STANDPUNKTE

IZA Standpunkte Nr. 68

Evidenzbasierte Wirtschaftspolitik in Deutschland: Defizite und Potenziale

Bernhard Boockmann
Claudia M. Buch
Monika Schnitzer

April 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Evidenzbasierte Wirtschaftspolitik in Deutschland: Defizite und Potenziale

Bernhard Boockmann
IAW, Universität Tübingen und IZA

Claudia M. Buch
Universität Magdeburg, IWH und CESifo

Monika Schnitzer
Universität München und CEPR

IZA Standpunkte Nr. 68
April 2014

IZA

Postfach 7240
53072 Bonn

Tel.: (0228) 3894-0
Fax: (0228) 3894-180
E-Mail: iza@iza.org

IZA Standpunkte Nr. 68
April 2014

ZUSAMMENFASSUNG

Evidenzbasierte Wirtschaftspolitik in Deutschland: Defizite und Potenziale[1]

Moderne Evaluationsmethoden auf der Basis ökonometrischer Verfahren und randomisierter Feldexperimente machen es für immer mehr Politikbereiche möglich, die Wirksamkeit wirtschaftspolitischer Maßnahmen zu überprüfen. Gleichwohl sind diese Methoden in der deutschen Evaluationspraxis nicht der Standard. Andere Länder sind Deutschland in dieser Hinsicht voraus. Gerade vor dem Hintergrund knapper öffentlicher Mittel ist eine Verbesserung der Evaluationspraxis dringend geboten, um die Mittelverwendung auf solche Maßnahmen fokussieren zu können, deren Wirksamkeit nachgewiesen ist. In diesem Beitrag werden institutionelle Voraussetzungen für methodisch valide Wirkungsanalysen diskutiert und mögliche Schritte hin zu einer stärker evidenzbasierten Wirtschaftspolitik in Deutschland vorgeschlagen.

JEL-Codes: A11, D09

Schlagworte: kausale Wirkungsanalysen, experimentelle Wirtschaftsforschung

Kontaktadresse:

Claudia Buch
Institut für Wirtschaftsforschung Halle (IWH)
Kleine Märkergasse 8
D-06108 Halle (Saale)
E-mail: claudia.buch@iwh-halle.de

[1] Dieser Beitrag wurde für die Zeitschrift „Perspektiven der Wirtschaftspolitik“ verfasst. Die Autoren danken zahlreichen Gesprächspartnern für wichtige Hintergrundgespräche zur Vorbereitung dieses Beitrags. Unser Dank gilt vor allem Markus Nagler, Gesine Stephan und Achim Wambach. Alle Ungenauigkeiten, Fehler und Einschätzungen gehen allein zu Lasten der Autoren dieses Beitrags.

> "*Some policies do seem to have many upsides and few downsides, such as allowing more skilled immigrants, strengthening the education systems, and eliminating unwise regulation. But when we move beyond such simple broad policies towards specific entrepreneurship strategies like clustering, our ignorance becomes obvious. The best path forward involves experimentation and evaluation*."
>
> Chatterji, Glaeser, Kerr „Clusters of Entrepreneurship and Innovation" (2013)

1 Hintergrund

Kausale Wirkungsanalysen auf der Basis statistisch-ökonometrischer Methoden und randomisierter Feldexperimente sind unter Wissenschaftlern weltweit zum „state of the art" geworden. Die in den letzten Jahrzehnten entwickelten Verfahren ermöglichen es, die Wirksamkeit wirtschaftspolitischer Maßnahmen zu überprüfen und somit die erforderliche Informationsgrundlage für eine evidenzbasierte Politik zu schaffen. Die Nutzung dieser Methoden wurde durch den verbesserten Zugang zu Mikrodaten in den vergangenen Jahren in einigen Bereichen, insbesondere in Bezug auf den Arbeitsmarkt, deutlich erleichtert. Die Politik, so könnte man meinen, sollte diese neuen Methoden begrüßen und über Länder- und Politikbereiche hinweg zum Standard ihrer Evaluationspraxis machen.

