Durch Zufall zum Erkenntnisgewinn
Emu trifft Pinguin

Quantitative Evaluationsstudien zielen meist darauf ab, die Wirkung arbeitsmarktpolitischer Maßnahmen auf die Arbeitsmarktchancen der Geförderten zu bestimmen. Dazu vergleichen sie in der Regel die Teilnehmer einer Maßnahme mit ähnlichen, aber nicht mit der Maßnahme geförderten Personen. Hierbei kann es sinnvoll sein, sich den Zufall zunutze zu machen: Mit einer Zuweisung in eine Teilnehmer- und eine Kontrollgruppe nach dem Zufallsprinzip lassen sich Verzerrungen vermeiden, die aus den Entscheidungen der Arbeitslosen, der Vermittlungsfachkräfte oder der beauftragten Träger resultieren können.

Mit den sogenannten Hartz-Reformen hat sich in Deutschland eine neue Kultur der Evaluation aktiver Arbeitsmarktpolitik entwickelt. Politik und Arbeitsverwaltung hinterfragen inzwischen regelmäßig, ob der Einsatz von Maßnahmen der aktiven Arbeitsmarktpolitik die gewünschte Wirkung erzielt. Dies gilt zum Beispiel für die Förderung beruflicher Weiterbildung, für Lohnkostenzuschüsse oder „Ein-Euro-Jobs". Quantitativ ausgerichtete Forschungsarbeiten haben in der Regel zum Ziel, den kausalen Effekt einer Förderung auf ausgewählte Ergebnisvariablen – insbesondere die Beschäftigungschancen – zu bestimmen. Der Evaluationsbedarf beschränkt sich dabei nicht auf die Analyse und Beurteilung gesetzlich etablierter Instrumente der aktiven Arbeitsmarktpolitik. Er umfasst auch innovative Ansätze der aktiven Arbeitsförderung und Veränderungen im Vermittlungs- und Beratungsprozess.

Joshua Angrist und Jörn-Steffen Pischke verdeutlichen die für Wirkungsanalysen zentrale Messproblematik in ihrem aktuellen Ökonometrie-Lehrbuch anhand einer einfachen Frage: Können Krankenhäuser die Gesundheit der Patienten verbessern? Einerseits erhalten Menschen dort notwendige medizinische Versorgung, andererseits sind sie der Gefahr ausgesetzt, sich bei anderen Patienten mit weiteren Krankheiten anzustecken. Befragt man nun Personen nach einem Krankenhausaufenthalt zu ihrem Gesundheitszustand und vergleicht diesen mit dem einer Gruppe von Personen, die nicht stationär behandelt wur-

den, lässt sich das Ergebnis erahnen: Menschen, die nicht im Krankenhaus waren, sind im Mittel gesünder. Ebenso einfach lässt sich dies erklären: Menschen ohne Krankenhausaufenthalt waren im Durchschnitt von Anfang an gesünder. Um die kausale Wirkung des Krankenhausaufenthalts zu ermitteln, müsste also bekannt sein, wie gesund die ehemaligen Patienten ohne die Behandlung gewesen wären.

Vor einer ähnlichen Problematik stehen Arbeitsmarktforscher, wenn sie die kausale Wirkung der Teilnahme an einer Maßnahme auf Ergebnisvariablen wie die Beschäftigungschancen ermitteln wollen. Denn die Förderung könnte zum Beispiel von vornherein auf Personen mit besonders geringen Beschäftigungschancen ausgerichtet sein. Das Ergebnis der Teilnehmer nach einer Förderung muss daher mit dem Ergebnis verglichen werden, das sich bei ihnen ohne diese Maßnahme eingestellt hätte. Letzteres ist aber nicht bekannt, sondern muss mit Hilfe einer Vergleichsgruppe geschätzt werden. Diese setzt sich aus Personen zusammen, die den Teilnehmern ähneln, aber nicht an der interessierenden Maßnahme teilgenommen haben.

