Quelle: Gerhard Kleinhenz (Hrsg.) (2002): IAB-Kompendium Arbeitsmarkt- und Berufsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 250, S. 415-427.

Soziale Experimente und Modellversuche: Ein Beitrag zur Evaluation von Neuansätzen in der Arbeitsmarktpolitik

Michael Müntnich, Frank Wießner

Forschung als ein gesetzlicher Auftrag

Mit der Einführung des SGB III[1] im Jahre 1998 wurde die Bundesanstalt für Arbeit erstmals ausdrücklich aufgefordert, Lage und Entwicklung des Arbeitsmarktes sowie die Wirkungen der aktiven Arbeitsförderung zu beobachten, zu untersuchen und auszuwerten. Spätestens mit diesem Schritt wurde der Grundstein für eine institutionalisierte und auf Dauer angelegte Evaluation arbeitsmarktpolitischer Maßnahmen gelegt. In unmittelbarer Folge darauf wurden entsprechende Forschungsaktivitäten des IAB deutlich verstärkt.[2]

Die Ansprüche an die Arbeitsmarktforschung zum Zwecke der Politikberatung reichen jedoch viel weiter: Während der § 280 SGB III sich mit der Erstellung von Statistiken, Arbeitsmarkt- und Berufsforschung allgemeiner Art und einschlägiger Berichterstattung begnügt, wird von den Politikverantwortlichen in zunehmendem Maße auch Unterstützung bei der Entwicklung und Ausgestaltung des künftigen arbeitsmarktpolitischen Instrumentariums eingefordert. Zudem wirft auch und vor allem der immense Ressourceneinsatz zur Bekämpfung der Arbeitslosigkeit Fragen nach der Wirksamkeit der Aufwendungen auf. Im Jahre 2000 wurden beispielsweise für die Ermessensleistungen der aktiven Arbeitsförderung nach dem SGB III mehr als 27,3 Mrd. DM (entspricht knapp 14 Mrd. €) ausgegeben. Für das derzeit noch laufende Haushaltsjahr 2001 liegt der Ansatz für den so genannten *Eingliederungstitel*[3] bei gut 14,2 Mrd. €.

Gründe für die Entwicklung neuer Ansätze in der Arbeitsmarktpolitik gibt es mehr als genug. Weitaus schwieriger gestaltet sich jedoch aus Sicht des Gesetzgebers die Aufgabe, aus der Vielzahl der Vorschläge, die in die aktuelle politische Diskussion einfließen, jene auszuwählen, die nicht nur Erfolge versprechen, sondern zugleich bei gegebenem Mit-

---

1 Drittes Buch Sozialgesetzbuch.

2 Aller Voraussicht nach werden die kodifizierten Impulse für die Wirkungsforschung durch das neue JOB-AQTIV-Gesetz zusätzlich verstärkt werden.

3 Der so genannte Eingliederungstitel im Haushalt der Bundesanstalt für Arbeit enthält die Ermessensleistungen der aktiven Arbeitsförderung. Darunter fallen unter anderem Trainingsmaßnahmen, Mobilitätshilfen, Unterhaltsgeld, Eingliederungs- und Einstellungszuschüsse, Sozialplanmaßnahmen, ABM, die Freie Förderung, u. a. m.

teleinsatz Vorteile maximieren ohne allzu große negative Begleiterscheinungen, so genannte Emergenzeffekte, aufzuwerfen. Wie die Erfahrungen aus verschiedenen Modellversuchen zeigen, wird es zunehmend schwierig, solche Policy-Ansätze zu finden, die gleichermaßen marktkonform wie auch anreizkompatibel sowie verträglich mit bestehenden arbeitsmarktpolitischen Instrumentarium sind und zugleich die Initiativen und Aktivitäten anderer Politikfelder integrieren können. So müssen beispielsweise Kombilöhne im Niedriglohnbereich mit dem Sozialhilfebezug harmonisiert werden und zugleich muss sichergestellt sein, dass andere Förderprogramme, wie etwa Arbeitsbeschaffungs- oder Strukturanpassungsmaßnahmen, nicht wesentlich attraktivere Konditionen für Arbeitgeber oder Arbeitnehmer bieten.

Lösungen gesucht - Modelle versucht

Der so genannten Ex ante-Evaluation, der Bewertung möglicher Policy-Effekte durch Vorüberlegungen, sind logischerweise Grenzen gesetzt. Wo die bisherigen nachvollziehbaren Erfahrungen enden, fangen oft Visionen oder Spekulationen an. Und auch die „klassische“ wissenschaftliche Begleitforschung, sei sie nun „ongoing“ oder „ex post“, kann zu innovativen Ansätzen nur begrenzten Erkenntnisgewinn beisteuern. Denn um auswertbare Daten zu generieren, muss ein politisches Programm erst längere Zeit - oft sogar über Jahre hinweg - in der Praxis erprobt werden. Weil solche Untersuchungsansätze retrospektiv sind, können politikrelevante Evaluationsbefunde kaum zeitnah, sondern immer nur mit einem gewissen time-lag versehen, bereitgestellt werden. Dies ist insbesondere dann der Fall, wenn zur Untersuchung der Nachhaltigkeit von Programmeffekten oder einfach nur zur Betrachtung von Entwicklungen im Zeitverlauf tiefergehende Analysen einen größeren Beobachtungszeitraum erfordern. Solange jedoch der Ausgang einer Initiative ungewiss ist, kann eine flächendeckende Einführung erhebliche Risiken bergen. Neben den allgemeinen monetären Belastungen für Beitragszahler und öffentliche Hände sind hier vor allem Fehlanreize und Fehlallokationen und in deren Gefolge individuelle oder auch kollektive Nachteile zu bedenken. Oftmals ist es im Nachhinein auch schwierig, erkannte Fehlsteuerungen ohne größere Beeinträchtigung zu korrigieren. Was liegt also näher, als innovative Ansätze zunächst im Kleinen, in Form von Modellversuchen, zu erproben?