Die Realität sieht allerdings anders aus. Zwar gibt es zahlreiche wissenschaftliche Arbeiten[2] und allgemeinverständliche Beiträge[3], die sich mit Verfahren und Anwendungen der quantitativen Wirkungsanalyse auseinandersetzen. In Deutschland werden solche Analysen jedoch nur in wenigen Bereichen und von wenigen staatlichen Stellen regelmäßig durchgeführt bzw. in Auftrag gegeben. Auch findet hierzulande kaum eine systematische Kosten-Nutzen-Analyse auf der Basis von quantitativen Wirkungsabschätzungen statt. Und schließlich werden quantitative Evaluationen dort, wo sie durchgeführt werden, oft nicht systematisch in die politischen Entscheidungsprozesse einbezogen.

Warum Wirkungsanalysen in der Praxis meist nicht zum Einsatz kommen, hat vielfältige Gründe. In diesem Beitrag wird diskutiert, wie ein Dialog zwischen Politik, Wissenschaft und der breiten Öffentlichkeit über eine methodisch besser fundierte Wirtschaftspolitik befördert werden könnte. Anhand von Beispielen aus der Evaluationspraxis in den

---

[2] Einen Überblick über den Stand der Forschung im Bereich der Arbeitsmarktevaluation in Deutschland geben unter anderem Bernhard et al. (2009), Bräuninger et al. (2013), Bundesministerium für Arbeit und Soziales und Institut für Arbeitsmarkt- und Berufsforschung (2011), Eichhorst und Zimmermann (2007), Heyer et al. (2011) und Jacobi und Kluve (2007).

[3] Beispielhaft seien hier ein aktuelles Gutachten des Wissenschaftlichen Beirats beim BMWI (2013), als Lehrbuch Bauer et al. (2009) sowie die Arbeiten von Arni (2012), Kugler et al. (2014) und Schmidt (2009) genannt.

Vereinigten Staaten und aus dem Bereich der Arbeitsmarktpolitik in Deutschland wird im ersten Teil aufgezeigt, wie kausale Wirkungsanalysen erfolgreich eingesetzt werden können und welche Faktoren dazu beigetragen haben, dass dies in der politischen Praxis tatsächlich geschieht. Im zweiten Teil werden institutionelle Rahmenbedingungen diskutiert, die eine evidenzbasierte Wirtschaftspolitik befördern können.

Die bisherige Evaluationspraxis in Deutschland ist, ähnlich wie in den meisten anderen europäischen Ländern, vor allem auf die Ex ante-Evaluation wirtschaftspolitischer Maßnahmen ausgerichtet. Es gibt keine ausgeprägte Kultur der Evaluation von wirtschaftspolitischen Maßnahmen im Sinne einer Identifikation *kausaler* Effekte (Smith 2009), die naturgemäß erst nach Durchführung einer Maßnahme erfolgen kann. Dabei kann die genaue Wirkung von Maßnahmen vorab nicht mit Exaktheit bestimmt werden (Manski 1995). Wir leben in einer unsicheren Welt und die Erkenntnis darüber, was in der politischen Praxis funktioniert („what works“)[4] und was nicht, muss über einen Prozess des Experimentierens und Evaluierens erlangt werden. Gerade das macht die Ex post-Wirkungsanalyse unverzichtbar.

Andere Länder sind uns im Einsatz und methodischen Anspruch solcher Ex post-Analysen deutlich voraus. Ein anschauliches Beispiel dafür sind die Fördermaßnahmen, die im Zusammenhang mit der jüngsten Finanzkrise zur Stabilisierung beschlossen wurden. In den Vereinigten Staaten wurden mit Beginn der Maßnahmen im Rahmen des *American Recovery and Reinvestment Act* systematische Wirkungsanalysen geplant, die relevanten Informationen begleitend zu diesen Fördermaßnahmen erhoben und transparent auf einer Webseite dargestellt (http://www.recovery.gov/). In Deutschland sind vergleichbare Informationen über die finanzpolitischen Maßnahmen der Jahre 2008 und 2009 auf der Mikroebene hingegen nicht öffentlich; sie stehen selbst der vom Bundesfinanzministerium in Auftrag gegebenen Ex post-Evaluation nicht zur Verfügung.