Die Wahl der Vergleichsgruppe ist kein triviales Problem: Bei der Abgrenzung zwischen den Teilnehmern und den passenden Vergleichspersonen darf die Teilnahmeentscheidung nicht bereits durch das erwartete Ergebnis beeinflusst worden sein – also der Krankenhausaufenthalt nicht durch den erwarteten Gesundheitszustand bzw. die Teilnahme an einer Maßnahme nicht durch die erwarteten Arbeitsmarktchancen. Andernfalls liegt eine sogenannte Selektionsverzerrung vor.

Die experimentelle Idee

Die Evaluation arbeitsmarktpolitischer Maßnahmen erfolgt häufig mit Hilfe statistischer Matching-Verfahren. Dabei werden aus großen Individualdatensätzen nachträglich Vergleichspersonen ermittelt, die den Teilnehmern in wichtigen Merkmalen ähnlich sind – sogenannte „statistische Zwillinge". Hierdurch lässt sich das Selektionsproblem dann vermeiden, wenn alle zentralen Merkmale, die sowohl die Teilnahme an einer Maßnahme als auch das Arbeitsmarktergebnis beeinflussen, in den verwendeten Datengrundlagen verfügbar sind. Auch bei einer guten Datengrundlage kann dies zwar plausibel begründet, aber nicht bewiesen werden. Für die Teilnahme und das Ergebnis können allerdings auch Merkmale, die nicht beobachtet werden können – wie zum Beispiel die Motivation – eine Rolle spielen.

Experimente gehen hingegen grundsätzlich anders an das Selektionsproblem heran und versuchen, es vorab durch eine zufällige Zuweisung in die Förderung zu lösen. Wiederum bietet sich eine Parallele zur Medizin an: Bei der Erprobung eines neues Medikamentes erhält häufig eine Teilnehmergruppe das neue Mittel, während eine Kontrollgruppe ein alternatives Medikament oder ein Placebo erhält. Die Zuweisung zur Teilnehmer- und Kontrollgruppe erfolgt dabei mit Hilfe eines Zufallsmechanismus, um vergleichbare Gruppen zu erhalten.

Vor der Einführung neuer oder neu ausgestalteter arbeitsmarktpolitischer Instrumente bietet es sich an, diese im Rahmen von Modellversuchen mit zufallsgesteuerter Personenzuweisung zu erproben. In diesen Projekten nehmen Arbeitslose mit einer vorab festgelegten Wahrscheinlichkeit – unter ansonsten realen Bedingungen – an einer bestimmten Fördermaßnahme teil. Durch die zufällige Zuweisung lassen sich von vornherein potenzielle Selektionsprobleme vermeiden, die aus den Entscheidungen der Arbeitslosen, der Vermittlungsfachkräfte oder der beauftragten Träger resultieren könnten.

Ist die Zufallszuweisung ethisch vertretbar?

Hier stellt sich unmittelbar die Frage nach der ethischen Vertretbarkeit eines solchen Vorgehens. Grenzen des experimentellen Ansatzes müssen grundsätzlich überall dort beachtet werden, wo der stark gesetzlich regulierte Rahmen der bundesdeutschen Arbeitsförderung Pflichtleistungen definiert: Natürlich dürfen Personen ihnen zustehende Leistungen nicht verweigert werden. Ethische Einwände beziehen sich häufig darauf, dass vergleichbare Personen unterschiedlich behandelt werden und eine Gruppe besser als die andere gestellt werden könnte – und das nicht als Ergebnis einer Ermessensentscheidung, sondern auf Zufallsbasis.

Andererseits lässt sich der Einsatz einer Maßnahme mit unbekannter, eventuell sogar negativer Wirkung eben-

so aus ethischen Gründen anzweifeln. Sinnvoll und zu rechtfertigen ist die zufällige Zuweisung nur dann, wenn vorab nicht bekannt ist, ob eine Maßnahme tatsächlich die Arbeitsmarktchancen der Geförderten verbessert oder sie vielleicht sogar verschlechtert. Dann können Modellversuche mit Zufallszuweisung, an denen eine begrenzte Anzahl von Personen teilnimmt, Entscheidungshilfen für die Einführung neuer Maßnahmen, die Fortführung erfolgreicher Maßnahmen oder die Einstellung erfolgloser Maßnahmen liefern.

