Sonderdruck aus:

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung

Jeffrey Smith

Evaluation aktiver Arbeitsmarktpolitik: Erfahrungen aus Nordamerika

33. Jg./2000 **3**

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (MittAB)

Die MittAB verstehen sich als Forum der Arbeitsmarkt- und Berufsforschung. Es werden Arbeiten aus all den Wissenschaftsdisziplinen veröffentlicht, die sich mit den Themen Arbeit, Arbeitsmarkt, Beruf und Qualifikation befassen. Die Veröffentlichungen in dieser Zeitschrift sollen methodisch, theoretisch und insbesondere auch empirisch zum Erkenntnisgewinn sowie zur Beratung von Öffentlichkeit und Politik beitragen. Etwa einmal jährlich erscheint ein „Schwerpunktheft", bei dem Herausgeber und Redaktion zu einem ausgewählten Themenbereich gezielt Beiträge akquirieren.

Hinweise für Autorinnen und Autoren

Das Manuskript ist in dreifacher Ausfertigung an die federführende Herausgeberin
Frau Prof. Jutta Allmendinger, Ph. D.
Institut für Arbeitsmarkt- und Berufsforschung
90478 Nürnberg, Regensburger Straße 104
zu senden.

Die Manuskripte können in deutscher oder englischer Sprache eingereicht werden, sie werden durch mindestens zwei Referees begutachtet und dürfen nicht bereits an anderer Stelle veröffentlicht oder zur Veröffentlichung vorgesehen sein.

Autorenhinweise und Angaben zur formalen Gestaltung der Manuskripte können im Internet abgerufen werden unter http://doku.iab.de/mittab/hinweise_mittab.pdf. Im IAB kann ein entsprechendes Merkblatt angefordert werden (Tel.: 09 11/1 79 30 23, Fax: 09 11/1 79 59 99; E-Mail: ursula.wagner@iab.de).

Redaktion

Ulrike Kress, Gerd Peters, Ursula Wagner, in: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), 90478 Nürnberg, Regensburger Str. 104, Telefon (09 11) 1 79 30 19, E-Mail: ulrike.kress@iab.de: (09 11) 1 79 30 16, E-Mail: gerd.peters@iab.de: (09 11) 1 79 30 23, E-Mail: ursula.wagner@iab.de: Telefax (09 11) 1 79 59 99.

Herstellung

Satz und Druck: Tümmels Buchdruckerei und Verlag GmbH, Gundelfinger Straße 20, 90451 Nürnberg

Verlag

W. Kohlhammer GmbH, Postanschrift: 70549 Stuttgart: Lieferanschrift: Heßbrühlstraße 69, 70565 Stuttgart: Telefon 07 11/78 63-0;
Telefax 07 11/78 63-84 30: E-Mail: waltraud.metzger@kohlhammer.de, Postscheckkonto Stuttgart 163 30.
Girokonto Städtische Girokasse Stuttgart 2 022 309.
ISSN 0340-3254

Bezugsbedingungen

Die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" erscheinen viermal jährlich. Bezugspreis: Jahresabonnement 52,- € inklusive Versandkosten: Einzelheft 14,- € zuzüglich Versandkosten. Für Studenten, Wehr- und Ersatzdienstleistende wird der Preis um 20 % ermäßigt. Bestellungen durch den Buchhandel oder direkt beim Verlag. Abbestellungen sind nur bis 3 Monate vor Jahresende möglich.

Zitierweise:

MittAB = „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" (ab 1970)
Mitt(IAB) = „Mitteilungen" (1968 und 1969)
In den Jahren 1968 und 1969 erschienen die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" unter dem Titel „Mitteilungen", herausgegeben vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit.

Internet: http://www.iab.de

Evaluation aktiver Arbeitsmarktpolitik: Erfahrungen aus Nordamerika

*Jeffrey Smith**

Dieser Beitrag untersucht die Lehren, die in Amerika mit der Evaluation von aktiven Arbeitsmarktprogrammen gemacht wurden und die möglicherweise auf europäische Staaten übertragen werden könnten, die auf diesem Gebiet erst in jüngster Zeit aktiv werden. Dabei werde ich zwei Fragestellungen der Evaluation betrachten. Die erste Fragestellung betrifft die institutionellen Rahmenbedingungen, innerhalb derer Evaluation stattfindet: Welche Individuen, Unternehmen oder Organisationen führen Evaluationen durch, wer sind die Auftraggeber und wer evaluiert - explizit oder implizit - die Evaluatoren. Ich werde argumentieren, dass diese institutionellen Rahmenbedingungen eine entscheidende Rolle dabei spielen, ob Evaluationen objektiv durchgeführt werden und somit der Politik und öffentlichen Meinung von Nutzen sein können.

Die zweite Fragestellung betrifft die Wahl zwischen verschiedenen ökonometrischen Evaluationsmethoden. Ich werde dabei insbesondere auf die Notwendigkeit eingehen, bei der Wahl einer ökonometrischen Methode und der Interpretation der Ergebnisse berücksichtigen zu müssen, dass der Einfluss eines Programms zwischen verschiedenen Individuen variieren kann. Ich werde dann Kosten und Nutzen von sozialen Experimenten und die Einsatzmöglichkeiten der jüngeren Propensity Score Matching Methoden betrachten. Zum Abschluss werde ich auf die Bedeutung von allgemeinen Gleichgewichtseffekten in der Evaluationsforschung eingehen.

Gliederung

1 Einführung

Für jemanden, der mit Evaluationsforschung beschäftigt ist, ist dies eine aufregende Zeit. Die Kapitel von Heckman, LaLonde und Smith (1999) und Angrist und Krueger (1999) in der neuesten Auflage des *Handbook of Labor Economics* geben einen Überblick über die rasante Entwicklung auf diesem Gebiet und die dadurch ausgelöste lebhafte intellektuelle Diskussion. Auch auf Seiten der Politik ist eine außerordentliche Begeisterung festzustellen. Regierungen, die bisher zum größten Teil eine ernsthafte Evaluation ihrer aktiven Arbeitsmarktpolitik verhindert haben, sind zunehmend an diesem Gebiet interessiert.

Dieser Beitrag präsentiert einige Bemerkungen, die mit zwei wichtigen Fragestellungen der Evaluation aktiver Arbeitsmarktpolitik zusammenhängen. Die erste Fragestellung, die die industrielle Organisationsstruktur der Evaluationsindustrie betrifft, wird zwar selten diskutiert, ist aber nichtsdestotrotz mitentscheidend für den langfristigen Erfolg von Evaluationsmaßnahmen und ihren Möglichkeiten, der Politik als Entscheidungshilfe zu dienen.

Die zweite traditionellere Fragestellung betrifft die neueren Entwicklungen ökonometrischer Methoden, die in der Evaluationsforschung angewandt werden. Dieser zweite Teil baut auf Arbeiten von James Heckman, mir und anderen auf, die in Heckman, LaLonde und Smith (1999) dargestellt werden.

Bezüglich der ökonometrischen Methoden weise ich auf vier wesentliche Punkte hin. Zunächst werde ich argumentieren, dass es wichtig ist, sowohl in Bezug auf wirtschaftspolitische Empfehlungen als auch aus ökonometrischer Sicht, interpersonelle Unterschiede in den Effekten von Programmen zu berücksichtigen. Zweitens diskutiere ich Stärken und Schwächen von sozialen Experimenten, wobei ich zu dem Schluss komme, dass Experimente ein wichtiges Instrument der Evaluation darstellen, das weder vorschnell abgelehnt noch unkritisch akzeptiert werden sollte.

Drittens diskutiere ich die neuesten Entwicklungen bei den nichtexperimentellen Matching-Methoden, deren Mittelpunkt die Technik des Propensity Score-Matchings bildet. Ich weise darauf hin, dass Matching höchst sinnvoll sein kann, wenn die Daten die Anwendung rechtfertigen, dass es jedoch auch auf sehr restriktiven Annahmen aufbaut.

Es setzt ferner auf Seiten des Forschers eine schwierige Einschätzung über die Matching-Variablen voraus, die letztlich nur durch Rückgriff auf a-priori vorhandenes oder theoreti-

* Jeffrey Smith ist Professor für Wirtschaftswissenschaften an der Universität von Western Ontario in London, Ontario, Kanada. Seine Forschungsschwerpunkte umfassen Evaluation von sozialen Programmen, wie z. B. Job Training für Benachteiligte. Er hat darüber hinaus auch wissenschaftliche Beiträge zu Themen wie dem Einfluß einer universitären Ausbildung auf die Arbeitsmarktchancen oder über die Möglichkeiten, Teilnehmer an öffentlichen Programmen mit Hilfe von Profilingmethoden auszuwählen. Der Beitrag liegt in der alleinigen Verantwortung des Autors.

Der Autor dankt Dan Black, Michael Lechner und Miana Plesca für wertvolle Kommentare.

sches Wissen erfolgen kann. Obwohl Matching-Verfahren sich in der letzten Zeit in der angewandten Literatur einer zunehmenden Beliebtheit erfreuen, stellen sie somit nicht den „Königsweg" dar und vermögen das Evaluationsproblem nicht in jeder Situation zu lösen.

Abschließend wende ich mich dem Problem der allgemeinen Gleichgewichtseffekte zu. Diese Effekte kommen zustande, wenn Programme das Ergebnis und das Verhalten von Teilnehmern und Nichtteilnehmern beeinflussen. Wie in der jüngsten Arbeit von Heckman, Lochner und Taber (1998) und anderen gezeigt wurde, kann die Berücksichtigung von allgemeinen Gleichgewichtseffekten zu Schlussfolgerungen führen, die deutlich von denen abweichen, die aufgrund einer partial-analytischen Betrachtung gewonnen wurden. Die methodischen Schwierigkeiten, die mit einer solchen allgemeinen Gleichgewichtsanalyse verbunden sind, bedeuten jedoch, dass diese Analysen sowohl in der wissenschaftlichen Literatur als auch im politischen Umfeld umstritten bleiben werden. Trotz dieser Kontroverse sollten Forscher diesen Effekten jedoch ihre Aufmerksamkeit widmen, wenn auch nur indirekt, indem die Sensitivität von Kosten-Nutzen-Analysen hinsichtlich verschiedener Annahmen überprüft wird. Solche Sensitivitäts-Analysen würden eine Verbesserung gegenüber den meisten gegenwärtigen partial-analytischen Untersuchungen bedeuten, die diese Effekte einfach ignorieren.

2 Bemerkungen über die Evaluationsindustrie und Evaluationspolitik

In diesem Abschnitt werde ich die Organisationsstruktur der Evaluationsindustrie in Nordamerika beschreiben, und weise auf einige Aspekte hin, die als Modell auch für andere Länder dienen könnten.[1] Der Blick nach Nordamerika ist aus dem Grund gerechtfertigt, da die Evaluationsforschung und deren Anwendung dort am meisten fortgeschritten sind.[2] Zugleich möchte ich jedoch betonen, dass nicht alles auf dem Gebiet der Evaluation in Nordamerika perfekt ist. So hinkt die Evaluationsforschung in Soziologie und Psychologie noch Jahre (wenn nicht sogar Jahrzehnte) hinter den Wirtschaftswissenschaften hinterher, sowohl in den statistischen Methoden als auch in konzeptionellen Fragen (vgl. z.B. Rossi/ Freeman 1993). Zugleich berücksichtigen Wirtschaftswissenschaftler nicht ausreichend genug wichtige Ergebnisse der Programmimplementierung und Ergebnismessung aus anderen Wissenschaften. In diesem Abschnitt werde ich die positiven Erfahrungen hervorheben, die in Nordamerika gemacht wurden und die sich auch in anderen Ländern bewähren könnten.