Das Evaluierungsproblem

Soll also nun anhand eines Pilotprojektes herausgefunden werden, ob der arbeitsmarktpolitische Eingriff auch tatsächlich zielführend ist und demgemäss nachfolgend flächendeckend eingeführt werden sollte, so erfordert dies zunächst eine formale Bewertung des Programmerfolges. Eine solche Bewertung gerät jedoch leicht zum intellektuellen Kraftakt, erfordert sie doch die Auseinandersetzung mit einem überaus komplexen Problem:

Die Ermittlung des *Nettoeffektes* der Förderung setzt nämlich die Kenntnis voraus, wie es dem *einzelnen* Teilnehmer ergangen wäre, hätte er *nicht* an der Maßnahme teilgenommen. Da jedoch eine einzelne Person niemals gleichzeitig in zwei verschiedenen Situationen beobachtet werden kann, ist diese Vergleichsgröße faktisch nicht beobachtbar. Aufgrund der daraus resultierenden *Kontrafaktizität* bleibt Aussagen über mögliche Programmwirkungen oft auch ein hypothetischer Beigeschmack.

Ein vordergründig einfacher Ausweg aus diesem Dilemma könnte in der Bildung einer geeigneten *Vergleichsgruppe* bestehen, d. h. man vergleicht die Maßnahmeteilnehmer einfach mit Nicht-Teilnehmern. Ein solches Verfahren wirft jedoch neue Probleme auf, weil neben den offensichtlichen - und damit kontrollierbaren - Merkmalen der Teilnehmer sicherlich auch noch viele unbeobachtete, weil nicht beobachtbare Unterschiede zwischen Teilnehmern und Nichtteilnehmern vorliegen können.[4] Da eine hinreichende Vergleichbarkeit mehr oder minder willkürlich gegenüber gestellter Stichproben nicht abgesichert ist, erwächst aus dem Evaluierungsproblem mithin ein *Selektionsproblem*. Formal akzeptabel ist die Vergleichsgruppenbildung erst dann, wenn hinreichend detaillierte Informationen über alle Individuen verfügbar sind und zudem Teilnehmer und Nicht-Teilnehmer auch tatsächlich im Hinblick auf diese Charakteristika übereinstimmen.

Selbstselektion und Fremdselektion

Oftmals wirken jedoch bei der Teilnehmerauswahl für ein bestimmtes Programm verschiedene Selektionsmechanismen, die sich der Beobachtung weitgehend entziehen und damit verzerrte Befunde verursachen können.[5] So könnte die *Selbstselektion* von Teilnehmern beispielsweise begründet sein in der Erwartung, mit Hilfe einer bestimmten Maßnahme die individuellen Beschäftigungsaussichten verbessern zu können. Umgekehrt könnte aber auch ein eher negatives Image einer Maßnahme (wie dies etwa immer wieder von ABM berichtet wird) dazu führen, dass bestimmte Personen daran möglichst nicht teilnehmen wollen, beispielsweise um eine (vermeintliche) Stigmatisierung zu vermeiden.

Selektion kann aber auch exogen, also seitens Dritter erfolgen, wie etwa durch das Arbeitsamt oder Maßnahmeträger. Denkbar wäre hier einerseits eine Positivselektion, so genanntes „Creaming“, wodurch überproportional viele Teilnehmer, die bestimmte, mutmaßlich besonders erfolgversprechende Merkmale aufweisen, in den Genuss der Förderung kommen könnten. Eine weitere Variante von *Fremdselektion* könnte darin beste-

---

[4] Diese so genannte unbeobachtete Heterogenität soll vor allem mittels ökonometrischer Schätzverfahren kontrolliert werden.

[5] Vgl. z. B. Schömann 1996: 130.

hen, bevorzugt solche Antragsteller zu fördern, die eher benachteiligt sind. In diesem Falle etwa würde es sich um eine negative Auslese solcher Personen handeln, die zuvor schon aufgrund einer Verknüpfung ungünstiger Risiken arbeitslos geworden und wegen verschiedener vermittlungshemmender Merkmale vielleicht auch längere Zeit geblieben wären.

