, soll mit TrEffeR eine dauerhafte Unterstützung der Vermittlung erreicht werden.

Allein die Hartz-Evaluierung beinhaltet Implementations- und Makroanalysen sowie ein Benchmarking. Anhand der in beiden Projekten durchgeführten Mikroanalysen mit anonymisierten Individualdaten können Aussagen zur Integrationswirkung von Instrumenten der aktiven Arbeitsmarktpolitik bei Geförderten in Relation zu einer ungeförderten Vergleichsgruppe getroffen werden. Idealerweise würde zu jeder geförderten Person eine ungeförderte Vergleichsperson mit gleichen Charakteristika verwendet. Eine solche Person ist jedoch meist nicht vorhanden, daher wird auf eine Vergleichsperson zurückgegriffen, die statistisch ähnlich ist. Auf diese Mikroanalysen und ihre Ergebnisse wird im Folgenden näher eingegangen. Der Anhang informiert ergänzend über alternative Vergleichsgruppenkonzepte.

Mikroanalysen

TrEffeR soll im Gegensatz zur Hartz-Evaluierung nicht nur Aussagen über durchschnittliche Maßnahmenwirkungen treffen, sondern auch Wirkungen alternativer Maßnahmen im Einzelfall prognostizieren. Daher benötigt TrEffeR Daten von möglichst allen Personen in Maßnahmen, um auch für atypische Einzelfälle statistisch valide Prognosen treffen zu können. Angesichts des Umfangs erforderlicher Daten scheiden Befragungen

BLICKPUNKT ARBEIT UND WIRTSCHAFT bietet Entscheidungsträger/innen kompakte und systematische Auswertungen von Ideen und Erkenntnissen aus Wissenschaft, Politik und Praxis. Dabei liegt der Fokus auf dem Themenfeld Arbeitsmarkt.

Tabelle 1: Charakteristika der Hartz-Evaluierung und des TrEffeR-Projekts

Kriterium	Hartz-Evaluierung	TrEffeR-Projekt
Ziel	retrospektive Evaluation der <u>durchschnittlichen</u> Wirkung arbeitsmarktpolitischer Instrumente, die von den Hartz-Reformen betroffen waren	1. retrospektive Evaluation der <u>durchschnittlichen</u> Wirkung arbeitsmarktpolitischer Instrumente 2. prospektive Orientierungshilfe bei der Zuweisung in Maßnahmen für Vermittler/innen <u>im Einzelfall</u>
Untersuchungsmethoden	Implementations-, Mikro- und Makroanalysen sowie Benchmarking	Mikroanalysen
Projektstand	fast beendet	im Aufbau, erste Ergebnisse verfügbar
Dauerhaftigkeit	einmaliges Projekt (Zwischenbericht 2005, Endbericht 2006)	permanent
Mikroanalysen		
Daten	anonymisierte Individualdaten der Bundesagentur für Arbeit sowie Befragungsdaten	anonymisierte Individualdaten der Bundesagentur für Arbeit
Stichprobengröße	„kleine“ Stichprobe (wegen Befragungen)	„große“ Stichprobe
untersuchte Instrumente	reformierte und neue arbeitsmarktpolitische Instrumente (u.a. alle „großen“ Instrumente) und beschäftigungspolitische Maßnahmen	(fast) alle arbeitsmarktpolitischen Instrumente, die von der Bundesagentur für Arbeit administriert werden
Erfolgskriterium	insb. Arbeitsmarktintegration	keine faktische Arbeitslosigkeit, d.h. weder arbeitslos noch Maßnahmeteilnahme (Arbeitsmarktintegration ohne Befragungen nicht zeitnah verfügbar)
Vergleichsgruppe(n) (VG) (vgl. Anhang)	Arbeitslose, die während des gesamten Beobachtungszeitraums nicht an der untersuchten, möglicherweise aber an einer anderen Maßnahme teilgenommen haben	a) Arbeitslose, die während des gesamten Beobachtungszeitraums an keiner Maßnahme teilgenommen haben b) wie nebenstehend c) Arbeitslose, die nicht zum gleichen Zeitpunkt wie die geförderten Zielpersonen eine Maßnahme begonnen haben, möglicherweise aber später an der untersuchten oder einer anderen Maßnahme teilgenommen haben
Verfahren der VG-Bildung	<i>Propensity Score Matching</i>	alternativ mehrere Matching-Verfahren (<i>Propensity Score</i> , genetisch)
Differenzierungsmöglichkeiten	insb. Geschlecht, Ost/West, Alter, Maßnahmengattung (teilweise kombiniert)	feinere Differenzierungen nach Personengruppen, regional und nach Maßnahmengattungen auch kombiniert möglich und notwendig
Aktualität	zeitnah nach Ende eines festgelegten Beobachtungszeitraums (nach Maßnahmeteilnahme)	grundsätzlich zeitnah nach Ende eines festgelegten Beobachtungszeitraums (nach Maßnahmeteilnahme); derzeit nur Beobachtungszeiträume bis Ende 2004 möglich
Publikationen	<u>Zentrale Ergebnisse</u> : <i>BLICKPUNKT ARBEIT UND WIRTSCHAFT</i> 1-4/2006; <u>Zwischenbericht 2005</u> : Deutscher Bundestag [2006]; <u>8 Einzelberichte der Evaluatoren</u> : Links unter http://www.wipol.de/hartz	STEPHAN, RÄSSLER und SCHEWE [2006]; RÄSSLER [2006]; STEPHAN, SCHEWE und RÄSSLER [2006]