2.1 Wer sind die Evaluatoren?

Im Wesentlichen gibt es vier Gruppen in Nordamerika, die Evaluationsforschung betreiben. Jede Gruppe hat ihre Vor- und Nachteile bezüglich der Kenntnis der methodischen Literatur, der Fähigkeit, Studien zeitgerecht durchzuführen und der Anreize, sich bei der Programmevaluation auch von anderen Motiven beeinflussen zu lassen außer der wissenschaftlichen Redlichkeit.

Die erste Gruppe besteht aus den Mitarbeitern staatlicher Stellen. Die meisten verfügen über wissenschaftlich ausgebildetes Personal, oftmals mit Doktorgraden in Wirtschaftswissenschaften oder verwandten Disziplinen, die „im Hause"-Evaluationen durchführen. Die Vorteile liegen auf der Hand – das Personal ist sowohl mit den Programmen, die ihre Behörden durchführen als auch mit den Daten, die dabei zu Evaluationszwecken gesammelt werden, vertraut. Ein Nachteil besteht darin, dass diese Personen zusätzlich noch mit anderen Aufgaben betraut sind, und daher oftmals nicht über den neuesten Kenntnisstand der Evaluationsmethoden verfügen. Ebenso besteht die Gefahr, dass es innerhalb der Behörde starke Anreize geben kann, positive Ergebnisse zu produzieren.

Die zweite Gruppe besteht aus professionellen Beratungsunternehmen. Die Spanne reicht dabei von Unternehmen, die sich ausschließlich mit Evaluation oder ähnlichen empirischen Studien beschäftigen, wie z.B. Mathematica oder Westat, bis hin zu großen Beratungsunternehmen, die über kleinere Abteilungen verfügen, die Evaluation betreiben, wie z.B. Goss Gilroy. Diese Unternehmen stehen unter dem Zwang, Profite zu erwirtschaften, was langen methodischen Überlegungen im Wege steht, so dass sich diese Unternehmen auf die „20 Prozent der Methoden konzentrieren, mit denen 80 Prozent der Arbeit erledigt werden kann". Diese Konzentration bedeutet in der Regel jedoch auch, dass die Arbeit termingerecht und professionell erledigt und zudem in einer Form präsentiert wird, die für den Auftraggeber verständlich und leicht umsetzbar ist. Wenn jedoch bestimmte Stellen wiederholt Geschäfte mit diesen Beratungsunternehmen machen und positive Evaluationsergebnisse eher honoriert werden als negative, können Anreizprobleme entstehen, insbesondere wenn der Anteil des Umsatzes, den diese Beratungsunternehmen mit diesen Stellen erwirtschaften, hoch ist.

Gemeinnützige Beratungsunternehmen, wie „Manpower Demonstration Research Corporation" (MDRC) in den USA und die „Social Research and Demonstration Corporation", ihr kanadischer Ableger, ähneln im Allgemeinen kommerziellen Beratungsunternehmen, mit einigen Unterschieden in der Schwerpunktsetzung. Gemeinnützige Unternehmen streben oft eine führende Rolle in der Entwicklung neuer Politikideen an und suchen dann Quellen, die bereit sind, die Durchführung dieser Pilotprojekte und entsprechender Evaluationen zu finanzieren. Gemeinnützige Unternehmen, zumindest diejenigen am oberen Ende, wie z.B. MDRC, haben i.d.R. einen akademischen Hintergrund. Da die Finanzierung dieser Unternehmen zu einem Großteil durch Stiftungen erfolgt, wird auch das oben erwähnte Anreizproblem geringer sein. Die niedrigere Bezahlung speziell bei kleineren Beratungsunternehmen kann jedoch bewirken, dass die Positionen vorwiegend von Personen besetzt werden, die eher an einem bestimmten politischen Ergebnis interessiert sind als an der Beantwortung der Frage, ob ein spezielles Programm tatsächlich wirkt.

Die letzte Gruppe, die Evaluationsstudien durchführt, besteht aus akademischen Wissenschaftlern, vorwiegend Wissenschaftler aus den Bereichen Wirtschaftswissenschaften, Soziologie und Psychologie. Verglichen mit anderen Gruppen sind Akademiker näher an dem neuesten methodischen Stand. Zugleich sind sie berüchtigt dafür, Aufträge nicht rechtzeitig zu erledigen, und die Ergebnisse in einer Art zu präsentieren, die zwar wissenschaftlichen Standards gerecht wird, jedoch zu technisch für die Auftraggeber ist. Akademiker haben zugleich das geringste Anreizproblem, da die meisten nur einen kleinen Teil ihres Einkommens mit der Durchführung von Evaluationsstudien bestreiten.

[1] Ich verwende den Begriff "Nordamerika" in dem Sinn, in dem er oft in Kanada gebraucht wird, wo darunter Kanada und die Vereinigten Staaten fallen, nicht jedoch Mexiko.

[2] Vgl. Riddell (1991) für eine frühere, ausführliche Einschätzung der nordamerikanischen Evaluationsindustrie, die sowohl auf organisatorische als auch auf ökonometrische Gesichtspunkte eingeht.

In großen Evaluationsstudien arbeiten diese unterschiedlichen Gruppen zusammen. In der „U.S. National JTPA Study" z.B. führten ein kommerzielles Unternehmen, Abt Associates, zusammen mit zwei gemeinnützigen Unternehmen, MDRC und dem „National Opinion Research Center" (NORC), und akademischen Top-Forschern gemeinsam die Evaluation durch. Gängige Praxis in Kanada ist, dass Unternehmen, seien es kommerzielle oder gemeinnützige, Evaluationen im Auftrag von staatlichen Behörden durchführen und dann selbst mit akademischen Forschern zusammenarbeiten, die sie beim Untersuchungsdesign und der Interpretation der Resultate beraten. Mein Eindruck ist, dass diese Kooperationen oftmals die Stärken der verschiedenen Gruppen kombinieren und deren Schwächen teilweise ausgleichen können und somit in der Praxis vorangetrieben werden sollten.

2.2 Wer zahlt für Evaluationen?

In Nordamerika, bestehen nicht nur verschiedene Gruppen, die Evaluationen durchführen, sondern auch verschiedene Gruppen, die diese finanzieren. Die größte Gruppe stellen dabei natürlich staatliche Behörden dar, die bestimmte Programme durchführen. Die föderale Struktur der USA und Kanada bewirkt dabei, dass das gleiche Programm sowohl auf föderaler, bundesstaatlicher als auch auf Bezirksebene evaluiert werden kann und manchmal auch wird. Dies trifft insbesondere bei Arbeitsmarktprogrammen zu, da diese in beiden Ländern auf föderaler Ebene, von den einzelnen Bundesstaaten und auch von Behörden auf Bezirksebene durchgeführt werden.

Die zweite Gruppe, die Evaluationen finanziert, besteht aus staatlichen Forschungseinrichtungen, die unabhängig von den Behörden operieren, die die Programme durchführen. Als Beispiel kann die „U.S. National Science Foundation" (NSF) genannt werden sowie der „Social Science and Humanities Research Council" in Kanada. Diese finanzieren zwar vorwiegend methodische Forschung, bei Evaluationsstudien kann diese jedoch unmittelbare und wichtige praktische Konsequenzen nach sich ziehen. Die methodischen Forschungen von James Heckman und anderen Wissenschaftlern, die Daten aus dem „U.S. National Job Training Partnership Act" (JTPA) nutzten, wurden zum Teil von der NSF finanziert. Zusätzlich zu ihrem methodischen Aspekt, konnten diese Arbeiten die Sichtweise politischer Entscheidungsträger hinsichtlich der Ergebnisse des JTPA-Experiments beeinflussen, indem sie auf die Sensitivität der Schätzergebnisse des Experiments (Heckman/ Smith 2000) und die wichtige Rolle von Substitutionseffekten innerhalb der Kontrollgruppe hin zu anderen Trainingsprogrammen (Heckman/ Hohmann/ Smith/ Khoo 2000) hingewiesen haben. In Abschnitt 3 werde ich die Gelegenheit haben, auf diese Arbeit näher einzugehen.

Die letzte Finanzierungsquelle sind private Stiftungen. Einige von ihnen sind an der Implementierung von speziellen Programmen interessiert – z.B. Erziehungsprogramme für Kinder aus Familien mit geringem Einkommen. Diese Stiftungen finanzieren Evaluationsstudien, die darauf abzielen, einen Datensatz zu erstellen, der dann die empirische Rechtfertigung liefern soll, diese Programme auszuweiten.

Die Existenz von verschiedenen Finanzierungsquellen, unter ihnen solche, die nicht an der Durchführung von Programmen der aktiven Arbeitsmarktpolitik beteiligt sind, ist mitentscheidend für das Aufkommen von divergierenden Sichtweisen hinsichtlich der Effizienz von Programmen. Die Existenz von Finanzierungsquellen, die kein direktes Interesse an dem Überleben bestimmter Programmen haben, bedeutet zudem, dass Forscher nicht fürchten müssen, aufgrund negativer Ergebnisse von zukünftigen Evaluationsstudien ausgeschlossen zu werden.

2.3. Wer evaluiert die Evaluatoren?

Ein wichtiger Aspekt bei Evaluationen ist die Qualitätskontrolle. Wer oder was kann sicherstellen, dass Evaluatoren die richtigen Fragen stellen und die besten Methoden verwenden, um sie zu beantworten? Diese Qualitätskontrolle hat zwei Aspekte. Der eine Aspekt betrifft die Frage, wie jemand, der eine Evaluationsstudie finanziert, wissen kann, ob er eine gute oder eine schlechte Evaluation erhalten hat? Der zweite Aspekt betrifft das Anreizproblem, auf das bereits hingedeutet wurde. Unterliegen Institutionen oder Personen dem Anreiz, die empirischen Ergebnisse zu Gunsten des Programms auszulegen? Dies sind wichtige Fragen, die jedoch zu wenig Interesse auf Seiten der Forschung erfahren haben. Ich werde hier lediglich einige Beobachtungen aus der Praxis vorstellen.

Die erste Qualitätskontrolle besteht natürlich aus den Personen bei den staatlichen Behörden, die die Evaluationsstudie in Auftrag gegeben haben. Nach meiner Erfahrung kann qualifiziertes Personal, das sowohl in der Lage ist, bei der Gestaltung des Evaluationsprogramms mitzuhelfen als auch die Evaluation, die es von ihren Auftragnehmern erhält, zu beurteilen, die Qualität und den Wert der Evaluationsstudien wesentlich steigern. Meiner Meinung nach wäre es eine schlechte Idee, die meisten Evaluationsstudien innerhalb der Behörden selbst durchzuführen. Es ist jedoch wichtig, dass in diesen Behörden Personen mit den technischen und organisatorischen Fähigkeiten vorhanden sind, die notwendig sind, um die Evaluationsarbeit der Auftragnehmer beeinflussen zu können.

Die akademische Welt spielt ebenfalls eine wichtige Rolle in diesem Prozess der Qualitätskontrolle. Zunächst können Wissenschaftler beauftragt werden, Studien, die von kommerziellen oder gemeinnützigen Unternehmen im Auftrag von staatlichen Behörden durchgeführt wurden, zu begutachten. Zweitens haben die meisten Personen, die in Behörden oder Unternehmen Evaluationsstudien durchführen, einen akademischen Hintergrund. Die Werte, die sie dort vermittelt bekommen – idealerweise eine starke Hingabe zur empirischen Wahrheit – wirken als ein interner Qualitätskontrollmechanismus, der sie bei ihrer Arbeit leiten wird.