Ohne explizite Zusatzinformationen ist aus der Forschungsperspektive nicht nachvollziehbar, inwieweit und nach welchen Kriterien innerhalb des gesetzlich vorgegebenen Rahmens eine *zusätzliche* Auswahl hinsichtlich des Zuganges zu einem bestimmten Förderprogramm getroffen wird. Die Abschätzung der direkten (Netto-)Wirkungen der Förderung wird dadurch erschwert und manövriert damit Wirkungsforscher und Programmgestalter gleichermaßen in ein geradezu klassisches Dilemma: Denn gerade die Flexibilität und Offenheit eines Programms, die den Umsetzern Gestaltungsspielräume eröffnen und eine Anpassung an spezifische Situationen ermöglichen sollen, beeinträchtigen dessen Evaluierbarkeit. Zumindest während der Phase modellhafter Erprobung sind im Hinblick auf die Einschätzung der Programmwirkungen auch solche Aspekte zu berücksichtigen.

Auch kann die eigentliche Durchführung eines solchen Experimentalansatzes an verschiedenen Stellen ebenfalls Friktionen aufwerfen, etwa in Gestalt administrativer Probleme bei den Arbeitsämtern. Programmumsetzer könnten unter Umständen (bewusst oder unbewusst) dazu tendieren, der Vergleichsgruppe quasi als „Kompensation" zusätzliche Serviceleistungen anzubieten. Björklund/Regnér bezeichnen dies als *substitution bias*. Auf der Ebene der Programmteilnehmer ist auch denkbar, dass allein das Bewusstsein der Teilnahme an einem Experiment Einfluss auf das Verhalten der Probanden ausüben kann. Diese Art der Verzerrung ist als *„Hawthorne-Effekt"*[6] bekannt.

Die zweite Gruppe macht den Unterschied

Aufgrund der skizzierten Verzerrungsmechanismen ist es nicht möglich, die Programmwirkung durch einen einfachen Vergleich des Endzustandes der Untersuchungspopulation mit ihrem Ausgangszustand zu ermitteln, da ein solches Vorgehen nicht zu eindeutigen Befunden führt. Denn es steht nicht zweifelsfrei fest, ob die registrierten Effekte tatsächlich und ausschließlich auf den Eingriff zurückzuführen sind und keine anderen (Neben-)Ursachen haben. Zugleich ist unklar, ob nicht die *vermeintliche* Wirkung genauso gut ohne den Eingriff hätte eintreten können. Solche „Ein-Gruppen-Pläne" sind gekenn-

[6] Mit „Hawthorne-Effekt" wird das Phänomen bezeichnet, dass bereits das Bewusstsein, an einem wissenschaftlichen Experiment teilzunehmen, Verhaltensänderungen bei den Versuchsteilnehmern hervorrufen kann.

zeichnet von geringer interner Validität[7] bei gleichzeitig weiten Interpretationsspielräumen, die eben gerade durch die Evaluationsforschung eingeengt werden sollen.

Die Beliebigkeit bei der Deutung von Ergebnissen lässt sich im Prinzip leicht reduzieren durch die Einführung einer weiteren Untersuchungspopulation, der so genannten *Kontrollgruppe*. Beim „Zwei-Gruppen-Plan" wird der Maßnahmeeffekt dann ermittelt durch den Vergleich von Treatment-Group versus Non-Treatment-Group. Ein solches Vorgehen bietet zudem den Vorteil, neben den relativ leicht ermittelbaren Bruttoeffekten eines Programms zusätzlich auch dessen Nettoeffekte, d. h. die durch die Maßnahme *zusätzlich* erzielten Wirkungen, bestimmen zu können. Ein schwerwiegendes Alltagsproblem beeinträchtigt jedoch solche Untersuchungsdesigns: Denn in der sozialen Realität kann es „echte" Non-Treatment-Groups" kaum geben. Selbst wenn das Arbeitsamt etwa für diejenigen, die nicht an einer Maßnahme teilnehmen, keinerlei weitere Aktivitäten unternehmen würde (was kaum vorstellbar ist), so bliebe dieser Gruppe gegenüber der Experimentalgruppe doch viel mehr Zeit für Eigeninitiativen. Zudem lauert auch beim Zwei-Gruppen-Plan wiederum eine methodologische Falle: Denn die Schlussfolgerung, die Unterschiede zwischen den beiden Subpopulationen repräsentierten die *tatsächlichen* Treatment-Effekte, ist nur dann zulässig, wenn gewährleistet ist, dass die beiden Gruppen ursprünglich im Hinblick auf *alle untersuchungsrelevanten* Merkmale auch tatsächlich vergleichbar sind. Unkontrollierte Heterogenität beeinträchtigt wiederum die Qualität der Evaluation und die Generalisierbarkeit der Befunde, zugleich ist jedoch auch die Vorstellung exakt identischer Vergleichsgruppen schlichtweg utopisch. Allerdings sinkt die Wahrscheinlichkeit des Auftretens unkontrollierbarer Unterschiede zwischen den beiden Stichproben erheblich, wenn man die Entscheidung über die Gruppenzuordnung dem *Zufall* überlässt.