Modellversuche in der Praxis

In den USA sind entsprechende Untersuchungsdesigns bereits seit langem Standard, wie Burt Barnow in einem aktuellen Artikel in der Zeitschrift für ArbeitsmarktForschung (ZAF) ausführt. So gab es in den 1960er und 1970er Jahren eine Reihe von Modellversuchen zur sogenannten „Negativen Einkommenssteuer", einem Kombilohn-Ansatz. Im Bereich der aktiven Arbeitsmarktpolitik nutzte der Job Training Partnership Act (JTPA) klassische experimentelle Designs, um zu prüfen, welche Maßnahmen zur Integration von schwer vermittelbaren Personen in den Arbeitsmarkt beitragen.

Auch die Bundesagentur für Arbeit (BA) kann entsprechende Modellversuche selbst initiieren und steuern, um wichtige Erkenntnisse für den Einsatz arbeitsmarktpolitischer Instrumente zu gewinnen. Ein wichtiges Thema ist zum Beispiel, ob die BA und private Unternehmen Vermittlungsdienstleistungen ähnlich effektiv und effizient erbringen. Zentrales Ziel der Einbindung privater Vermittlungsdienstleister ist es, den Markt für Vermittlungsdienstleistungen wettbewerblicher zu organisieren. Die Nachfrage nach privaten Vermittlungsdienstleistungen erfolgt dabei entweder über Gutscheine oder durch Ausschreibungen. Im zweiten Fall entscheiden die Vermittlungsfachkräfte bzw. Fallmanager darüber, welche Arbeitsuchenden an die privaten Vermittlungsdienstleister überwiesen werden.

Welches Problem kann bei der nachträglichen Bildung einer Vergleichsgruppe durch statistisch-ökonometrische Standard-Matching-Verfahren auftreten? Vermittler könnten den privaten Vermittlungsdienstleistern Personen zuweisen, die sich – für den Forscher nicht beobachtbar – systematisch von den verbleibenden Personen unterscheiden. Dies können zum Beispiel besonders arbeitsmarktferne Personen sein, oder aber Personen, die aufgrund vergleichsweise positiver Beschäftigungschancen nach Meinung des Vermittlers keine Unterstützung durch die Arbeitsagentur benötigen. Die BA hat dieses Problem einer möglicherweise selektiven Vermittlerzuweisung in zwei Projekten aufgegriffen.

Emu trifft Pinguin – Elektronischer Münzwurf sorgt für Vergleichbarkeit

Nachdem der Gesetzgeber im Jahr 2003 die „Beauftragung von Trägern mit Eingliederungsmaßnahmen" eingeführt hatte (§ 421i des Sozialgesetzbuchs III), nutzte die BA diese bis zum Jahr 2008 geltende Regelung, um die bekannten Probleme bei der Vergleichsgruppenbildung aufzulösen. Die konkreten Inhalte und die Art der Durchführung der Maßnahmen wurden nicht durch die Arbeitsagenturen vorgegeben; in einem wettbewerblichen Verfahren zwischen Trägern sollte das beste Integrationskonzept für eine bestimmte Zielgruppe identifiziert werden. Die bisherigen Erfahrungen mit der Vorgabe von Leistungszielen sowie deren Regulierung und Kontrolle durch die Arbeitsverwaltung reichten für Empfehlungen an die Politik jedoch nicht aus. Daher startete im Jahr 2007 das Modellprojekt „Private Arbeitsmarktdienstleister wirksamer einbinden", an dem insgesamt etwa 9.500 Arbeitslosengeld-Empfänger beteiligt waren bzw. sind.