Drittens: In Nordamerika gibt es eine starke Fluktuation zwischen akademischen, staatlichen und unternehmerischen Stellen. Dies umfasst z.B. auch höchste Positionen beim U.S. Department of Labor, die in jüngster Zeit unter anderem von Lawrence Katz aus Harvard, Alan Krueger aus Princeton und Harry Holzer aus Michigan State besetzt wurden. Auch auf niedrigerer Ebene gibt es Ökonomen, die oft Positionen in beiden Bereichen besetzen, z.B. solche, die eine Stelle bei der Rand Corporation und der volkswirtschaftlichen Abteilung der „University of California, Los Angeles" (UCLA) innehaben. Personen aus staatlichen Behörden oder Unternehmen, die darauf hoffen, wieder in die akademische Welt zurückzukehren, oder bereits einen Teil ihres Lebens dort verbracht haben, haben einen starken Antrieb, die akademische Qualität ihrer Arbeit hochzuhalten. Zugleich trägt die starke Fluktuation an Individuen zwischen diesen beiden Bereichen dazu bei, dass methodische Neuentwicklungen schneller in die Praxis Einzug halten. Das nordamerikanische Experiment verdeutlicht, wie wichtig es sein kann, einen Austausch von Forschern zwischen der akademischen, staatlichen und unter-

nehmerischen Welt zuzulassen, und insbesondere führenden Wissenschaftlern (zeitweise) Rollen im Forschungsmanagement und der Politik zuzuweisen.

Da die meisten Evaluationsforscher bei staatlichen oder privaten Stellen daran interessiert sind, den Kontakt zu der akademischen Welt aufrechtzuerhalten oder evtl. dorthin zurückzukehren, wirken ihre Bemühungen, die Ergebnisse in wissenschaftlichen Zeitschriften zu publizieren, als eine zusätzliche Qualitätskontrolle. Ökonomen bei größeren Beratungsunternehmen publizieren zusätzlich zu den Gutachten, die sie für ihre Auftraggeber anfertigen, routinemäßig die Ergebnisse ihrer Evaluationsforschung in wissenschaftlichen Zeitschriften. Dieser Wunsch, ein Teil der akademischen Welt zu bleiben, und Anerkennung für die eigene Arbeit durch Publikationen zu erhalten, dient als Gegengewicht zu den Anreizen, dem Auftraggeber wunschgemäße Ergebnisse zu liefern. Staatliche Behörden können diese Anreize durch Publikation der eigenen Ergebnisse noch verbessern, indem sie in den Evaluationsverträgen Mittel vorsehen, die es den Forschern ermöglichen, ihre Ergebnisse in wissenschaftlichen Zeitschriften zu veröffentlichen.

2.4 Die Bedeutung von Daten

Ein aktuelles Thema in der wissenschaftlichen Literatur ist die Bedeutung guter Daten (vgl. z.B. Heckman/ Ichimura/ Smith/ Todd 1998). Vieles deutet darauf hin, dass relativ viel Zeit darauf verwendet wurde, sich über die Wahl geeigneter Schätzer Gedanken zu machen und zu wenig Zeit, um über die Qualität der Daten nachzudenken, auf die die Schätzer dann angewendet werden. Dabei sind die besten ökonometrischen Methoden nicht in der Lage, schlechte Daten auszugleichen.

Es gibt eine Reihe von Dingen, die staatliche Behörden tun können, um die Qualität und Menge der Daten, die für die wissenschaftliche Evaluationsforschung geeignet wären, zu verbessern. Zunächst einmal sollten Daten aus Evaluationen der gesamten Forschungsgemeinschaft zum Zwecke der Überprüfung der Ergebnisse und für Analysen, die auf andere Fragen abzielen, zugänglich gemacht werden. Jeder Evaluationsvertrag sollte von daher finanzielle Mittel für die Anfertigung eines öffentlich nutzbaren Datensatz enthalten. Datensätze von vergangenen Evaluationen in den USA, die der Forschungsgemeinschaft zur Verfügung gestellt wurden, wie z.B. die „National Supported Work Demonstration" (NSW) und die zahlreichen „MDRC Work-Welfare"- Experimente, bildeten die Basis für wichtige und wesentliche methodologische Neuentwicklungen. Natürlich müssen geeignete Maßnahmen bezüglich der Zustimmung der Befragten und der Wahrung der Vertraulichkeit getroffen werden.

Zweitens können staatliche Behörden Daten, die als Teil von bestimmten Programmen gesammelt werden, zur wissenschaftlichen Forschung weitergeben, nachdem wiederum angemessene Sicherheitsmaßnahmen getroffen wurden. Um den größtmöglichen Nutzen aus diesen Daten zu ziehen, sollte die amtliche Datenerhebung dabei im Hinblick auf Evaluationsfragen ausgestaltet werden. Das könnte z.B. bedeuten, zusätzliche Variablen zu sammeln (z.B. fehlen in vielen amtlichen Daten Angaben zur Bildung) oder die Datenerhebungsprozedur so zu gestalten, dass die Zahl der fehlenden oder ungültigen Werte bei Schlüsselvariablen möglichst gering gehalten wird. In Nordamerika wurden administrative Daten ohne zusätzliche Erhebungen zu Evaluationszwecken genutzt, darüber hinaus wurden diese Daten mit anderen Datensätzen zusammengeführt, um auf diese Weise möglichst einfach und kostengünstig auch Längsschnittsangaben über Einkommen und Erwerbstätigkeit zu erhalten. In der Tat basieren die einzigen validen Schätzungen über die langfristigen Effekte (vier oder mehr Jahre nach der Maßnahme) von Beschäftigungs- und Trainingsmaßnahmen auf solchen zusammengefügten administrativen Daten – vgl. Couch (1992) für das „NSW Demonstration" und U.S. General Accounting Office (1996) für das JTPA Experiment.

Neben administrativen Daten und Umfragen, die speziell für eine Evaluation gesammelt wurden, haben allgemeine Paneldatensätze eine wichtige Rolle in der Evaluationsforschung in den USA gespielt. Der Prototyp der Paneldatensätze, an die ich in diesem Zusammenhang denke, sind die „U.S. National Longitudinal Surveys". Z.B. dient der „National Longitudinal Survey of Youth", der auch Informationen über die Kinder der Befragten sammelt, als Datengrundlage für die wohl beste Analyse, die zu dem „U.S. Head Start Programm" für benachteiligte Kinder im Vorschulalter existiert (vgl. Currie/ Thomas 1995). Berücksichtigt man die Anzahl an Untersuchungen, die mit diesen Datensätzen arbeiten, kann man zu dem Eindruck gelangen, dass Paneldatensätzen diesen Typs wohl die besten Datensätze darstellen, die zur Zeit in den Sozialwissenschaften verfügbar sind.

3 Bemerkungen zu jüngsten Entwicklungen bei den Evaluationsmethoden

In diesem Abschnitt gehe ich auf einige wichtige neuere Themen zu den Evaluationsmethoden ein. Ich werde dabei die wichtigsten Probleme vorstellen und verweise auf die neueste Literatur, in der der interessierte Leser mehr Details erfahren kann.

3.1 Heterogenität

Ein Großteil des konzeptionellen Fortschritts in der Evaluationsliteratur resultiert aus einem sorgfältigen und formalen Nachdenken über Modelle, in denen die Wirkung von Programmen zwischen den Personen differiert. Die Berücksichtigung von heterogenen Wirkungen im Rahmen von Evaluationen, macht deutlich, dass es nicht einen einzigen interessierenden Parameter gibt, sondern mehrere. Es wird ferner deutlich, dass konsistente Schätzer für einen Parameter nicht auch konsistente Schätzer für andere Parameter sein müssen.

Um dies deutlicher zu erkennen, verwenden wir folgende einfache Notation. Y_1 sei das Ergebnis, dass eine Person erzielt, nachdem er oder sie an einem Programm teilgenommen hat. Mit diesem Ergebnis kann das Einkommen, die Beschäftigung, die Gesundheit oder eine andere Größe gemeint sein, die mit dem Programm beeinflusst werden soll. Mit Y_0 soll das gleiche Ergebnis während der gleichen Zeit für die Nichtteilnahme der Person bezeichnet sein. Natürlich kann eine Person entweder an dem Programm teilnehmen oder nicht, so dass nur eines dieser zwei potenziellen Ergebnisse für jede Person beobachtet wird. Nichtsdestotrotz macht es konzeptionell Sinn, für jede Person diese beiden Ergebnisse zu unterscheiden und die Differenz zwischen ihnen als die Wirkung des Programms auf diese Person anzusehen. Anders formuliert, die Wirkung eines Programms auf eine daran teilnehmende Person besteht aus der Differenz dieser beiden Ergebnisse. Formal kann die Wirkung auf eine Person *i* folgendermaßen definiert werden:

$$\Delta_i = Y_{1i} - Y_{0i},$$

wobei Δ_i die Wirkung auf die Person *i* bezeichnet.

Wir können nun zwischen verschiedenen Möglichkeiten unterscheiden, wie die Wirkung eines Programms zwischen verschiedenen Personen variiert. Im einfachsten Fall gibt es keine Variation und das Programm hat für jeden Teilnehmer den gleichen Effekt. In der obigen Notation würde für alle Personen i gelten: $\Delta_i = \Delta$. Obwohl diese „gemeinsame Effekt"-Annahme höchst unwahrscheinlich ist, mag sie in einigen Zusammenhängen als gute Approximation dienen (und in anderen nur schlecht passen). Diese Annahme war lange Zeit prägend für die ökonometrische und angewandte Literatur über Programmevaluation.

Unter weniger restriktiven Annahmen variieren die Maßnahmeneffekte über die Personen, jedoch haben vor der Teilnahme weder der potenzielle Teilnehmer noch die für das Programm verantwortlichen Personen Informationen über diese personenspezifische Wirkungskomponente. Programme haben somit verschiedene Effekte bei verschiedenen Personen, niemand kann jedoch vorhersagen, wer mehr oder weniger von der Teilnahme profitiert, so dass diese Variation in den Wirkungen keinen Einfluss auf die Partizipationsentscheidung hat. Unter diesen Annahmen hat die Variation in den Maßnahmenwirkung auch nur geringe politische Implikationen.

Unter allgemeineren Annahmen variieren die Maßnahmeneffekte zwischen den Personen, wobei zusätzlich noch entweder die Teilnehmer oder die Verantwortlichen des Programms oder aber beide a-priori Informationen über den individuellen Wert dieser Maßnahme haben. Unter diesen Annahmen beeinflusst die personenspezifische Wirkungskomponente die Teilnahme an dem Programm. Daraus resultieren auch wichtige politische Implikationen, da Änderungen der politischen Rahmenbedingungen, die unterschiedliche Personengruppen in das Programm ein- oder ausschließen, einen unterschiedlichen durchschnittlichen Effekt haben werden.

Um zu sehen, warum die Unterschiede in den Maßnahmenwirkungen relevante Politikimplikationen haben können, wollen wir im Folgenden drei Parameter betrachten, die für politische Entscheidungsträger von Interesse sein könnten. Diese Parameter sollen sich dabei auf ein freiwilliges Programm beziehen, an dem nur ein Teil und nicht die ganze interessierende Grundgesamtheit teilnimmt, z.B. ein freiwilliges Job-Training Programm für Personen, die Sozialhilfe beziehen. Ein Parameter von Interesse ist der Effekt, den das Programm auf die aktuellen Teilnehmer hat. In der Literatur wird dieser Parameter „Treatment on the Treated" (TT) genannt oder in unserem Beispiel der Effekt des „Trainings auf die Trainierten". Zusammen mit Informationen über die Kosten des Programms, und unter Vernachlässigung anderer allgemeiner Gleichgewichtseffekte außer den Steuereffekten, gibt dieser Parameter einen Hinweis auf die Frage, ob das Programm beendet werden sollte oder nicht. In einer strikten Kosten-Nutzen-Betrachtung sollte ein Programm, für das die „Treatment on the Treated"-Wirkung unterhalb der Kosten des Programms liegt (inklusive der Nettowohlfahrtsverluste, die durch die Steuerfinanzierung des Programms entstehen) beendet werden.