Quelle: Eigene Darstellung auf Basis der angegebenen Publikationen.

aus Kostengründen aus; TrEffeR kann deshalb nur auf anonymisierte Individualdaten der Bundesagentur für Arbeit zurückgreifen. Daraus lässt sich eine Arbeitsmarktintegration nur unvollständig und nur mit einem Zeitverzug von etwa anderthalb Jahren nach Ende des Kalenderjahres nachvollziehen. Daher wird als Erfolg angesehen, wenn eine Person weder arbeitslos ist noch an einer Maßnahme teilnimmt (keine „faktische“ Arbeitslosigkeit); dies bedeutet nicht notwendigerweise auch eine Arbeitsmarktintegration. Da TrEffeR als dauerhaftes Evaluations- und Steuerungsinstrument fungieren soll, ist eine weitgehende Standardisierung der Methode über verschiedene Maßnahmen und Personengruppen erforderlich. Ab 2005 bestehen überdies wegen der Einführung der Grundsicherung für Arbeitsuchende in den sog. Optionskommunen erhebliche Datenlücken, so dass Analysen derzeit nur bis Ende 2004 möglich sind.

Die Hartz-Reformen zielten auf eine bessere Arbeitsmarktintegration. Diese muss deswegen als Erfolgskriterium bei der Hartz-Evaluierung herangezogen werden. Die hierfür erforderlichen Befragungen begrenzen aus Kostengründen den Umfang der betrachteten Förderfälle. Eine kleine Stichprobe erlaubt nur wenige kombinierte Differenzierungen etwa nach Geschlecht, Region und Alter. Die Hartz-Evaluierung kann eher als TrEffeR flexibel Besonderheiten von Maßnahmetypen berücksichtigen.

Derzeit werden bei TrEffeR zwei alternative Erfolgskriterien betrachtet: (1) keine faktische Arbeitslosigkeit zu einem Stichtag und (2) die Dauer, in der im Beobachtungszeitraum keine faktische Arbeitslosigkeit vorliegt. Die Interpretation ist unterschiedlich, die abgebildeten Wirkungen sind jedoch identisch.

Bei TrEffeR werden alternativ mehrere Matching-Verfahren eingesetzt, um ihre statistische Qualität im Vergleich besser beurteilen zu können. Den Vorzug erhält ein Verfahren, das im Vergleich zum „Besten“ nur eine leicht geringere statistische Qualität und ähnliche Ergebnisse bei deutlich geringerem Rechenaufwand aufweist. Die Merkmale, die beim Matching berücksichtigt werden, werden so ausgedrückt, dass sie nur zwei Ausprägungen („Ja“ und „Nein“) haben (sog. Dummy-Variablen). Beispielsweise geht nicht das Alter in Jahren ein, sondern es werden mehrere überschneidungsfreie Altersgruppen definiert, wobei eine Person jeweils zu einer Altersgruppe dazugehört („Ja“) oder nicht („Nein“). Bei einem solchen Vorgehen muss kein spezifischer Zusammenhang zwischen den zugrunde gelegten Merkmalen und der Maßnahmeteilnahme unterstellt werden. Würde z.B. direkt

das Alter in Jahren eingehen, so wäre damit die nicht überprüfbare Annahme eines linearen Zusammenhangs verbunden; dann muss ein/e 40-Jährige/r eine doppelt so hohe Wahrscheinlichkeit wie ein/e 20-Jährige/r haben, an einer bestimmten Maßnahme teilzunehmen.