Randomisierung - der Zufall hilft

Im so genannten *randomisierten* Feldexperiment werden die Untersuchungspersonen *zufällig* auf mindestens zwei Gruppen aufgeteilt. Analog zum Zwei-Gruppen-Plan erhält die Experimentalgruppe das Treatment[8], dessen Effekt getestet werden soll, während die Kontrollgruppe keiner Intervention ausgesetzt wird. Ein probater Weg zur Gewinnung vergleichbarer Experimental- und Kontrollgruppen besteht nun darin, die Untersuchungseinheiten zufallsgesteuert auf diese beiden Gruppen zu verteilen. Hat ein jeder Proband

---

[7] Ein Experiment weist dann interne Validität auf, wenn die Variation der abhängigen Größe eindeutig auf die (manipulative) Variation der erklärenden Variablen zurückgeführt werden und das Ergebnis damit möglichst eindeutig interpretiert werden kann.

[8] Denkbar ist es auch, mit qualitativ unterschiedlichen oder quantitativ abgestuften Treatments zu arbeiten.

a priori die gleiche Chance, in die Treatment-Group bzw. in die Non-Treatment-Group zu gelangen, so besteht auch für jede theoretisch mögliche Kombination aus beobachteten und unbeobachteten Eigenschaften die gleiche Wahrscheinlichkeit, in der einen wie auch in der anderen Gruppe aufzutreten.

Damit stellt das Verfahren sicher, dass die beiden Gruppen *im Durchschnitt* identisch sind und sich zumindest *nicht systematisch* voneinander unterscheiden. Und weil die aus einer solchen Zufallszuweisung resultierende Experimental- und Kontrollgruppe sich eben nur zufällig voneinander unterscheiden, folgt weiterhin, dass auch die relevanten Einflussgrößen zufällig auf die beiden Gruppen verteilt sind und damit keine *systematische*, sondern allenfalls eine zufällige Verzerrung der Befunde verursachen. Beobachtet man nun die interessierenden abhängigen Größen für die Experimental- als auch für die Kontrollgruppe, so können etwaige beobachtete Unterschiede ursächlich auf die Intervention zurückgeführt werden.[9]

Soziale Experimente

Einen möglichen Ausweg aus dem Evaluierungsdilemma bieten praktische Experimente. Als Experiment bezeichnet man allgemein die systematische und wiederholbare Generierung von Daten durch kontrollierte und isolierte Variation von Faktoren, die - idealerweise theoriegeleitet - als relevant für die untersuchten Zustände oder Prozesse betrachtet werden. Dabei ist es essenziell, das Untersuchungsdesign als Zwei-Gruppen-Plan vorzusehen, eine geeignete Vergleichsgruppe schon vorab festzulegen. Im Wesentlichen wird darauf abgezielt, möglichst immer nur eine Determinante kontrolliert, d. h. messbar und nachvollziehbar, zu variieren und dabei gleichzeitig alle weiteren als wesentlich erachteten Einflussfaktoren konstant zu halten. Als mögliche Steuerungsgrößen kommen neben ökonomischen auch außerökonomische, insbesondere psychologische Merkmale in Frage. Konkret würde man beispielsweise wissen wollen, wie das Ergebnis einer arbeitsmarktpolitischen Maßnahme, d. h. der Verbleib der Maßnahmeteilnehmer, beeinflusst wird von Mechanismen zur Teilnehmerselektion, von individuellen Charakteristika der Teilnehmer oder etwa einer Variation von Dauer und Inhalten des Programms oder der Art der Betreuung durch den Maßnahmeträger.

---

[9] Vgl. z. B. Rossi/Freeman/Hofmann 1999: 262. Genau genommen sind hier jedoch noch zusätzliche „Designeffekte“, solche Auswirkungen also, die auf den Untersuchungsplan zurückzuführen sind, zu beachten. Verzerrungen können beispielsweise verursacht werden durch den Test an sich bzw. das Wissen der Probanden um die Test-Situation („Reaktivität“), eingeschränkte bzw. unzureichende Reliabilität und Validität des Untersuchungsinstrumentariums oder auch so genannte „fehlende Werte“ (missing value Problem) bei Einzelmerkmalen. Vgl. auch Bamberg et al. 2000: 123.

Die zum Zwecke der statistischen Absicherung gegen Zufallseinflüsse notwendige Wiederholbarkeit wird durch eine entsprechende Kontrolle der Entscheidungsumwelt erreicht. Gerade hier liegt aber das Problem: Denn nur selten lässt sich die Komplexität wirtschafts- oder sozialwissenschaftlicher Evaluationsfragen soweit reduzieren, dass sie im Rahmen von Laborexperimenten, wie sie aus den Naturwissenschaften bekannt sind, beantwortet werden könnten.[10] In der jüngeren Vergangenheit jedoch zeichnet sich eine neue Entwicklung ab, die kontrollierten Experimenten durchaus ihren Platz in den Sozialwissenschaften zubilligt.