Oft ist jedoch die Frage, ob ein Programm beendet werden soll, nicht die einzige und nicht einmal die vorrangige Frage, die beantwortet werden soll. Das Ziel könnte z.B. auch darin bestehen, eine Politikmaßnahme zu evaluieren, die auf eine 10prozentige Reduktion der an einer Maßnahme teilnehmenden Personen abzielt. Diese Reduktion könnte dabei entweder durch eine Selbstbeteiligung an den Kosten des Trainingsmaterials oder durch eine Rationierung der verfügbaren Plätze erreicht werden. Der Parameter, der in diesem Fall von Interesse ist, ist nicht die Wirkung auf all diejenigen Personen, die derzeit an dem Programm teilnehmen, sondern vielmehr auf diejenigen, die von diesem Programm ausgeschlossen werden.

Bei heterogenen Maßnahmeneffekten könnte es durchaus sein, dass der durchschnittliche Maßnahmeneffekt für diese Gruppe nicht über die eingesparten Kosten hinausgeht, während die 90 Prozent der übrigen Teilnehmer einen über die Kosten hinausgehenden Nutzen von diesem Programm haben. Wenn diejenigen, die am meisten von dem Programm profitieren, auch diejenigen sind, die am bereitwilligsten daran teilnehmen (und so am ehesten bereit sind, sich an den Kosten des Trainingsmaterials zu beteiligen oder sich um Plätze zu bemühen), dann ist dieses Ergebnis in der Tat zu erwarten. Selbst ein einfaches ökonomisches Modell der Programmteilnahme deutet darauf hin, dass wenn die potenziellen Teilnehmer eine bestimmte Vorstellung von ihrem personenspezifischen Nutzen aus dem Programm haben, diejenigen mit dem größten Nutzen c.p. auch am ehesten daran teilnehmen werden.

Dieser marginale Wirkungsparameter ist ein Beispiel für das, was Imbens und Angrist (1994) als „Local Average Treatment Effect" oder LATE bezeichnet haben. Er ist ein marginaler Teilnahmeparameter, ausgelöst durch eine bestimmte Politikmaßnahme, in diesem Fall z.B. die Teilnahmereduktion in Form einer Einführung einer Selbstbeteiligung am Unterrichtsmaterial. LATE misst den durchschnittlichen Einfluss des Programms auf diejenigen Personen, deren Teilnahmestatus sich aufgrund der durchgeführten Politikmaßnahme ändert.

Anstatt ein Programm zu beenden oder zu kürzen, könnte auch eine Politikmaßnahme in Betracht kommen, die auf eine Ausweitung des Programms auf alle geeigneten Personen abzielt. In unserem Beispiel würde das heißen, dass alle Sozialhilfeempfänger verpflichtet werden, an dem Job Training Programm teilzunehmen. Die zu beantwortende Frage lautet nun, ob ein solches verbindliches Programm einer Kosten-Nutzen-Analyse standhalten würde. Der Wirkungsparameter, der dabei von Interesse ist, wird in der Literatur als „Average Treatment Effect" (ATE) bezeichnet. Dieser Parameter gibt den durchschnittlichen Maßnahmeneffekt auf alle geeigneten Personen wieder, anstatt nur auf diejenigen, die freiwillig teilnehmen. Wenn wir wiederum an unser einfaches Modell der Programmteilnahme denken, in dem diejenigen mit den größten erwarteten Nutzen auch tatsächlich teilnehmen, können wir erwarten, dass der ATE kleiner sein wird als „Treatment on the Treated"-Effekt.

In einer Welt homogener Maßnahmenwirkungen werden natürlich alle drei Wirkungsparameter – TT, LATE und ATE – gleich sein. Diese Homogenität ist eine Erklärung für die Attraktivität der Annahme identischer Effekte, wie unrealistisch sie auch sein mag. In der Realität werden diese Maßnahmen jedoch unterschiedlich sein. Wenn zudem die Teilnehmer oder die für die Programme Verantwortlichen Informationen über diese unterschiedlichen Wirkungen haben, können sich diese Unterschiede auch in konkreten Politikempfehlungen niederschlagen.

Heterogene Programmeffekte haben außerdem Auswirkungen auf einige weitverbreitete nichtexperimentelle Evaluationsstrategien, wie z.B. die Instrumentalvariablen-Methode. Heckman, LaLonde und Smith (1999) und Heckman (1997) diskutieren diese Aspekte ausführlicher. Schließlich können

außer TT, LATE und ATE eine Reihe anderer interessierender Parameter definiert werden, wie z.B. die Varianz der Maßnahmenwirkung zwischen den Teilnehmern. Heckman, Smith und Clements (1997) diskutieren die Probleme bei der Schätzung solcher Parameter.

3.2 Soziale Experimente

Soziale Experimente haben sich zur ersten Wahl bei der Evaluation sozialer Programme in den USA entwickelt. Großangelegte Evaluationen, wie z.B. die „National JTPA Study" in den USA (vgl. Bloom und andere 1997) und die „Self-Sufficiency"-Projekte in Kanada haben einen Wandel in der Sichtweise und vor allen Dingen im Handeln der Politiker bewirkt. Mit einigen Ausnahmen, wie z.B. dem „RESTART"-Experiment in Großbritannien (vgl. z.B. White/ Lakey 1992 und Dolton/ O´Neill 1996), einigen sozialen Experimenten zur Evaluation von Trainingsprogrammen in Norwegen (vgl. Torp et al. 1993), und einem kleinen Experiment in Schweden, das in Björklund und Regnér (1996) beschrieben wird, haben diese Methoden erst seit kurzem Einzug in die meisten europäischen Ländern gehalten. In diesem Abschnitt untersuche ich die Kosten und Nutzen sozialer Experimente, und komme zu dem Schluss, dass sie ein wichtiges Werkzeug zur Evaluation darstellen, aber eines, das einer sorgfältigen Implementierung und Interpretation bedarf. Für eine zusätzliche (mehr technische) Diskussion sozialer Experimente vgl. Björklund und Regnér (1996), Burtless und Orr (1986), Burtless (1995), Heckman und Smith (1993, 1995, 1996a,b) und Heckman, LaLonde und Smith (1999).

Idealerweise gehen soziale Experimente von Personen aus, die unter normalen Umständen an einem Programm teilnehmen würden und weisen sie einer der beiden folgenden Gruppen zu. Die erste Gruppe, Teilnehmergruppe genannt, nimmt an dem Programm wie gewöhnlich teil, während die zweite Gruppe, die sog. Kontrollgruppe, davon ausgeschlossen wird. Experimentelle Kontrollgruppen unterscheiden sich von traditionellen nichtexperimentellen Kontrollgruppen, die sich üblicherweise aus Nichtteilnehmern zusammensetzen, weil sie, abgesehen von gewöhnlichen Stichprobenvariationen, die gleiche Verteilung von beobachtbaren und unbeobachtbaren Charakteristika aufweisen wie die Teilnehmergruppe. Auch die Erzeugung einer solchen experimentellen Kontrollgruppe unterscheidet sich von der Erzeugung einer typischen nichtexperimentellen Kontrollgruppe. In nichtexperimentellen Evaluationen werden statistische Techniken angewandt, um die Ergebnisse von Personen, die nicht an dem Programm teilnehmen, an die hypothetischen Ergebnisse anzunähern, die man beobachten würde, wenn die Teilnehmer nicht an dem Programm teilgenommen hätten. Im Gegensatz dazu produzieren Experimente direkt dieses hypothetische Gegenereignis, indem einige potenzielle Teilnehmer an der Teilnahme gehindert werden.

Ein Ergebnis dieser zufälligen Zuordnung besteht darin, dass unter bestimmten Bedingungen ein einfacher Vergleich des durchschnittlichen Ergebnisses der Teilnehmer- und Kontrollgruppe eine konsistente Schätzung des Maßnahmeneffektes des Programms auf die Teilnehmer wiedergibt. Unter Verwendung der vorhergehenden Terminologie bewirkt ein soziales Experiment eine konsistente Schätzung des „Treatment on the Treated"-Parameters. Bei geschicktem Design kann das soziale Experiment auch zur Schätzung des „Average Treatment"-Effektes genutzt werden, wie z.B. in dem britischen „RESTART"-Experiment, wo Personen zufällig die Teilnahme an einem ansonsten obligatorischen Experiment verwehrt wurde. Ähnlich kann eine zufällige Zuordnung innerhalb einer Teilgruppe, wie in der Evaluation von Arbeitslosigkeitsversicherungsansprüchen (Teilnahmezuordnung basierend auf der vorhergesagten Dauer der Arbeitslosigkeit) bei Black, Smith, Berger und Noel (2000) zur Schätzung des LATE-Parameters genutzt werden.

Abgesehen von der Tatsache, dass soziale Experimente unter bestimmten später zu diskutierenden Bedingungen konsistente Schätzer der TT-Wirkung produzieren, weisen sie weitere Vorteile gegenüber normalen nichtexperimentellen Methoden auf. Als Erstes sind soziale Experimente politischen Entscheidungsträgern einfach zu erklären. Die meisten gebildeten Personen verstehen die Ideen, die hinter einer zufälligen Zuordnung stecken.[3]

Zum Zweiten sind soziale Experimente weniger kontrovers als nichtexperimentelle Methoden. In Nordamerika haben die weit streuenden Schätzungen der Wirkung des „Comprehensive Employment and Training Act" (vgl. dazu Barnow 1987) zu einem ernsthaften Zweifel an nichtexperimentellen Methoden geführt. In diesen Evaluationsstudien kamen verschiedene Forscher, die den gleichen Datensatz genutzt haben, zu deutlich verschiedenen Schlussfolgerungen über die Effektivität des Programms.[4] Im Gegensatz dazu sind Experimente in der Lage, „eine Zahl" zu liefern anstatt einer ganzen Palette von verschiedenen Schätzern, die man oftmals bei nichtexperimentellen Schätzern erhält. Berücksichtigt man jedoch, dass auch experimentelle Schätzergebnisse sensitiv sein können und auf einer Reihe von Annahmen basieren, wird diese Eindeutigkeit fragwürdig und sollte nicht überbetont werden (Heckman/ Smith 2000). Abgesehen von dieser Sensitivität bleiben experimentelle Schätzer verglichen mit nichtexperimentellen, aufgrund ihrer einfachen und nachvollziehbaren Methode, den nichtexperimentellen jedoch überlegen.

Drittens kann man bei Experimenten nicht betrügen. Wenn also die Person, Firma oder Organisation, die eine Evaluationsstudie durchführt, eine Präferenz für ein Ergebnis hat, das dem Programm eine hohe oder geringe Wirksamkeit attestiert, so bieten Experimente weniger Möglichkeiten der Manipulation. Im Gegensatz dazu könnte ein nichtexperimenteller Forscher, der die Literatur und die dort beschriebenen Verzerrungen der verschiedenen nichtexperimentellen Schätzer kennt, diese Informationen nutzen, um eine Schätzstrategie zu wählen, die zu dem gewünschten Ergebnis führt. Die Durchführung eines Experiments zur Evaluation eines Programms erschwert solche Manipulationen, da nunmehr keine Möglichkeit besteht, zwischen verschiedenen Schätzern zu wählen.