TrEffeR ist insgesamt bemüht, einseitige realitätsferne Annahmen zu vermeiden (alternative Matching-Verfahren und Vergleichsgruppen, Verwendung von Dummy-Variablen). Ergänzend wäre wünschenswert, auch Alternativen zum Matching anzuwenden, beispielsweise ein (weiterentwickeltes) Selektionsmodell nach HECKMAN [1979]. Damit können die vermutlich oft bestehenden *unbeobachteten* systematischen Unterschiede zwischen Geförderten und Vergleichsgruppe kontrolliert und so eine Verzerrung der Ergebnisse vermieden werden. So können etwa unbeobachtbare systematische Motivationsunterschiede zwischen den beiden Gruppen bestehen.

Zentrale Evaluierungsergebnisse

Mit dem TrEffeR-Modell wurden wie auch bei der Hartz-Evaluierung bisher mit Überbrückungsgeld geförderte Existenzgründer/innen, Arbeitsbeschaffungsmaßnahmen sowie mit einem Eingliederungszuschuss (Lohnkostenzuschuss) geförderte Beschäftigte und einzelne Maßnahmetypen der Förderung beruflicher Weiterbildung untersucht. Insgesamt bestätigen hier die TrEffeR-Ergebnisse die Zwischenergebnisse der Hartz-Evaluierung. Darüber hinaus wurden mit dem TrEffeR-Modell auch Trainingsmaßnahmen mit einer Dauer von bis zu acht Wochen analysiert.

Nach den TrEffeR-Ergebnissen sind mit Überbrückungsgeld geförderte Existenzgründer/innen (6 Monate Förderung) ebenso wie mit einem Eingliederungszuschuss bei Einarbeitung (4 bis 6 Monate Förderung) bzw. bei erschwelter Vermittlung (7 bis 12 Monate Förderung) geförderte Beschäftigte nach Ende der Förderung seltener faktisch arbeitslos als die jeweilige Vergleichsgruppe. Hingegen führen Arbeitsbeschaffungsmaßnahmen auch noch zwei Jahre nach ihrem Beginn zu einem höheren Risiko, faktisch arbeitslos zu sein.

Die Förderung beruflicher Weiterbildung wirkt je nach Maßnahme und Zielgruppe heterogen. So führen Bildungsmaßnahmen zur Qualifikationserweiterung mit einer Dauer von vier bis sechs Monaten in Ostdeutschland bereits nach sechs Monaten zu einer Reduktion des Risikos, faktisch arbeitslos zu sein. In Westdeutschland haben sie jedoch in den ersten zwei Jahren nach Maßnahmebeginn keine nachweisbare

Wirkung. Längere Maßnahmen mit bis zu einem Jahr wirken bei Männern in Ostdeutschland positiv.

Auch die Wirkung von Trainingsmaßnahmen ist heterogen. Während mit betrieblichen Maßnahmen Geförderte bereits sehr kurzfristig ein nachhaltig geringeres Risiko haben, faktisch arbeitslos zu sein, sind derartige Effekte bei außerbetrieblichen Trainingsmaßnahmen zur Vermittlung von Kenntnissen oder zur Überprüfung der Verfügbarkeit nicht feststellbar bzw. teilweise sogar negativ.

Ausblick: Orientierungshilfe für die Vermittlung

TrEffeR soll künftig - neben der retrospektiven Evaluierung - anhand von Prognosen im Einzelfall für Vermittler/innen der Bundesagentur für Arbeit eine Orientierungshilfe für die Zuweisung in Maßnahmen bereitstellen. Hierfür sind statistisch zuverlässige Aussagen über die Arbeitsmarktwirkung von Maßnahmen auch im ggf. atypischen Einzelfall erforderlich. Hierfür steht noch eine methodische Lösung aus. Dabei operiert TrEffeR bereits jetzt auf einem hohen methodischen Niveau.

Sowohl für die Orientierungshilfe als auch für eine fortlaufende retrospektive Evaluierung ab 2005 müssen noch die in den Optionskommunen bestehenden Datenlücken geschlossen werden.

Für den Einsatz der Orientierungshilfe kommt es darauf an, den Vermittlungsfachkräften den Nutzen zu verdeutlichen und die prognostische Qualität plausibel zu machen. Überdies sollte - wie geplant - gewährleistet werden, dass TrEffeR-Ergebnisse den Maßnahmeeinsatz nicht vorgeben, sondern die Entscheidungen vor Ort nur unterstützen. Es liegt nahe, die Orientierungshilfe zunächst regional begrenzt umzusetzen und ihre Wirksamkeit zu prüfen.