Eine Frage der Ethik

Zugleich stößt hier aber die Methode der Zufallszuordnung an ihre Grenzen. Denn ein Experimentalansatz, der den Vergleich mit einer Situation ohne die Förderung herzustellen sucht, kann in rechtlicher wie auch in ethischer Hinsicht problematisch werden. In der bundesdeutschen Arbeitsförderung herrscht die Auffassung vor, dass individuelle Lebensverläufe nicht durch exogen herbeigeführte Zufallsentscheidungen beeinflusst werden sollten. Auch gilt es als ethisch nicht akzeptabel, Menschen aus wissenschaftlichen Gründen die nach dem aktuellen Kenntnisstand bestmögliche Förderung vorzuenthalten, insbesondere dann nicht, wenn diese für die Teilnehmer möglicherweise die einzige Chance darstellt, auf dem Arbeitsmarkt wieder Fuß zu fassen.[11] Dies ist aber bei den Modellversuchen auch überhaupt nicht der Fall, da die Versuchsteilnehmer von den herkömmlichen Hilfestellungen keineswegs aus Forschungsgründen ausgeschlossen werden. Deshalb sind auch die meisten Modelle als *Ergänzung* oder zumindest *Alternative* zum bisher bestehenden Instrumentarium angelegt. Bei Fördermaßnahmen, auf die im Rahmen bestehender Gesetze quasi ein Rechtsanspruch vorliegt, wäre ein Ausschluss ohnehin kaum möglich.[12] Da die im Experiment erprobten Ansätze zudem stets eine Verbesserung für die Teilnehmer erbringen sollen, stellt sich faktisch eher umgekehrt die Frage, ob das experimentelle Treatment nicht sogar eine Bevorzugung der Teilnehmer gegenüber den Nichtteilnehmern darstellen könnte.

---

[10] Bis in die achtziger Jahre hinein wurde die Auffassung vertreten, dass sich aus der Unmöglichkeit der Durchführung von Laborexperimenten in den Sozialwissenschaften ein grundlegender Unterschied zwischen den Natur- und Sozialwissenschaften ergibt. Editorial zu „Ökonomie und Gesellschaft“, 13. Jahrbuch, 1996: 7.

[11] Vgl. z. B. Burtless 1995: 187.

[12] Zu ethischen Bedenken und rechtlichen Problemen experimenteller Ansätze siehe Björklund/Regnér 1996: 92.

Quasiexperimente in der sozialwissenschaftlichen Forschungspraxis

Viele der Fragestellungen, mit denen die Wirkungsforschung im Bereich der Arbeitsmarktpolitik konfrontiert wird, bieten jedoch aufgrund der faktisch vorliegenden Rahmenbedingungen keine hinreichenden Gestaltungsmöglichkeiten für entsprechende Kontrollgruppen-Designs. Eine hinreichende Randomisierung von Experimental- und Kontrollgruppe ist deshalb nicht bzw. allenfalls mit Einschränkung möglich. Wenn die Zuordnung der Untersuchungsteilnehmer zu den unabhängigen Variablen bereits exogen gegeben ist und sich die Zufallsauswahl lediglich auf die interne Zusammensetzung der jeweils zu vergleichenden Teilpopulationen beschränkt, werden solche Untersuchungsdesigns als „natürliche Experimente“ oder „Quasiexperimente“ bezeichnet.

Da sich jedoch die untersuchbaren Unterschiede eben auf bereits existierende, also nicht erst künstlich herzustellende Teilpopulationen beziehen, für die wiederum unterschiedliche individuelle Einstiegsvoraussetzungen schon vor der Intervention anzunehmen sind, erscheint die Heterogenität zwischen der Experimental- und der Kontrollgruppe unvermeidlich. Als Folge ist deshalb praktisch zwangsläufig mit einer Art von Selektionsverzerrung zu rechnen. Die spannende Frage, die sich daraus für den Sozialwissenschaftler ergibt, lautet nun: Welcher Art ist diese Verzerrung?

Entsprechend versucht man, in Ermangelung einer „echten“ Kontrollgruppe, vergleichbare Datensätze mit unterschiedlichen Methoden nach Untersuchungseinheiten (d. h. Personen), die denen der Maßnahmepopulation ähneln, zu durchsuchen und auf diesem Wege zu einer „quasi-experimentellen“ Kontrollgruppe oder zumindest Vergleichsgruppe zu gelangen.

Vergleiche mit solchermaßen konstruierten Gruppen weisen jedoch wiederum die bekannten Schwächen hinsichtlich ihrer internen Validität auf. Die Untersuchungsergebnisse sind hier - konzeptionell bedingt - viel freier interpretierbar als bei „echten“ Experimenten und dementsprechend konkurriert häufig eine Vielzahl von Alternativerklärungen miteinander. Zudem bergen quasiexperimentelle Ansätze die Gefahr, dass unabhängige Variablen mit anderen ebenfalls relevanten Einflussgrößen korrespondieren, ohne dass dies in der Untersuchung deutlich wird. Mit Hilfe verschiedener Techniken sollen vor allem personengebundene *Störvariablen*, aber auch andere *Kontextgrößen*, wie z. B. die regionale Arbeitsmarktsituation, der Familienkontext oder auch Untersuchungseffekte möglichst unter Kontrolle gehalten werden.[13] Zu den gebräuchlichsten Methoden zählen:

[13] Vgl. Bortz 1984: 404.

- Konstanthaltung der Störgrößen; jedoch oft auf Kosten der externen Validität;[14]
- Parallelisierung der Stichproben in Bezug auf die Störgrößen (d. h. die beiden Stichproben sollen hinsichtlich der Störgrößen annähernd gleiche Mittelwerte und Streuungen aufweisen), ebenfalls zulasten der externen Validität und
- Bildung von so genannten „matched samples", indem die Untersuchungseinheiten von Experimental- bzw. Kontrollgruppe einander paarweise in Bezug auf die zu kontrollierenden Merkmale zugeordnet werden.