Zum Vierten bieten Experimente die Möglichkeit, nichtexperimentelle Schätzer zu kalibrieren und die Effizienz verschiedener nichtexperimenteller Schätzstrategien zu untersuchen. LaLonde (1986) nutzt in seiner Arbeit Daten des „U.S. National Supported Work Demonstration"-Experimentes (NSW), um die Verzerrungen einer gängigen Schätzstrategie zu untersuchen, bei der eine Kontrollgruppe aus einem exis-

[3] Natürlich bleiben alle anderen komplexen Probleme im Zusammenhang mit Wirkungsschätzern, seien es nun experimentelle oder nichtexperimentelle, bestehen. Dazu gehören Fragen, in welchem Ausmaß geschätzte Parameter eines Programms und einer Grundgesamtheit auf ein anderes Programm oder eine andere Grundgesamtheit übertragen werden können.

[4] Es bleibt anzumerken, dass ein Teil dieser Unterschiede auf Wahlmöglichkeiten hinsichtlich der Handhabe der Daten zurückzuführen ist und nicht auf verschiedene nichtexperimentelle Schätzer. Vgl. Dickinson, Johnson und West (1987).

tierenden nationalen Datensatz ausgewählt wird, um dann nichtexperimentelle Techniken anzuwenden.[5] Sein Ergebnis, dass nichtexperimentelle Schätzer oft von experimentellen Schätzern abweichen, hat wesentlich zu der Hinwendung zu Experimenten in Nordamerika beigetragen.

In einer neueren Arbeit benutzen Dehejia und Wahba (1999a,b) und Smith und Todd (2000) den selben NSW Datensatz, um die Leistungsfähigkeit des Propensity Score Matchings zu untersuchen, das ich im Abschnitt 3.3 näher vorstellen werde. Heckman, Ichimura, Smith und Todd (1996, 1998) und Heckman, Ichimura und Todd (1997) nutzen Daten der National JTPA Studie, um Matching-Methoden und Probleme der Selektivitätsverzerrungen zu untersuchen. Heckman und Hotz (1989) schließlich finden, dass eine Auswahl unter verschiedenen nichtexperimentellen Schätzern, die sich auf Spezifikationstests stützt, die mit diesen Methoden verbundenen Verzerrungen verringern kann.[6]

Obwohl soziale Experimente eine Reihe von Vorteilen gegenüber gewöhnlichen nichtexperimentellen Methoden aufweisen, stellen sie keineswegs eine einfache Lösung für alle Evaluationsprobleme dar. Die folgenden Ausführungen dieses Abschnitts befassen sich mit Einschränkungen und möglichen Problemen sozialer Experimente. Diese Probleme wurden in dem letzten Jahrzehnt ausgiebig in der nordamerikanische Literatur behandelt, haben jedoch auch einen Einfluss auf die Ausgestaltung von Experimenten in Europa ausgeübt.

Wir beginnen mit der Anmerkung, dass mit sozialen Experimenten nicht alle interessierenden Parameter geschätzt werden können. Zunächst gibt es persönliche Charakteristika, wie z.B. Geschlecht oder Familieneinkommen, die einer zufälligen Zuordnung entgegenwirken. Zum anderen sind soziale Experimente zwar i.d.R. sehr gut geeignet, die Wirkung der Maßnahme auf die Teilnehmer zu schätzen (TT-Effekt), jedoch weniger geeignet, allgemeine Gleichgewichtseffekte zu berücksichtigen. Ich werde diese allgemeinen Gleichgewichtseffekte genauer in Abschnitt 3.4 diskutieren. Schließlich erfordert z.B. die Schätzung der Variation in den Maßnahmenwirkungen zwischen verschiedenen Individuen auch in einem partiellen Gleichgewichtskontext zusätzliche nichtexperimentelle Annahmen. Heckman, Smith und Clements (1997) diskutieren diesen Aspekt detaillierter.

Zum Zweiten kann eine zufällige Zuordnung die Wirkung eines Programms unterbrechen, so dass die geschätzte Maßnahmenwirkung nicht die Wirkung des Programms unter normalen Umständen wiedergibt. Wir betrachten in diesem Zusammenhang drei Beispiele. Als erstes: Wenn die Anzahl an Personen, die an einem Programm teilnehmen, während eines Experimentes die gleiche bleibt wie zu anderen Zeiten auch, werden die für das Programm Verantwortlichen zusätzliche potenzielle Teilnehmer anwerben müssen. Auf diese zusätzlichen Teilnehmer, die zufällig in die Gruppe der Teilnehmer und die Kontrollgruppe aufgeteilt werden, kann das Programm jedoch einen unterschiedlichen Einfluss haben als auf die normalerweise Teilnehmenden.

Ferner kann die zufällige Zuordnung einen Einfluss auf das Antwortverhalten innerhalb der Teilnehmer- und Kontrollgruppe haben, das sich zudem noch von dem Verhalten in nichtexperimentellen Evaluationen unterscheiden kann. Potenzielle Teilnehmer, denen die Teilnahme an dem Programm verweigert wurde, können sich auch weigern, an der Befragung teilzunehmen. Schließlich kann sich die Drohung, aufgrund der zufälligen Zuordnung nicht an dem Programm teilnehmen zu können, auch negativ auf Aktivitäten auswirken, die Teilnehmer im Vorfeld des Programms unternehmen und die einen Einfluss auf den Maßnahmenerfolg haben können.

Drittens können Experimente manchmal teurer sein als nichtexperimentelle Methoden. Eine zufällige Zuordnung verursacht zusätzliche Kosten für Training und Kontrolle des Personals, das diese zufällige Zuordnung vornehmen muss, und des Weiteren Kosten, die durch die Unterrichtung und Überzeugung der potenziellen Teilnehmer entstehen, die ihrer Teilnahme an einem sozialen Experiment zustimmen müssen. Dieser Punkt darf jedoch nicht überbetont werden, wie von Heckman, LaLonde und Smith (1999, vgl. Abschnitt 8.1) hervorgehoben wird. Nichtexperimentelle Evaluationen sind zwar kostengünstig, da sie auf existierende nationale Datensätze zur Konstruktion einer Kontrollgruppe zurückgreifen. Die Nutzung solcher Datensätze bedeutet jedoch fast immer, dass die Kontrollgruppe nicht aus dem gleichen lokalen Arbeitsmarkt gezogen wurde wie die Teilnehmergruppe, und dass evtl. wichtige Variablen unterschiedlich gemessen wurden. Wenn diese Faktoren wichtig sind im Hinblick auf die Vermeidung von Verzerrungen, gehen die eingesparten Kosten mit verzerrten Schätzern einher. Heckman, Ichimura, Smith und Todd (1998) präsentieren empirische Evidenz, dass diese Verzerrungen beträchtlich sein können.

Zum Vierten kann eine zufällige Zuordnung politische Kontroversen und eine negative öffentliche Meinung auslösen. In der „U.S. National JTPA Study" z.B. mussten die Evaluatoren ca. 200 der ungefähr 600 existierenden JTPA Trainingszentren in den USA kontaktieren und 1 Mio. US$ an Budgetmitteln aufwenden, um schließlich 16 Zentren zu überzeugen, freiwillig an dem Experiment teilzunehmen. Doolittle und Traeger (1990) weisen darauf hin, dass der Haupthinderungsgrund für die Teilnahme eine mögliche negative öffentliche Meinung war.

Schließlich wird die Interpretation der experimentellen Schätzer auch dadurch erschwert, dass Teilnehmer eines Experimentes das Programm vorzeitig verlassen und Nichtteilnehmer an anderen ähnlichen Programmen teilnehmen können. Wenn es nur um das Problem des vorzeitigen Ausscheidens der Teilnehmer geht, gibt es eine Reihe von Methoden, die dennoch die Schätzung des „Treatment on the Treated"-Effekt ermöglichen (vgl. Bloom 1984 und Heckman/ Smith/ Taber 1998).

Wenn jedoch Individuen der Kontrollgruppe an anderen Programmen teilnehmen, die dem zu evaluierenden ähnlich sind, wird die Situation komplizierter. Die experimentellen Wirkungsschätzer vergleichen dann nämlich das zu evaluierende Programm mit anderen Programmen anstatt mit der „kein Programm"-Alternative. Wenn die anderen Programme ähnlich gut oder ähnlich schlecht arbeiten, wie dasjenige, das experimentell evaluiert werden soll, wird die geschätzte Maßnahmenwirkung null sein, unabhängig davon, wie effektiv das Programm in Wirklichkeit ist verglichen mit der „kein Programm"-Alternative. Heckman, Hohmann, Smith und Khoo (2000) zeigen, dass die korrekte Interpretation in diesem Fall äußerst kompliziert ist und eine Anwendung von nichtexperimentellen Methoden auf die experimentellen Daten notwen-

[5] Vgl. auch Fraker und Maynard (1987) und LaLonde und Maynard (1987).

[6] Abschnitt 8.4 von Heckman, LaLonde und Smith (1999) diskutieren die Einschränkungen dieser Spezifikationstests. Vgl. Regnér (2001) und Raaum und Torp (2001) für neuere Anwendungen der Evaluation europäischer Programmen der aktiven Arbeitsmarktpolitik.

dig ist, um eine Schätzung der Wirkung des Programms relativ zu der „kein Programm"-Alternative zu erhalten.

Zum Abschluss sei angemerkt, dass sich experimentelle Methoden in Nordamerika trotz allem als sehr erfolgreich erwiesen haben. Sie konnten überzeugende Schätzungen der Wirkung von Pilot- und bereits existierenden Programmen liefern. Zugleich musste man mit zunehmender Erfahrung mit Experimenten jedoch erkennen, dass in der Praxis das Design und die Interpretation von Experimenten schwieriger ist, als es auf den ersten Blick scheint. Fragen wie Randomization-Bias, vorzeitiges Ausscheiden von Teilnehmern aus dem Programm und Substitution in alternative Programme innerhalb der Kontrollgruppe erschweren die Entwicklung und Interpretation experimenteller Evaluationen. Diese Nachteile führen keineswegs zu dem Schluss, dass Experimente vermieden werden sollten. Sie sind bloß ein Hinweis darauf, dass das Diktum von Burt Barnow gilt, wonach „Experimente kein Substitut für Denken sind."

3.3 Matching

Innerhalb der nichtexperimentellen Evaluationsforschung sozialer Programme konnte man in den letzten Jahrzehnten Zeuge werden, wie sich verschiedene Schätzmethoden wechselnder Beliebtheit erfreut haben. Vor zwei Jahrzehnten, dominierte das bivariate Selektionsmodell von Heckman (1979) die Literatur. Dieses Modell versucht, Selektionsverzerrungen aufgrund von unbeobachtbaren Variablen zu berücksichtigen, d.h. es berücksichtigt die Tatsache, dass Variablen, die nicht von dem Ökonometriker beobachtet werden, sowohl die Teilnahme an dem Programm als auch die interessierende Ergebnisvariable beeinflussen können. Vor zehn Jahren wurde Selektionsverzerrung aufgrund von unbeobachtbaren Variablen zwar berücksichtigt, die damals vorherrschende Methode war jedoch die Differenz-in-Differenzen-Methode, deren Annahme darin besteht, dass die Selektivität auf einer fixen unbeobachtbaren Komponente der Ergebnisvariable beruht. In den letzten Jahren haben, nicht zuletzt aufgrund reichhaltigerer Daten, die Bedenken bezüglich Selektivitätsverzerrung aufgrund von unbeobachtbaren Daten ab- und zugleich das Interesse an Matching-Methoden zur Programmevaluation zugenommen.