Gegenüber der Regelbetreuung wurden bei den Projektteilnehmern verschiedene Elemente variiert. Dies waren unter anderem eine individuell auf die Person zugeschnittene Zuweisungsdauer (statt einheitlicher Zuweisungsdauern) und ein persönlicher Ansprechpartner der Agenturen beim Träger, um Anliegen direkt vor Ort zu klären. Aus den Erfahrungen vergangener Modellversuche war das Problem der selektiven Vermittlerzuweisung bekannt. Aus diesem Grunde wurde ein sogenannter „Elektronischer Münzwurf" entwickelt. Diese Zufallszuweisung ist nicht manipulierbar. Damit ist die Vergleichbarkeit von

Teilnehmer- und Kontrollgruppe gewährleistet – was Auswertungen der beobachtbaren Merkmale beider Gruppen bestätigen. Erste quantitative Befunde aus diesem Projekt wird die Zentrale der Bundesagentur für Arbeit im Jahr 2011 vorlegen; die qualitative Begleitforschung erfolgt durch das IAB.

Im Projekt „Interne ganzheitliche Unterstützung zur Integration im SGB III", kurz „Pinguin" genannt, werden seit Mitte 2009 in drei Arbeitsagenturen Arbeitslose mit besonderen Vermittlungshemmnissen zufällig entweder an einen privaten Arbeitsmarktdienstleister oder aber an ein internes Projektteam der Agentur zugewiesen (vgl. Abbildung). Sowohl der private Dienstleister als auch das interne Projektteam erbringen ganzheitliche Beratungs- und Vermittlungsleistungen. Die Zufallszuweisung ist nicht manipulierbar und erfolgt ebenfalls durch den „Elektronischen Muenzwurf" (EMu). Das IAB evaluiert derzeit in Kooperation mit der Zentrale der BA die Effektivität dieses Modellprojektes; erste Ergebnisse werden voraussichtlich im Jahr 2011 publiziert. Begleitend erfolgt eine qualitative Evaluation durch das Soziologische Forschungsinstitut an der Universität Göttingen.

Solche Modellprojekte und ihre Evaluation erfordern eine gute Abstimmung zwischen der Projektleitung, den beteiligten Arbeitsagenturen, dem Projektmonitoring und den beteiligten Forschern. Zentrale Aufgabe des Monitorings ist die laufende Projektbeobachtung (häufig anhand von Kennziffern), um festzustellen, ob das Projekt wie geplant läuft oder gegebenenfalls Anpassungsbedarf besteht. Erkenntnisgegenstand quantitativer Evaluationsstudien sind hingegen die „Netto"-Wirkungen des Projektes über einen längeren Zeitraum, während qualitiative Studien hinterfragen, warum und wie ein Projekt wirkt. Das IAB als Forschungsinstitut kann durch seine Arbeiten die kurzfristig ausgerichteten Monitoring-Aktivitäten durch solche längerfristig angelegten Evaluationen ergänzen. Die Evaluationsergebnisse kommen einerseits dem Informationsbedarf von Politik und Arbeitsverwaltung entgegen, tragen andererseits aber auch zur wissenschaftlichen Diskussion bei.

Durch die Zufallszuweisung Erkenntnisse gewinnen

Im Idealfall lassen sich bei einem entsprechenden Modellversuch Vermittlungs- und Förderprozesse genau definieren und beobachten, was bei einer nachträglichen Analyse mit statistischen Matching-Verfahren im Regelfall nicht gewährleistet ist. Bei einem (aus wissenschaftlicher Sicht) gelungenen Modellversuch können unterschiedliche Ergebnisse von Teilnehmer- und Kontrollgruppe kau-

Abbildung
Der Zuweisungsprozess im Modellprojekt „Pinguin"

sal allein auf eine bestimmte Förderung zurückgeführt werden – man spricht dann von „interner Validität". Hier trägt also der Zufall maßgeblich zum Erkenntnisgewinn bei und fördert damit die weitere Umsetzung erfolgreicher Konzepte.