Insbesondere beim letztgenannten Verfahren ist es das Ziel, retrospektiv auf Grundlage der über die Teilnehmer vorliegenden Informationen, eine *glaubhafte* kontrafaktische Vergleichsgruppe zu konstruieren. James Heckman, Nobelpreisträger in Ökonomie, hat das statistische Verfahren der paarweisen Zuordnung, das Matching, neu aufgegriffen und auf Evaluationsstudien angewandt. Ist die Information über die Individuen detailliert genug, so entsteht dadurch retrospektiv eine akzeptable Vergleichsgruppe, die einer randomisierten Kontrollgruppe ähnelt. Klar ist aber auch, dass die Qualität der Matches mit den Details der verfügbaren beobachteten Daten steht und fällt. Entscheidend ist dabei vor allem, ob tatsächlich alle für die Teilnahmeentscheidung und die Arbeitsmarktergebnisse relevanten Größen hinreichend erfasst worden sind. Allzu groß ist deshalb bei diesem Verfahren oft die Versuchung, nicht theoriegeleitet, sondern „empirizistisch" vorzugehen, und die „Zwillinge" eben aufgrund des vorliegenden anstelle des tatsächlich benötigten Datenmaterials zu bilden.

Exklusiv im Experiment: Implementationsaspekte und Prozessevaluation

Trotz aller Stärken, die ökonometrische Verfahren aufweisen, vor allem, wenn es um eindeutig messbare Größen wie Verbleib oder Einkommenssituation geht, stellen sie kein Allheilmittel der Wirkungsforschung dar. Die verbleibenden Informationslücken können jedoch gerade durch Modellversuche und deren wissenschaftliche Praxisbegleitung geschlossen werden. Erst der Praxistest und damit einhergehend die Prozessevaluation geben nämlich Aufschluss über erfolgs- und damit steuerungsrelevante Voraussetzungen, wie etwa hinsichtlich legislativer wie auch administrativer Rahmenbedingungen, der Anreizstrukturen oder der erforderlichen Personal- und Ressourcenkapazitäten für eine erfolgreiche Programmimplementation.

Zugleich können damit auch Sachverhalte erforscht werden, die der rein statistischen Analyse verborgen bleiben: So hängt die Einschätzung des Programmerfolges sicherlich

[14] Die externe Validität bezeichnet das Ausmaß der Generalisierbarkeit der ermittelten Befunde, d. h. inwieweit die Schlussfolgerungen aus dem Experiment auf andere Stichproben, situative oder örtliche Gegebenheiten usw. übertragbar sind.

auch in erheblichem Ausmaß von der Zielgruppenerreichung ab. Und *auf den zweiten Blick* kann die Einbeziehung spezifischer Adressaten die Kosten-Nutzen-Effizienz einer Maßnahme deutlich relativieren, insbesondere im Vergleich zu alternativen Mittelverwendungen. Entsprechend ist der Programmnutzen eben nicht nur am reinen Beschäftigungseffekt zu messen, sondern auch an der realisierten Allokation und Distribution von Ressourcen und Arbeitsmarktchancen. Umgekehrt steht und fällt der Programmerfolg nicht nur mit der Konzeption und Ressourcenausstattung, sondern auch mit der Akzeptanz der Maßnahme seitens der verschiedenen Arbeitsmarktakteure. D. h. um am Arbeitsmarkt Erfolge verbuchen zu können, muss ein Programm auch hinreichend den Motivationen, Interessen und Fähigkeiten der Betroffenen wie auch weiterer Arbeitsmarktakteure - und zwar sowohl auf der Angebots- als auch auf der Nachfrageseite des Arbeitsmarktes - entsprechen. Schon diese bei weitem unvollständige Auflistung weiterer Evaluationsaspekte verdeutlicht die Unverzichtbarkeit von Modellversuchen in der Arbeitsmarktpolitik.

Ein Beispiel: Der Modellversuch „Profiling und Case Management“

Ein kurzer Exkurs in die Forschungspraxis soll die Anwendungs- und Analysemöglichkeiten von Modellprojekten, zugleich aber auch deren Problemlagen und Grenzen illustrieren. Ziel eines IAB-Projektes[15] war es, zunächst über ein vorgeschaltetes „Profiling“ Arbeitslose verschiedenen Gruppen mit unterschiedlichen Arbeitslosigkeitsrisiken zuzuweisen. Darauf aufbauend sollte ein Weg gefunden werden, den Betroffenen möglichst bedarfsgerecht zu helfen. Deshalb wurde darüber hinaus in einer zweiten Projektkomponente die Wirkung eines so genannten „Case Managements“ bei Personen mit erhöhtem Risiko untersucht.

Dem auf drei Modell-Arbeitsämter und den Zeitraum vom 01.12.1999 bis 30.06.2001 begrenzten Projekt standen statistische Merkmale zu gesuchtem Beruf, Ausbildungsabschluss, Geschlecht, Alter und Gesundheit aus coArb zur Verfügung. Zusätzlich nahmen die Vermittler Einstufungen vor in Bezug auf Qualifikation, Mobilitätsbereitschaft, Motivation und Verhalten. Außerdem beurteilten sie das Risiko von Langzeitarbeitslosigkeit aufgrund ihrer persönlichen Berufserfahrung.