Matching-Methoden sind keineswegs neu. Einige Evaluationsstudien des „U.S. Comprehensive Employment and Training Act" (CETA), die in Barnow (1987) vorgestellt werden, nutzen bereits modifizierte Formen des Matching. Was neu hinzugekommen ist, ist die Nutzung von Propensity Score Matching-Methoden, entwickelt von Rosenbaum und Rubin (1983). Propensity Score Matching nutzt anstelle eines Vektors an beobachtbaren Charakteristika X, die geschätzte Teilnahmewahrscheinlichkeit $P(X)$, um Teilnehmer und Nichtteilnehmer eines Programms zu matchen. Rosenbaum und Rubin (1983) konnten zeigen, dass wenn Matching basierend auf X konsistente Schätzer hervorbringt, dies auch für Matching basierend auf $P(X)$ gilt.

Der Vorteil von Matching basierend auf $P(X)$ anstatt auf X besteht darin, dass $P(X)$ eine skalare Größe ist, während X mehrere Dimensionen haben kann. Wenn X mehrdimensional ist, kann das Matchen problematisch werden, da es sein kann, dass für einige Werte von X für die Teilnehmer keine geeigneten Matches gefunden werden. Dieses Problem wird vermieden (wenn auch nicht vollständig beseitigt, wie ich weiter unten ausführen werde), wenn das Matching basierend auf der skalaren Größe $P(X)$ geschieht.

Matching, basierend auf X oder $P(X)$, ist abhängig von der „Conditional Independence" Annahme. Diese Annahme besagt, dass die Teilnahme an einem Programm, gegeben X oder $P(X)$, unabhängig ist von dem Ergebnis bei Nichtteilnahme (X_0 in der Notation aus Abschnitt 3.1). Dies ist keine triviale Annahme. Sie verlangt, dass alle Variablen, die sowohl die Teilnahme als auch das Ergebnis bei Nichtteilnahme beeinflussen, berücksichtigt werden müssen. Um dieser Annahme in praktischen Anwendungen zu genügen, bedarf es des Zugangs zu reichhaltigen Daten. Sie setzt ferner sorgfältige, auf ökonomischer Theorie basierende Überlegungen voraus, welche Variablen die Teilnahme an einem Programm und die Ergebnisvariable beeinflussen und welche nicht.

An dieser Stelle mag sich der Leser fragen, worin sich Matching-Methoden von einfachen Regressionen unterscheiden. Schließlich bringt auch die Regression einer Ergebnisvariablen auf einen Teilnahmeindikator und auf X einen Schätzer für die Wirkung der Maßnahme hervor. Ich werde hier auf zwei wichtige Unterschiede eingehen. Zunächst ist Matching nichtparametrisch und vermeidet somit die Restriktion, die implizit in linearen Regressionen in Form der linearen Funktion enthalten ist. Dehejia und Wahba (1998) und Smith und Todd (2000), die Matching- und Regressionsschätzungen, basierend auf dem gleichen X, miteinander vergleichen, kommen zu dem Schluss, dass die Berücksichtigung von nichtlinearen Beziehungen eine Reduktion der Verzerrungen bewirken kann. Natürlich nehmen diese Unterschiede auch ab, wenn in der Regression zusätzlich noch Terme höherer Ordnung und Interaktionsterme berücksichtigt werden. Jedoch ist die Aufnahme solcher Terme (außer dem quadrierten Alter oder der quadrierten Ausbildungsdauer) nicht üblich in der Praxis.

Zum Zweiten verdeutlicht Matching das sog. Problem des gemeinsamen Definitionsbereichs. Der Definitionsbereich einer Verteilung ist die Menge aller Werte, für die diese Verteilung eine positive Dichte hat – d.h., die Menge aller Werte, für die die Wahrscheinlichkeit ungleich null ist. Dieser Aspekt ist wichtig für das Matching, da es in empirischen Studien vorkommen kann, dass für bestimmte Werte von X oder $P(X)$ für die Teilnehmergruppe keine Beobachtungen in der Gruppe der Nichtteilnehmer existieren.[7] In diesem Fall stimmt der Definitionsbereich der beiden Stichproben nicht überein. Darüber hinaus wird der gemeinsame Definitionsbereich, d.h. die Menge aller Werte, für die Beobachtungen in beiden Stichproben vorliegen, nicht alle Beobachtungen umfassen. Es bleibt anzumerken, dass es für die Schätzung des „Treatment-on-the-Treated"-Effekts irrelevant ist, ob es Beobachtungen innerhalb der Gruppe der Nichtteilnehmer gibt, für die keine entsprechenden Beobachtungen in der Gruppe der Teilnehmer vorliegen. Für die Schätzung dieses Effektes ist es lediglich notwendig, dass es für jeden Teilnehmer einen entsprechenden Nichtteilnehmer gibt. Für Werte von X, für die $P(X)$ gilt, werden die entsprechenden Teilnehmer notwendigerweise außerhalb des gemeinsamen Definitionsbereichs liegen, da für diesen Fall die Wahrscheinlichkeit, nicht am Programm teilzunehmen, null ist.

[7] Die Auswirkungen des Problems des gemeinsamen Definitionsbereichs hängen implizit von der Bereitschaft des Forschers ab, auch schwache (d.h. nur schwach vergleichbare) Matches hinzunehmen. Vgl. Heckman, Ichimura, Smith und Todd (1998) für eine ausführliche Diskussion dieses Problems und Möglichkeiten seiner Lösung.

Wenn diese soeben beschriebene Bedingung nicht zutrifft und es für einige Teilnehmer keine entsprechenden Nichtteilnehmer zum Matchen gibt, kann für diese Teilnehmer auch keine Maßnahmenwirkung geschätzt werden. Wenn die Maßnahmenwirkung zwischen verschiedenen Individuen variiert, würde man in diesem Fall auf einen falschen Maßnahmeneffekt schließen. In diesem Fall ist die Anwendung von anderen Schätzern ratsam, die Beobachtungen ohne gemeinsamen Definitionsbereich nicht aus der Analyse ausschließen. Matching weist auf das gemeinsame Definitionsbereichproblem in der Weise hin, dass es unmittelbar deutlich macht, wann diese Bedingung verletzt wird. Im Falle des Propensity Score Matchings verdeutlichen einfache Histogramme, wie die bei Heckman, Ichimura, Smith und Todd (1998), das Problem.[8] Im Gegensatz dazu wird dieses Problem bei der einfachen Durchführung einer Regressionsanalyse auf X noch nicht einmal untersucht.

Einige Vorbehalte gelten auch für Matching-Methoden. Ich beziehe mich an dieser Stelle auf die drei Wichtigsten. Zunächst einmal, obwohl Matching dem Forscher die Entscheidung über die Wahl einer adäquaten funktionalen Form abnimmt, nimmt es ihm dennoch nicht die Entscheidung der geeigneten Auswahl an Variablen ab, d.h. der Forscher muss entscheiden, welche Variablen in X aufgenommen werden sollen. Außer einem Vergleich der resultierenden Schätzer mit denen aus einem Experiment gibt es keine Handlungsanweisungen, die dem Forscher bei dieser Entscheidung helfen könnten.[9] Heckman, Ichimura, Smith und Todd (1998) zeigen, dass die Schätzer der Matching-Methoden sehr sensitiv sein können im Hinblick auf die Auswahl an Variablen, die für die Konstruktion von $P(X)$ verwendet werden.

Zum Zweiten kann die Wahl einer bestimmten Matching-Methode bei kleinen Stichproben große Auswirkungen haben. In der Literatur gibt es eine Reihe von unterschiedlichen Matching-Methoden (vgl. Heckman/ Ichimura/ Todd 1997 für eine ausführliche Diskussion). Die gebräuchlichste ist die „Nearest Neighbor Matching"-Methode, bei der derjenige Nichtteilnehmer zu einem Teilnehmer gematcht wird, der diesem bzgl. $P(X)$ am nächsten liegt. Das Ergebnis dieses nächsten Nachbarn dient dann als Approximation an das hypothetische Gegenereignis der Teilnehmers – d.h. es approximiert das Ergebnis, das eingetreten wäre, wenn er oder sie nicht an dem Programm teilgenommen hätte. „Nearest Neighbor Matching" kann ausgestaltet werden, indem mehr als ein nächster Nachbar mit und ohne Zurücklegen berücksichtigt wird, wobei unter „mit Zurücklegen" gemeint ist, dass ein Nichtteilnehmer als Kontrollindividuum für mehr als einen Teilnehmer dienen kann. Alternativen zu diesem „Nearest Neighbor Matching" sind das „Kernel Matching", bei dem ein gewichteter Durchschnitt aus den Zielvariablen der Beobachtungen gebildet wird, die am nächsten an einem Teilnehmer liegen oder „Local Linear Matching", bei dem eine Teilregression für jeden Teilnehmer durchgeführt wird, um das Gegenereignis zu erhalten. Alle diese Methoden sind konsistent, da sie sich mit wachsender Stichprobe mehr und mehr dem Ergebnis bei exaktem Matchen annähern. Bei kleiner Stichprobe können sie jedoch voneinander abweichen, wobei einige Methoden Eigenschaften aufweisen, die sie unter bestimmten Umständen geeigneter erscheinen lassen als andere.

Drittens ist es wichtig, die richtigen Standardfehler zu ermitteln. Die Schätzung der Propensity Scores (wenn Propensity Score Matching gebraucht wird) und das Matching selbst tragen beide zu einer erhöhten Variation bei, die über die normale Stichprobenvariation (vgl. die Diskussion in Heckman/ Ichimura/ Todd 1998) hinausgeht. Im Falle des „Nearest Neighbor Matchings" mit einem nächsten Nachbarn, führt die Betrachtung der gematchten Kontrollgruppe als gegeben, zu einer Unterschätzung der Standardfehler. In der Praxis berechnen daher die meisten Forscher Standardfehler mit der Bootstrap-Methode.

In den letzten Jahren sind einige Arbeiten veröffentlicht worden, die unter Zuhilfenahme von experimentellen Daten, die Leistungsfähigkeit von Matching-Methoden zu evaluieren versuchen. Dabei haben sich zwei Gruppen herausgebildet, die sich in ihrer Einschätzung voneinander unterscheiden. Die erste Gruppe von Arbeiten – Heckman, Ichimura, Smith und Todd (1996, 1998) und Heckman, Ichimura und Todd (1997) – nutzen Daten aus der „U.S. National JTPA Study". Diese Arbeiten kommen zu dem Schluß, dass die Verzerrung der Schätzer bei Matching-Methoden beträchtlich ist und in der gleichen Größenordnung liegt wie die geschätzte Maßnahmenwirkung bei Nutzung von experimentellen Schätzern selbst. Im Gegensatz dazu kommen Dehejia und Wahba (1998, 1999) unter Verwendung von Daten aus dem „U.S. National Supported Work Demonstration" zu einem etwas optimistischeren Schluss. Sie wenden „Propensity Score Matching"-Methoden auf eine Teilstichprobe der Daten an, die von LaLonde (1986) genutzt wurden. In ihrer gewählten Spezifikation fällt die Verzerrung der Matching-Methoden deutlich geringer aus. Smith und Todd (2000) argumentieren jedoch, dass die Ergebnisse von Dehejia und Wahba in erheblichem Maße von ihrer Wahl der Teilstichprobe und der X Variablen abhängen. Änderungen in diesem Zusammenhang führen zu Ergebnissen, die mehr denen entsprechen würden, die unter Rückgriff auf die Daten des JTPA Experimentes gewonnen wurden.