Schließlich könnten TrEffeR-Ergebnisse auch in Zielvereinbarungen einbezogen werden (vgl. hierzu *BLICKPUNKT ARBEIT UND WIRTSCHAFT 9/2006*).

Literatur

Deutscher Bundestag [2006]: „Bericht 2005 der Bundesregierung zur Wirksamkeit moderner Dienstleistungen am Arbeitsmarkt“, Unterrichtung durch die Bundesregierung, *Bundestagsdrucksache*, 16/505, 1. Februar 2006, Berlin.

HECKMAN, JAMES J. [1979]: „Sample Selection Bias as a Specification Error“, *Econometrica*, Vol. 47, No. 1, Januar 1979, S. 153-161.

RÄSSLER, SUSANNE [2006]: „Der Einsatz von Missing Data Techniken in der Arbeitsmarktforschung des IAB“, *Allgemeines Statistisches Archiv*, Jg. 90, im Erscheinen.

STEPHAN, GESINE, SUSANNE RÄSSLER und TORBEN SCHEWE [2006]: *Wirkungsanalyse in der Bundesagentur für Arbeit*, Konzeption, Datenbasis und ausgewählte Befunde, IAB Discussion Paper, Nr. 7/2006, Nürnberg.

STEPHAN, GESINE, TORBEN SCHEWE und SUSANNE RÄSSLER [2006]: „The effectiveness of German labor market programs using different control group concepts“, unveröffentlichte vorläufige Fassung, April 2006, Nürnberg.

Anhang: Alternative Vergleichsgruppenkonzepte

Die TrEffeR-Untersuchungen analysieren den Erfolg bei der Zielgruppe jeweils in Relation zu drei unterschiedlichen Vergleichsgruppen (vgl. Tabelle 1). Die Richtung der ermittelten Wirkungen ist bei fast allen Instrumenten für die drei verwendeten Vergleichsgruppen identisch, die Stärke ist jedoch unterschiedlich. Auch die Interpretation ist abweichend.

Die Vergleichsgruppen unterscheiden sich danach, ob die Vergleichspersonen während des Beobachtungszeitraums

- a) keine Maßnahme (Nicht-Partizipation),
- b) eine andere Maßnahme (Warten auf eine etwaige andere Maßnahme) oder
- c) eine andere oder sogar die gleiche Maßnahme (Warten auf eine etwaige Maßnahme)

wie die geförderte Zielperson begonnen haben können. Während im Fall (a) die Wirkung einer arbeitsmarktpolitischen Maßnahme im Vergleich zu vollständig fehlendem Einsatz des arbeitsmarktpolitischen Instrumentariums ermittelt wird, wird im Fall (c) der Einsatz einer Maßnahme zu einem bestimmten Zeitpunkt im Vergleich zum weiteren Abwarten mit einem etwaigen späteren Maßnahmeeinsatz verglichen. Der Fall (b), der auch bei der Hartz-Evaluierung angewendet wird, liegt dazwischen, jedoch näher an Fall (c).

Die Instrumente schneiden - zunächst überraschend - durchgehend im Fall (c) besser ab als im Fall (a). Hierfür kommen mehrere Ursachen in Betracht:

- Der spätere Einsatz arbeitsmarktpolitischer Instrumente bei den Vergleichspersonen könnte mit einer temporären Verlängerung der faktischen Arbeitslosigkeit einhergehen. Dies würde die Arbeitsmarktchancen der Vergleichsgruppe beeinträchtigen. Dadurch schneiden die Instrumente - berechtigterweise - im Fall (c) besser als im Fall (a) ab.

- Vergleichspersonen könnten deswegen später nie an einer Maßnahme teilnehmen, weil sie *ex-ante* systematisch bessere Arbeitsmarktchancen haben. Dies würde eine systematische Verzerrung der Ergebnisse im Fall (a) nach unten bedeuten, die inhaltlich nicht interpretiert werden kann.
- Im Fall (a) werden bei den Vergleichspersonen anders als bei den Zielpersonen für deren Auswahl auch *ex-post*-Informationen über eine spätere Maßnahmeteilnahme verwendet. Maßnahmeteilnahme gilt als faktische Arbeitslosigkeit und damit als Misserfolg. Überdies ist typischerweise eine Maßnahmeteilnahme mit vorhergehender Arbeitslosigkeit verbunden. Werden - wie im Fall (a) - *ex-post* spätere Maßnahmeteilnehmer/innen aus der Vergleichsgruppe ausgeschlossen, so betrifft dies systematisch Personen mit längerer faktischer Arbeitslosigkeit. Mithin schneiden im Fall (a) die Vergleichsgruppe systematisch besser und das betrachtete Instrument schlechter ab.¹ Diese Verzerrung kann inhaltlich nicht sinnvoll interpretiert werden.