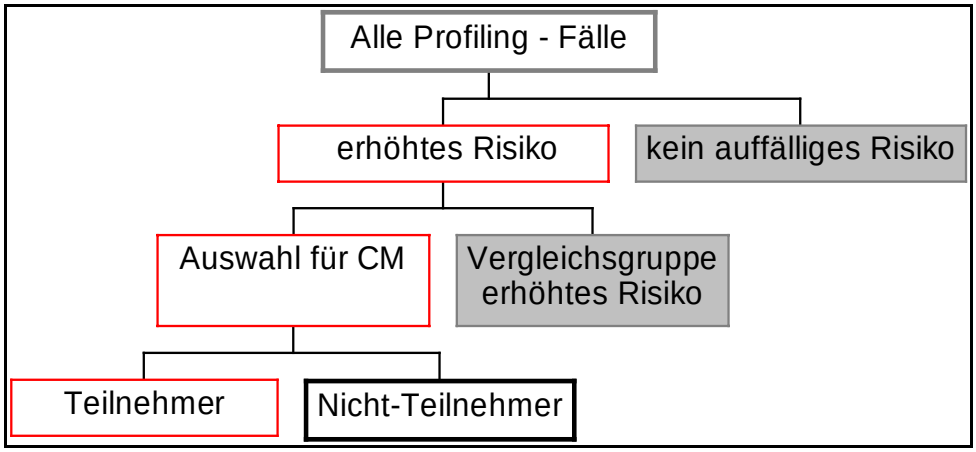
Aus der durchschnittlichen Arbeitslosigkeitsdauer für Personengruppen mit vergleichbaren statistischen Merkmalen des jeweiligen Amtes wurde zunächst ein statistischer Punktwert für das LZA-Risiko gebildet, der für alle Personen eines Amtes mit identi-

[15] IAB-Projekt 6/10-530.

schen Merkmalen gleich ist. Ab einem gewissen Punktwert wird erhöhte Gefahr von Langzeitarbeitslosigkeit vermutet.

Diese Gruppe wurde nun durch Zufallsauswahl (Randomisierung) aufgeteilt. Während die eine Hälfte der Arbeitslosen mit erhöhtem Risiko ein Angebot zur Teilnahme am Case-Management eines freien Trägers[16] erhielt (Experimentalgruppe), wurde die andere Hälfte (Vergleichsgruppe) von der Arbeitsvermittlung nach dem gewohnten Schema behandelt.

Abbildung: Kontrollgruppendesign



Die Gruppe ohne auffälliges Risiko und die Kontrollgruppe erhielten das übliche Vermittlungs- und Beratungsangebot und, wenn nötig, die Unterstützung durch arbeitsmarktpolitische Instrumente. Die für das Case-Management „ausgelosten" Personen wurden zu einer Informationsveranstaltung eingeladen und zur Teilnahme am Case-Management motiviert. Damit erfolgte die Teilnahme letztendlich auf freiwilliger Basis, sodass die oben geschilderten Selektivitätsverzerrungen in diesem Teilbereich der Untersuchungen zum Case-Management zumindest nicht ausgeschlossen werden können.

Das Untersuchungsdesign sah vor, ein Jahr nach Arbeitslosmeldung zu vergleichen, in welchem Umfang Langzeitarbeitslosigkeit eingetreten war für die Teilnehmer des Case-Managements, die Vergleichsgruppe und die übrigen Arbeitslosen, bei denen kein erhöhtes Risiko diagnostiziert worden war. Dadurch konnte sowohl die Verlässlichkeit der Risikoprognose als auch die Wirksamkeit des Case-Management-Ansatzes überprüft werden.

---

[16] Beim Träger wurden in Einzelbetreuung Stärken und Schwächen der Bewerber durchgesprochen, Unterstützung bei Bewerbungen angeboten und Lösungen für außer-qualifikatorische Probleme bei der Arbeitssuche erarbeitet (z. B. Schuldenberatung, Kinderbetreuung). Die Case-Manager sollten auch Stellen ausfindig machen, SIS-Hilfestellung geben oder den Bewerber für Maßnahmen der Vermittlung vorschlagen.

Lehren aus der Praxis

Trotz des ambitionierten Untersuchungsdesigns und seines Beitrages zum Erkenntnisgewinn bleiben einige Fragen zumindest teilweise unbeantwortet. So ist zunächst festzuhalten, dass ohne weitere Annahmen die Ermittlung der Verteilung des Maßnahmeerfolges in der Gruppe der Teilnehmer auch mittels experimenteller Daten nicht möglich ist (Björklund/Regnér 1996: 94). Denn genau so wie auch bei nicht-experimentellen Daten kann eine Person eben nicht gleichzeitig der Gruppe der Teilnehmer und der Gruppe der Nichtteilnehmer angehören. Die Studie verdeutlicht auch, dass der Gruppenzugang hinsichtlich des Case-Managements - nicht zuletzt aufgrund der geschilderten ethischen Bedenken - der Selbstselektion der Teilnehmer unterlag. Auch wenn dieser Tatbestand grundsätzlich bekannt ist, so kann seine Auswirkung auf die Befunde der Untersuchung nicht ohne weiteres statistisch kontrolliert werden.