3.4 Allgemeine Gleichgewichtseffekte

Allgemeine Gleichgewichtseffekte treten auf, wenn ein Programm auch andere Personen außer den Teilnehmern beeinflusst. Ein Programm, dass z.B. Langzeitarbeitslose bei ihrer Suche nach Arbeit unterstützt, mag zwar die Geschwindigkeit erhöhen, mit der die Teilnehmer Arbeit finden, zugleich kann es jedoch auch die Wiederbeschäftigungschancen von Kurzzeitarbeitslosen verringern. Dieser Effekt wird *Verdrängungseffekt* genannt (vgl. z.B. Calmfors 1994). In diesem Beispiel nehmen Langzeitarbeitslose, die aufgrund ihrer Teilnahme an dem Programm ihre Sucheffizienz verbessern konnten, die Arbeitsplätze ein, die ansonsten von Kurzzeitarbeitslosen besetzt worden wären. Mit diesem Effekt hängen auch der *Substitutionseffekt*[10], bei denen z.B. Lohnkostenzuschüsse für eine Gruppe von Arbeitern Arbeitgeber veranlasst, diese für andere Arbeiter zu substituieren, und der *Nettowohlfahrtsverlust-Effekt* zusammen, bei dem Aktivitäten auch ohne deren Subventionierung stattgefunden hätten. Calmfors (1994) weist zusätzlich auf die Notwendigkeit hin, *Steuereffekte* zu berücksichtigen, bei denen Steuern, die zur Finanzierung eines Programms erhoben wurden, Entschei-

[8] Vgl. die Figur 2 im ersten Fall und die Figuren 1 und 2 im zweiten Fall.

[9] Der „Balancing Test" vorgeschlagen von Rosenbaum und Rubin (1983) und angewandt von Dehejia und Wahba (1998,1999) und von Lechner (1999) hilft dem Forscher bei der Entscheidung, ob er Terme höherer Ordnung und Interaktionsterme, gegeben ein bestimmtes X berücksichtigen sollte oder nicht. Er hilft ihm jedoch nicht bei der Auswahl der Variablen, die in X aufgenommen werden sollen. Vgl. auch die Diskussion in Smith und Todd (2000).

[10] Diese Substitutionseffekte unterscheiden sich, trotz der ähnlichen Terminologie, konzeptionell von denen, die im Rahmen von sozialen Experimenten diskutiert wurden.

dungen sowohl der Teilnehmer als auch der Nichtteilnehmer beeinflussen. Eine vollständige Programmanalyse sowohl hinsichtlich der Nutzen und Kosten als auch distributiver Effekte muss diese allgemeinen Gleichgewichtseffekte berücksichtigen.[11]

Allgemeine Gleichgewichtseffekte sind nur in bestimmten Zusammenhängen von Bedeutung. Sie werden eine wichtigere Rolle spielen bei der Evaluation großer Programme (im Verhältnis zu der relevanten Grundgesamtheit) als bei der Evaluation kleiner Programme. Ein kleines Demonstrationsprogramm, das 100 Individuen in einem großen, städtischen Arbeitsmarkt umfasst, wird somit keine nennenswerten allgemeinen Gleichgewichtseffekte erzeugen. Auf der anderen Seite wird ein Programm, dass Universitätsstudenten hohe Unterstützungsleistungen bietet, höchstwahrscheinlich erhebliche allgemeine Gleichgewichtseffekte auslösen. Natürlich wird ein Programm, dass keine partiellen Gleichgewichtseffekte aufweist, auch keine allgemeinen Gleichgewichtseffekte hervorbringen können. Ein Trainingsprogramm, das nicht zu einer Erhöhung des Humankapitals seiner Teilnehmer beiträgt, wird auch nicht dazu führen, dass Nichtteilnehmer aus dem Arbeitsmarkt verdrängt werden (obwohl die Steuereinnahmen, die notwendig waren, um dieses Programm zu finanzieren, eine Änderung des Arbeitsangebotsverhaltens bewirken können).

Allgemeine Gleichgewichtseffekte verursachen Probleme für die Evaluationsforschung, weil die partiellen Methoden, die meistens angewandt werden, diese Effekte überhaupt nicht berücksichtigen oder, und vielleicht noch schlimmer, von ihnen verzerrt werden. Um zu sehen, wie diese Probleme entstehen können, sei die Evaluation eines Trainingsprogramms betrachtet, bei der das Einkommen einer Teilnehmerstichprobe und einer Kontrollgruppe verglichen wird. Wenn das Programm einen Verdrängungseffekt erzeugt, wird sich dieser Effekt bemerkbar machen in einem geringeren durchschnittlichen Einkommen der Mitglieder der Kontrollgruppe, von denen einige aus dem Arbeitsmarkt verdrängt wurden. Dies führt zu einer nach oben verzerrten Schätzung der Maßnahmenwirkung des Programms auf seine Teilnehmer. Aufgrund des Verdrängungseffekts ist die Wirkung auf die Teilnehmer eine nach oben verzerrte Schätzung der gesamtwirtschaftlichen Wirkung des Programms. Es sei angemerkt, dass dieses Problem auch bei Durchführung sozialer Experimente bestehen bleibt.

Um die potenzielle Bedeutung von allgemeinen Gleichgewichtseffekten bei der Politikevaluation zu illustrieren und um ein Gespür für die Größenordnungen zu vermitteln, sollen im Folgenden drei Beispiele betrachtet werden. Die beiden Ersten nutzten Daten des „U.S. Unemployment Insurance Bonus"-Experiments (UI), das in Meyer (1995) sorgfältig untersucht wurde. In diesem Bonusexperiment erhielten Bezieher von Arbeitslosengeld, die während einer bestimmten Periode – relativ kurz verglichen mit U.S. Standards und extrem kurz verglichen mit europäischen – nach dem Bezug von Arbeitslosengeld einen Arbeitsplatz finden und für eine Mindestzeit (normalerweise vier Monate) halten konnten, einen Geldzuschuss.

Das erste Beispiel, das betrachtet werden soll, geht auf Meyer (1995) zurück. Er weist darauf hin, dass in einem permanenten UI Bonusprogramm die Existenz des Bonus und die Regeln seiner Anwendung weithin bekannt werden würden. Daraus resultiert eine Änderung sowohl des Arbeiter- als auch des Unternehmenverhaltens hinsichtlich verschiedener Aspekte. Personen, die nur für eine kurze Zeit arbeitslos sind und einen Anspruch auf Arbeitslosengeld hätten, würden dieses aufgrund fixer Kosten vielleicht nicht in Anspruch nehmen. Diese fixen Kosten können sich dabei in der notwendigen Zeit und dem Aufwand, Arbeitslosengeld zu erhalten und vielleicht auch aufgrund einer möglichen Stigmatisierung, niederschlagen. Der Bonus würde dazu führen, dass einige dieser Personen sich um Arbeitslosengeld bemühen und dieses evtl. auch erhalten. Dies ist ein klassisches Beispiel für einen Nettowohlfahrtsverlust-Effekt, bei dem Personen einen Bonus erhalten, ohne dass sich dadurch die Länge ihrer Arbeitslosigkeit verkürzen würde. Dieser allgemeine Gleichgewichtseffekt würde zu einer Reduktion des Nettoeffekts des Programms relativ zu dem geschätzten Effekt führen.

In dem zweiten Beispiel schätzen Davidson und Woodbury (1993) ein strukturelles Mortensen Pissarides-Suchmodell, um den Verdrängungseffekt des Bonusprogramms zu schätzen. Sie finden einen beträchtlichen Verdrängungseffekt unter Arbeitslosen, die keinen Anspruch auf Arbeitslosengeld haben (und somit auch keinen Anspruch auf den Bonus), da sie in dem vorhergehenden Jahr zu wenig gearbeitet haben. Ihre Ergebnisse deuten darauf hin, dass 30 bis 60 Prozent der Bruttowirkung – d.h. des partiellen Gleichgewichtseffekts, so wie er aufgrund des Experiments zur Evaluation des Bonusprogramms geschätzt wurde – kompensiert wird durch den Verdrängungseffekt.

Das dritte Beispiel stammt von Heckman, Lochner und Taber (1998). Sie untersuchen ein Programm, mit dem der Besuch einer Fachhochschule oder Universität unterstützt werden soll. Sie entwickeln eine Modell mit rationalen Erwartungen, perfekter Voraussicht und überlappenden Generationen der U.S. Wirtschaft, das zudem heterogene Fähigkeiten (Ausbildungsniveaus in ihrem Fall) mit separaten und endogenen Preisen berücksichtigt. Unter Verwendung dieses Modellrahmens simulieren sie die Auswirkungen einer einkommensneutralen Erhöhung der gegenwärtigen finanziellen Unterstützungsleistung für Fachhochschul- oder Universitätsstudenten um 500$. Ihr partieller Gleichgewichtseffekt der Erhöhung der Teilnehmerzahl, berechnet bei fixen Einkommen, beträgt 5,3 Prozent im steady-state. Dieses Ergebnis steht im starken Gegensatz zu der Erhöhung unter Berücksichtigung von allgemeinen Gleichgewichtseffekte. Wenn man eine Veränderung der relativen Einkommen zulässt, beträgt der Effekt nur noch 0,46 Prozent. Der große Unterschied kommt dadurch zustande, dass durch eine zunehmende Zahl an Absolventen von Fachhochschulen oder Universitäten deren Einkommen im Arbeitsmarkt sinken, und das Einkommen der knapper werdenden Schulabgänger steigen würde. Diese Veränderungen in den relativen Einkommen dämpfen den Effekt der Unterstützungsleistungen – nach ihren Berechnungen um mehr als 90 Prozent.

Im Zusammenhang mit allgemeinen Gleichgewichtseffekten treten zwei weitere Problemfelder auf. Zunächst gewinnen eine Reihe weiterer Parameter an Bedeutung. Im allgemeinen Gleichgewichtskontext wird der Forscher neben den Parametern, die in Abschnitt 3.1 diskutiert wurden, auch noch an dem

[11] Es sei darauf hingewiesen, dass allgemeine Gleichgewichtseffekte sich von dem unterschieden, was manchmal als „Makroeffekt" bezeichnet wird und unter dem die Beeinflussung der Programmeffektivität durch den gesamtwirtschaftlichen Zustand einer Volkswirtschaft verstanden wird. Ein Programm mag z.B. eine höhere Effektivität haben, wenn die Arbeitslosenquote bei vier Prozent anstatt bei zehn Prozent liegt. Solche Effekte können unter bestimmten Umständen wichtig sein, zählen aber nicht zu den allgemeinen Gleichgewichtseffekten, wie sie in diesem Abschnitt definiert wurden.

Einfluss des Programms auf Nichtteilnehmer interessiert sein. Dieser Effekt kann dabei in zwei Teile zerlegt werden, z.B. in die Effekte, die über den Arbeitsmarkt und diejenigen, die über das Steuersystem wirken. Unter bestimmten Umständen, wie z.B. bei Heckman, Lochner und Taber (1998) beschrieben, können verschiedene Varianten des Local Average Treatment Effekts (LATE), definiert in Abschnitt 3.1, konstruiert werden. In ihrem Modell führt die Unterstützungspolitik dazu, dass einige Individuen von der Universität zur Fachhochschule wechseln und andere von der Fachhochschule zur Universität. Sie definieren LATE für jede Gruppe, wie auch einen Gesamt-LATE, bestehend aus einem gewichteten Durchschnitt dieser beiden Gruppen.