Was kann die Ergebnisse verzerren?

Ein Modellversuch, bei dem die Personenzuweisung zufallsgesteuert wird, kann jedoch auch seine Tücken haben. Im Jahr 1995 haben der spätere Nobelpreisträger James Heckman und Jeffrey Smith wichtige Aspekte in einem Überblicksartikel zusammengefasst. So kann sich die Teilnehmergruppe aufgrund des Modellversuchs anders zusammensetzen als es bei einem flächendeckenden Programm der Fall gewesen wäre (Randomisierungsverzerrung). Bei dem bereits angesprochenen Job Training Partnership Act in den USA mussten etwa die Kriterien für eine Förderung teils gelockert werden, um hinreichend viele Teilnehmer zu gewinnen. Weiterhin könnte die Kontrollgruppe auf alternative Leistungen ausweichen, die ohne Modellprojekt nicht verfügbar wären (Substitutionsverzerrung). Auch dies war im Job Training Partnership Act zu beobachten.

Zudem sollte das Untersuchungsdesign eines Projekts so ausgestaltet sein, dass sich mögliche Wirkungsfaktoren isolieren lassen. So ist es zum Beispiel wenig sinnvoll, die Teilnehmergruppe in einem Modellversuch gleichzeitig stärker zu betreuen, ihnen Gutscheine für Kinderbetreuungsleistungen auszuhändigen, sie einer neuen Variante von Aktivierungsmaßnahmen zuzuweisen und den Vermittlern im Erfolgsfall zusätzliche Leistungsprämien auszuzahlen. Gerade bei Modellversuchen können schließlich Lerneffekte auftreten, so dass möglicherweise die Startphase und Folgezeiträume separat ausgewertet werden müssen.

Lassen sich die Ergebnisse verallgemeinern?

Die Situation, die untersucht wird, sollte sich möglichst verallgemeinern lassen, sonst fehlt es an „externer Validität". Dies wäre etwa dann der Fall, wenn die Beteiligung auf freiwilliger Basis erfolgt. So könnten sich nur Arbeitsagenturen mit besonders motivierten oder interessierten Führungskräften und Mitarbeitern an einem Projekt beteiligen, deren Ergebnisse sich nicht ohne weiteres auf andere Agenturen übertragen lassen. Beim Job Training Partnership Act in den USA lehnten zum Beispiel 90 Prozent der Weiterbildungscenter eine Teilnahme ab. Darüber hinaus kann die Wirksamkeit arbeitsmarktpolitischer Maßnahmen mit den ökonomischen Rahmenbedingungen variieren. Dies spricht dafür, entsprechende Modellversuche regional möglichst breit zu streuen.

Auch der sogenannte „Hawthorne-Effekt" kann die Verallgemeinerbarkeit einschränken, wenn ein Modellversuch Verhaltensänderungen bewirkt. Die Bezeichnung geht auf die Hawthorne-Werke der Western Electric Company zurück, in denen Forscher in den Jahren 1924 bis 1932 die Arbeitsproduktivität untersuchten. Hierzu variierten sie systematisch die Arbeitsbedingungen (zum Beispiel die Beleuchtung). Die Veränderungen wurden mit der betroffenen Arbeitsgruppe täglich diskutiert. Im Ergebnis stieg die Produktivität bei jeder Veränderung – egal, in welche Richtung sie erfolgte. Die Forscher erklärten dies unter anderem durch die gestiegene Aufmerksamkeit, die die Teilnehmer erhielten. Im Kontext von Modellversuchen im Bereich der Arbeitsmarktpolitik könnten beteiligte Arbeitsagenturen auf Monitoring-Aktivitäten, die mit dem Projekt verbunden sind, reagieren. Dann wären die gemessenen Wirkungen zum Teil nicht der untersuchten Fördermaßnahme, sondern der besonderen Beobachtung zuzuschreiben – ein Umstand, der bei der nachträglichen Bildung von Teilnehmer- und Kontrollgruppe über statistische Matching-Verfahren nicht auftritt.