Zu bedenken ist auch, dass selbst kontrollierte Experimente verschiedenen, schwer kontrollierbaren Einflüssen unterliegen, wie etwa durch die Programmkonzeption bedingten „Designeffekten". D. h. von der Programmkonzeption oder der Teilnahme am Modellversuch könnten Wirkungen ausgehen, wie etwa eine allgemeine Aktivierung der Teilnehmer oder deren Sensibilisierung für bestimmte Tatbestände, die nicht ohne weiteres messbar sind. Auch kann die Generalisierbarkeit der Befunde nicht grundsätzlich angenommen werden, weil risikoaverse Personen die Teilnahme an Experimenten möglicherweise scheuen können. Umgekehrt könnten Personen ihre Gruppenzugehörigkeit antizipieren und ihr Verhalten entsprechend anpassen. Im konkreten Projektbeispiel konnten Verzerrungen aber auch auftreten, weil Personen nicht auf eine Aufforderung zur Programmteilnahme reagierten oder nach anfänglicher Bereitschaft eine weitere Teilnahme ablehnten. Weitere Einschränkungen erfährt die Aussagekraft, wenn Teilnehmer, egal aus welcher Gruppe, während der Laufzeit - im ungünstigsten Fall aus unbekannten Gründen - ausscheiden oder gleichzeitig an weiteren Maßnahmen teilnehmen.

Ausblick

In Deutschland werden Modellversuche im Rahmen der Arbeitsmarktpolitik und soziale Experimente ganz allgemein erst seit kurzem in nennenswertem Umfang durchgeführt. Oft sind für aussagefähige retrospektive Studien kaum hinreichend geeignete Datengrundlagen vorhanden, auch wenn dieses Defizit in absehbarer Zeit behoben, oder doch zumindest gelindert sein dürfte. Gleichwohl sind Modellversuche in Deutschland noch immer ein relativ junger Ansatz und entsprechend befindet sich die politische Diskussion um deren Konzeption und Interpretation wohl eher noch in einem Anfangsstadium. Dabei muss zwischen experimentellen und ökonometrischen Ansätzen nicht notwendigerweise eine Kluft oder ein Konkurrenzverhältnis bestehen, sondern vielmehr sollten die beiden

Herangehensweisen als komplementär betrachtet werden, mit jeweils originären Vorzügen und Nachteilen. Auch wenn nicht-experimentelle Forschungsdesigns auf Grundlage gegenwärtig im Aufbau befindlicher Datenbasen zu Maßnahmeteilnehmern vielversprechend erscheinen, haben Modellversuche einen unverzichtbaren Stellenwert in der Forschungslandschaft. Denn Implementationsaspekte, Friktionen, Nebenwirkungen und Wirkungsmechanismen können oft erst in der praktischen Betrachtung richtig verstanden werden. Die Wirkungsforschung wird sich deshalb vermutlich in Zukunft noch intensiver mit experimentellen Ansätzen beschäftigen müssen. Für die Evaluierbarkeit dieser Modellversuche wird es dabei entscheidend sein, schon bei der Entwicklung der Programme den Belangen der Forschung Rechnung zu tragen.

Literatur

Bamberg, S./Gumbl, H./Schmidt, P. (2000): Rational Choice und theoriegeleitete Evaluationsforschung. Opladen: Leske + Budrich.

Björklund, A./Regnér, H. (1996): Experimental Evaluation of European Labour Market Policy. In: Schmid, Günther/O'Reilly, Jacqueline/Schömann, Klaus (Hrsg.): International Handbook of Labour Market Policy and Evaluation. Cheltenham, UK/Brookfield, US: Edward Elgar. 89-114.

Bortz, J. (1984): Lehrbuch der empirischen Forschung für Sozialwissenschaftler. Berlin/ Heidelberg/New York/Tokyo: Springer-Verlag.

Burtless, G. (1995): The Case for Randomized Field Trials in Economic and Policy Research. Journal of Economic Perspectives, 9/1995. 64-84.

Heckman, J./Robb, R. (1985): Alternative Methods for Evaluating the Impact of Interventions, in: Heckman, J. & Burton Singer (eds.): Longitudinal Analysis of Labour Market Data, ESM 10, New York, 1985) S. 637.

Rossi P.H./Freeman, H.E./Hofmann, G. (1999): Programm-Evaluation. Einführung in die Methoden angewandter Sozialforschung, Stuttgart, Enke-Verlag.

Schmidt, C. (2000): Ein Nobelpreis für die Mikroökonometrie. Anmerkungen zum Nobelpreis für James J. Heckman und Daniel L. McFadden. Wirtschaftsdienst 2000/XI. 690-696.

Schömann, K. (1996): Longitudinal Designs in Evaluation Studies. In: Schmid, G./O'Reilly, J./Schömann, K. (Hg.): International Handbook of Labour Market Policy and Evaluation. Cheltenham, UK/Brookfield, US: Edward Elgar. 115-142.

Ökonomie und Gesellschaft (1996): Experimente in der Ökonomie", Frankfurt a.M., Campus-Verlag. 13. Jahrbuch.