Das zweite Problemfeld kreist natürlich um das Problem, wie diese allgemeinen Gleichgewichtseffekte geschätzt werden können. Ein Teil der Literatur nutzt Unterschiede in dem Programmumfang verschiedener Zuständigkeitsbereiche, um den Effekt zu schätzen. Ein neueres Beispiel hierzu ist Forslund und Krueger (1994). Der andere Teil der Literatur schätzt strukturelle allgemeine Gleichgewichtsmodelle. Beide Arbeiten von Davidson und Woodbury (1993) und Heckman, Lochner und Taber (1998) nutzen diese Modelle. Sie weisen den Vorteil auf, dass sie explizit Annahmen über die den allgemeinen Gleichgewichtseffekten zugrundeliegenden Mechanismen treffen. Sie bieten zudem einen Modellrahmen, der die Schätzung einer Reihe von interessierenden Parametern möglich macht. Der größte Nachteil dieser Modelle, abgesehen von deren rechentechnischer und konzeptioneller Komplexität, sind die strengen Annahmen, die sie über die funktionale Form der ökonomischen Beziehungen und der wichtigsten ökonomischen Parameter treffen.

Da strukturelle allgemeine Gleichgewichtsmodelle erst in den letzten Jahren in größerer Zahl in der Evaluationsliteratur auftauchen, bleiben ihre Schlussfolgerungen kontrovers und die Frage nach dem Vorteil gegenüber traditionelleren Methode (insbesondere im Hinblick auf die höheren Kosten ihrer Implementierung) eine offene Frage. Was jedoch unbestritten ist, unbeachtet der weitgehenden Ignoranz der Literatur, ist die Bedeutung von allgemeinen Gleichgewichtseffekten für die Evaluation von Instrumenten der aktiven Arbeitsmarktpolitik.

4. Schlussfolgerung

Viele Länder, die lange Zeit Fragen der Evaluation keine Aufmerksamkeit geschenkt haben, beginnen nun, ihre Meinung zu ändern. In dem ersten Teil dieser Arbeit habe ich die Organisationsstruktur der Evaluationsindustrie in Nordamerika beschrieben, dem Land, in dem angewandte Evaluation und die Sammlung von Daten am meisten vorangeschritten ist. Ich habe auf Eigenschaften dieser Industrie hingewiesen, die eine Schlüsselrolle spielen, wenn es darum geht, die Qualität von Evaluationen zu sichern, und die von daher geeignet sind, auch in anderen Ländern übernommen zu werden.

Sogar in Nordamerika haben die rapiden methodischen Entwicklungen in der Evaluationsforschung in den letzten zwei Jahrzehnten die Bemühungen des Staates in der Datensammlung und die Bereitschaft der Praxis, diese Methoden zu übernehmen, übertroffen. Im zweiten Teil dieser Arbeit habe ich einen kurzen Überblick über einige neuere methodische Entwicklungen und deren Auswirkungen für praktische Evaluationen und Politikmaßnahmen gegeben. Ich habe ferner auf die entsprechende Literatur in diesem Zusammenhang verwiesen. Die wichtigste Schlussfolgerung ist, dass es noch sehr viel Raum für Verbesserungen in der praktischen Evaluation gibt.

Literaturverzeichnis

Angrist, Joshua/ Alan Krueger (1999): Empirical Strategies in Labor Economics. In: Orley Ashenfelter/ David Card (eds.), Handbook of Labor Economics, Volume 3A. Amsterdam: North-Holland, 1277-1366.

Barnow, Burt (1987): The Impact of CETA Programs on Earnings: A Review of the Literature. In: Journal of Human Resources, 22, 157-193.

Björklund, Anders/ Håkan Regnér (1996): Experimental Evaluation of European Labour Market Policy. In: Günther Schmid/ Jacqueline O'Reilly/ Klaus Schömann (eds.), International Handbook of Labour Market Policy and Evaluation. Brookfield, VT: Edward Elgar, 89-114.

Black, Dan/ Jeffrey Smith/ Mark Berger/ Brett Noel (2000): Is the Threat of Reemployment Services More Effective Than the Services Themselves: Experimental Evidence from the UI System. Unpublished manuscript, University of Western Ontario.

Bloom, Howard (1984): Accounting for No-Shows in Experimental Evaluation Designs. In: Evaluation Review, 82(2), 225-246.

Bloom, Howard/ Larry Orr/ Stephen Bell/ George Cave/ Fred Doolittle/ Winston Lin/ Johannes Bos (1997): The Benefits and Costs of JTPA Title II-A Programs: Findings from the National Job Training Partnership Act Study. In: Journal of Human Resources, 32(3), 549-576.

Burtless, Gary (1995): The Case for Randomized Field Trials in Economic and Policy Research. In: Journal of Economic Perspectives, 9(2), 63-84.

Burtless, Gary/ Larry Orr (1996): Are Classical Experiments Needed for Manpower Policy. In: Journal of Human Resources, 21, 606-639.

Calmfors, Lars (1994): Active Labor Market Policy and Unemployment – A Framework for the Analysis of Crucial Design Features. In: OECD Economic Studies, 22(1), 7-47.

Couch, Kenneth (1992): New Evidence on the Long-term Effects of Employment and Training Programs. In: Journal of Labor Economics, 10(4), 380-388.

Currie, Janet/ Duncan Thomas (1995): Does Head Start Make a Difference? In: American Economic Review, 85(3), 341-364.

Davidson, Carl/ Stephen Woodbury (1993): The Displacement Effects of Reemployment Bonus Programs. In: Journal of Labor Economics, 11(4), 575-605.

Dehejia, Rajeev/ Sadek Wahba (1998): Propensity Score Matching Methods for Non-Experimental Causal Studies. NBER Working Paper #6829.

Dehejia, Rajeev/ Sadek Wahba (1999): Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. In: Journal of the American Statistical Association, 94(448), 1053-1062.

Dickinson, Kathryn/ Terry Johnson/ Richard West (1987): An Analysis of the Sensitivity of Quasi-Experimental Net Estimates of CETA Programs. In: Evaluation Review, 11, 452-472.

Dolton, Peter/ Donal O'Neill (1996): Unemployment Duration and the Restart Effect: Some Experimental Evidence. In: Economic Journal, 106(435), 387-400.

Doolittle, Fred/ Linda Traeger (1990): Implementing the National JTPA Study. New York: Manpower Demonstration Research Corporation.

Forslund, Anders/ Alan Krueger (1997): An Evaluation of the Swedish Active Labor Market Policy. In: Richard Freeman/ Birgitta Swedenborg/ Robert Topel (eds.), The Welfare State in Transition. Chicago: University of Chicago Press, 267-298.

Fraker, Thomas/ Rebecca Maynard (1987): The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. In: Journal of Human Resources, 22(2), 194-227.

Heckman, James (1979): Sample Selection Bias as a Specification Error. In: Econometrica, 47(1), 153-161.

Heckman, James (1997): Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator. In: Journal of Human Resources, 32(3), 441-461.

Heckman, James/ Neil Hohmann/ Jeffrey Smith/ Michael Khoo (2000): Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. In: Quarterly Journal of Economics, 115(2), 651-694.

Heckman, James/ V. Joseph Hotz (1989): Choosing Among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training. In: Journal of the American Statistical Association, 84(408), 862-874.

Heckman, James/ Hidehiko Ichimura/ Jeffrey Smith/ Petra Todd (1996): Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method. In: Proceedings of the National Academy of Sciences, 93(23), 13416-13420.

Heckman, James/ Hidehiko Ichimura/ Jeffrey Smith/ Petra Todd (1998): Characterizing Selection Bias Using Experimental Data. In: Econometrica, 66(5), 1017-1098.

Heckman, James/ Hidehiko Ichimura/ Petra Todd (1997): Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. In: Review of Economic Studies, 64(4), 605-654.

Heckman, James/ Robert LaLonde/ Jeffrey Smith (1999): The Economics and Econometrics of Active Labor Market Programs. In: Orley Ashenfelter/ David Card (eds.), Handbook of Labor Economics, Volume 3A. Amsterdam: North-Holland, 1865-2097.

Heckman, James/ Lance Lochner/ Christopher Taber (1998): Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents. In: Review of Economic Dynamics, 1(1), 1-58.

Heckman, James/ Jeffrey Smith (1993): Assessing the Case for Randomized Evaluation of Social Programs. In: Karsten Jensen/ Per Kongshoj Madsen (eds.), Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives. Copenhagen: Danish Ministry of Labour, 35-96.

Heckman, James/ Jeffrey Smith (1995): Assessing the Case for Social Experiments. In: Journal of Economic Perspectives, 9(2), 85-110.

Heckman, James/ Jeffrey Smith (1996a): Experimental and Nonexperimental Evaluation. In: Günther Schmid/ Jacqueline O'Reilly/ Klaus Schömann (eds.), International Handbook of Labour Market Policy and Evaluation. Brookfield, VT: Edward Elgar, 37-88.

Heckman, James/ Jeffrey Smith (1996b): Social Experiments: Theory and Evidence. In: Ökonomie und Gesellschaft, Jahrbuch 13: Experiments in Economics - Experimente in der Ökonomie. Frankfurt/Main, New York: Campus Verlag, 186-213.

Heckman, James/ Jeffrey Smith (2000): The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study. In: David Blanchflower/ Richard Freeman (eds.), Youth Employment and Joblessness in Advanced Countries. Chicago: University of Chicago Press for NBER, 331-356.

Heckman, James/ Jeffrey Smith/ Nancy Clements (1997): Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. In: Review of Economic Studies, 64(4), 487-537.

Heckman, James/ Jeffrey Smith/ Christopher Taber (1998): Accounting for Dropouts in Evaluations of Social Programs. In: Review of Economics and Statistics, 80(1), 1-14.

Imbens, Guido/ Joshua Angrist (1994): Identification and Estimation of Local Average Treatment Effects. In: Econometrica, 62(4), 467-476.

LaLonde, Robert (1986): Evaluating the Econometric Evaluations of Training Programs with Experimental Data. In: American Economic Review, 76(4), 604-620.

LaLonde, Robert/ Rebecca Maynard (1987): How Precise Are Evaluations of Employment and Training Programs: Evidence from a Field Experiment. In: Evaluation Review, 11, 428-451.

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

Meyer, Bruce (1995): Lessons from the U.S. Unemployment Insurance Experiments. In: Journal of Economic Literature, 33(1), 91-131.

Michalopoulos, Charles/ David Card/ Lisa Gennetian/ Kristen Harknett/ Philip Robins (2000): The Self-Sufficiency Project at 36 Months: Effects of a Financial Work Incentive on Employment and Income. Ottawa: Social Research and Demonstration Corporation.

Raaum, Oddbjørn/ Hege Torp (2001): Labour Market Training in Norway – Effect on Earnings. In: Labour Economics. Forthcoming.

Regnér, Håkan (2001): A Nonexperimental Evaluation of Training Programs for the Unemployed in Sweden. In: Labour Economics. Forthcoming.

Rosenbaum, Paul/ Donald Rubin (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects. In: Biometrika, 70(1), 41-55.

Rossi, Peter/ Howard Freeman (1993): Evaluation: A Systematic Approach, 5th Edition. Newbury Park, CA: Sage.

Torp, Hege/ Oddbjørn Raaum/ Erik Hernæs/ Harald Goldstein (1993): The First Norwegian Experiment. In: Karsten Jensen/ Per Kongshoj Madsen (eds.), Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives. Copenhagen: Danish Ministry of Labour, 97-140.

Riddell, Craig (1991): Evaluation of Manpower and Training Programmes: The North American Experience. In: OECD (ed.), Evaluating Labor Market and Social Programs. Paris: OECD. 43-72.

Smith, Jeffrey/ Petra Todd (2000): Is Propensity Score Matching the Answer to LaLonde's Critique of Nonexperimental Estimators? Unpublished manuscript, University of Western Ontario.

U.S. General Accounting Office (1996): Job Training Partnership Act: Long-Term Earnings and Employment Outcomes. Report No. GAO/HEHE-96-40.

White, Michael/ Jane Lakey (1992): The Restart Effect: Evaluation of a Labour Market Programme for Unemployed People. London, UK: Policy Studies Institute.