Gänzlich an ihre Grenzen stoßen schließlich auch Modellversuche mit zufallsgesteuerter Personenzuweisung, wenn es um indirekte Effekte von Fördermaßnahmen auf andere Personen geht. So könnten etwa Arbeitslose, für die öffentlich geförderte Beschäftigungsverhältnisse geschaffen werden, Beschäftigte in anderen Bereichen der Wirtschaft verdrängen. Dies wird bei einem einfachen Vergleich der Arbeitsmarktergebnisse von Teilnehmern und Kontrollgruppe ausgeblendet.

Fazit

Vielen Evaluationsstudien des IAB und auch der BA-internen Wirkungsanalyse TrEffeR (Treatment Effects and PRediction) liegt ein Vergleich der Ergebnisse von Maßnahmeteilnehmern und nachträglich gebildeten Gruppen von „statistischen Zwillingen" zugrunde. Diese Methodik schont insofern den Geldbeutel der Beitrags- und Steuerzahler, als sie die umfangreichen administrativen Datenbestände der BA nutzt. Dadurch ist außerdem die externe Validität, also die Verallgemeinerbarkeit der Ergebnisse gesichert. Darüber hinaus treten keine „Hawthorne-Effekte" auf. Die interne Validität der Befunde ist jedoch nicht abschließend nachweisbar.

Gezielt eingesetzte Modellversuche mit einer zufallsgesteuerten Personenzuweisung können ergänzend wichtige Informationen zur Wirksamkeit arbeitsmarktpolitischer Maßnahmen bereitstellen: Verzerrungen, die die interne Validität der Befunde gefährden, lassen sich bei ihnen weitestgehend ausschließen. Sie könnten und sollten daher deutlich häufiger als bisher genutzt werden, um innovative Arbeitsmarktinstrumente vor einer Flächeneinführung zu erproben. Im Vergleich zu dem erwarteten Nutzen dürften die hierfür entstehenden Kosten in vielen Fällen vertretbar und vergleichsweise gering ausfallen. Damit die externe und interne Validität gesichert ist, muss das Design jedoch vorab sorgfältig zwischen den Entscheidungsträgern in Politik und Verwaltung sowie den an der Umsetzung und der Evaluation Beteiligten abgestimmt werden.

Literatur

Angrist, Joshua D.; Pischke, Jörn-Steffen (2008): Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Barnow, Burt (2010): Setting Up Social Experiments: The Good, the Bad, and the Ugly. Zeitschrift für ArbeitsmarktForschung Vo. 43, Nr. 2 (im Erscheinen). Online-Fassung unter http://www.springer.com/economics/labor/journal/12651.

Heckman, James J.; Smith, Jeffrey A. (1995): Assessing the Case for Social Experiments. Journal of Economic Perspectives 9, S. 85-110.

Müntnich, Michael; Wießner, Frank (2002): Soziale Experimente und Modellversuche: Ein Beitrag zur Evaluation von Neuansätzen in der Arbeitsmarktpolitik. In: Kleinhenz, Gerhard (Hg.): IAB-Kompendium Arbeitsmarkt- und Berufsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung 250, S. 415-427.

Die Autoren

Michael Müntnich
ist fachlicher Leiter „Produkt- und Programmanalyse" im Geschäftsbereich Strategie / Weiterentwicklung / Arbeitsmarkt der Bundesagentur für Arbeit.
michael.muentnich@arbeitsagentur.de

Torben Schewe
ist Leiter „Produkt- und Programmanalyse" im Geschäftsbereich Strategie / Weiterentwicklung / Arbeitsmarkt der Bundesagentur für Arbeit.
torben.schewe@arbeitsagentur.de

Prof. Dr. Gesine Stephan
ist Leiterin des Forschungsbereichs „Arbeitsförderung und Erwerbstätigkeit" am IAB.
gesine.stephan@iab.de