Während die beiden erstgenannten Begründungen nur unter spezifischen Bedingungen gelten, ist die letztgenannte in jedem Fall relevant.

Insgesamt ist der Fall (a) der inhaltlich interessantere, weil er den Einsatz mit dem Nicht-Einsatz arbeitsmarktpolitischer Instrumente vergleicht, jedoch sind die Ergebnisse hierfür verzerrt (die Wirkungen scheinen schlechter als sie tatsächlich sind). Hingegen sind die Ergebnisse für Fall (c) nicht von diesen Verzerrungen betroffen, allerdings vergleichen sie lediglich den Beginn einer Maßnahme zu einem bestimmten Zeitpunkt mit einem etwaigen späteren Maßnahmeinsatz statt mit dem Nicht-Einsatz. Wünschenswert wäre ein verzerrungsfreier Vergleich zwischen Einsatz und Nicht-Einsatz einer Maßnahme. Die Hartz-Evaluierung hat als Kompromiss zwischen den beiden Ansätzen den Fall (b) gewählt.

Überdies ist bei allen Vergleichsgruppen zu beachten, dass beim Überbrückungsgeld und den Eingliederungszuschüssen die Förderung untrennbar mit einer Erwerbstätigkeit im ersten Arbeitsmarkt verbunden ist. Daher unterscheiden sich die Vergleichspersonen nicht nur im Hinblick auf eine Maßnahmeteilnahme systematisch von den Zielpersonen, sondern auch hinsichtlich einer Erwerbstätigkeit im ersten Arbeitsmarkt. Deshalb muss beim Überbrückungsgeld und bei den Eingliederungszuschüssen offen bleiben, inwieweit für die Unterschiede zwischen Ziel- und Vergleichspersonen die Förderung und inwieweit die Erwerbstätigkeit im ersten Arbeitsmarkt verantwortlich ist.

¹ Zur Entwicklung dieser Begründung war ein Anfang 2006 gehaltener Vortrag von Dr. Hilmar Schneider (IZA) im Zusammenhang mit der Hartz-Evaluierung hilfreich.

BLICKPUNKT ARBEIT UND WIRTSCHAFT

Nr. 10/2006, 10. Juli 2006:

Kaltenborn, Bruno, und Kathrin Göggel

Arbeitsmarktpolitik: Innovative Evaluierungs- und Steuerungsansätze mit TrEffeR

Nr. 9/2006, 10. Juli 2006:

Kaltenborn, Bruno

Hartz IV: Zentrale Steuerung und lokale Autonomie

Nr. 8/2006, 6. April 2006:

Kaltenborn, Bruno, Petra Knerr und

Juliana Schiwarov

Hartz IV: Leistungen von Arbeitsgemeinschaften und Optionskommunen

Nr. 7/2006, 6. April 2006

(Aktualisierung von Nr. 3/2005):

Kaltenborn, Bruno, und Juliana Schiwarov

Hartz IV: Föderaler Finanzstreit vorerst beigelegt

Nr. 6/2006, 6. April 2006

(Aktualisierung von Nr. 2/2005):

Kaltenborn, Bruno, und Juliana Schiwarov

Hartz IV: Ausgaben deutlich unterschätzt

Nr. 5/2006, 6. April 2006

(Aktualisierung von Nr. 1/2005):

Kaltenborn, Bruno, und Juliana Schiwarov

Hartz IV: Deutlich mehr Fürsorgeempfänger/innen

Nr. 4/2006, Februar 2006:

Kaltenborn, Bruno, Petra Knerr und

Juliana Schiwarov

Agenturen für Arbeit: Systematisierung des Ressourceneinsatzes

Nr. 3/2006, Februar 2006:

Kaltenborn, Bruno, Petra Knerr und

Juliana Schiwarov

Hartz: Bilanz der Arbeitsmarkt- und Beschäftigungspolitik

Impressum

BLICKPUNKT ARBEIT UND WIRTSCHAFT, Jg. 2, Nr. 10/2006

Internet: <http://www.wipol.de>

Herausgeber: Dr. Bruno Kaltenborn

Erscheinungsweise: unregelmäßig

Erscheinungsort: Berlin

ISSN 1861-9436

Alle Rechte vorbehalten.