Dass die Evaluationskultur in Europa im Vergleich mit den Vereinigten Staaten so schwach ausgeprägt ist, muss überraschen. Denn auf nationaler wie auch auf EU-Ebene werden in Europa große Beträge für gezielte wirtschaftspolitische Fördermaßnahmen ausgegeben. Beispielsweise wird auf europäischer Ebene aktuell eine Reihe von Maßnahmen diskutiert, die zum Ziel haben, das Wachstum zu stimulieren sowie insbesondere kleine und mittelständische Unternehmen zu fördern (EU Kommission 2013). So gibt es Bestrebungen, im Rahmen sogenannter Vertragspartnerschaften Strukturreformen über Ländergrenzen hinweg zu finanzieren – um positive externe Effekte, die von Reformen in einem Land auf

[4] Für Übersichten über Studien im Sozial- und Bildungsbereich vgl. http://www.whatworks.edu.au/ oder https://www.gov.uk/government/publications/what-works-evidence-centres-for-social-policy

andere Länder ausgehen, auszugleichen.[5] Ein solcher finanzieller Ausgleich setzt aber voraus, dass die Wirkungen wirtschaftspolitischer Maßnahmen hinreichend gut quantifizierbar sind.

Die vorhandenen Möglichkeiten zur Abschätzung der Wirksamkeit politischer Interventionen werden also bisher nicht ausgeschöpft. Noch seltener wird die Frage nach der Relation von Kosten und Nutzen gestellt; es wird zu wenig gefragt, ob eine Maßnahme fiskalisch und gesamtwirtschaftlich effizient ist. Kosten-Nutzen-Analysen scheitern in Deutschland in der Regel an Datenbeschränkungen, die dazu führen, dass die durch bestimmte Maßnahmen hervorgerufenen Effekte meist nicht mit Angaben zu Kosten oder Nutzen verglichen werden können. Aber die Verfügbarkeit von Daten alleine würde noch nicht ausreichen. Kosten und Nutzen müssten zusätzlich umfassender definiert werden. Oft unterliegt politischen Entscheidungen ein zu enger Begriff der Kosten und Nutzen wirtschaftspolitischer Maßnahmen. Im Zuge einer gesetzgeberischen Maßnahme werden üblicherweise nur Kosten, die unmittelbar durch die Verwaltung der Maßnahme entstehen, betrachtet. Die ökonomischen Kosten in Form veränderter Anreize und Auswirkungen auf Marktergebnisse werden nicht Betracht gezogen, sie können aber immens sein.

Die Notwendigkeit, „Gesetzesfolgekosten“ breiter im Sinne einer Ex Post Evaluation zu definieren, wurde erst unlängst vom Normenkontrollrat betont. Dieses Gremium hat die Aufgabe, die Bundesregierung bei der Umsetzung ihrer Maßnahmen zum Bürokratieabbau und zur besseren Rechtsetzung zu unterstützen. Es geht also um die Abschätzung *unmittelbarer* Gesetzesfolgekosten im Vorfeld einer Maßnahme (ex ante-Verfahren). Als eine zentrale, im Bundeskanzleramt angesiedelte Institution hat der Normenkontrollrat in den Behörden das Bewusstsein dafür geschärft, die unmittelbar mit einer geplanten Maßnahme entstehenden Kosten in den Planungen zu berücksichtigen. Darüber hinaus weist der Normenkontrollrat in seinem jüngsten Jahresbericht auf die Notwendigkeit einer stärker ausgeprägten Evaluationskultur in Deutschland hin: „Aus Sicht des NKR ist es jedoch ebenso wichtig, nach etwa drei bis fünf Jahren zu prüfen, wie sich Gesetze und Verordnungen in der Praxis bewährt haben – also eine Evaluierung vorzunehmen.“ (Normenkontrollrat 2013: S. 65). In diesem Sinne hat der Staatssekretärsausschuss der Bundesregierung im Januar 2013 beschlossen, alle Gesetze mit Erfüllungskosten ab einem Schwellenwert von einer Million Euro verpflichtend im Nachhinein (ex post) evaluieren so lassen (ebenda).

Die Notwendigkeit von kausalen Wirkungsanalysen wird also zum Teil durchaus erkannt und erste organisatorische Schritte tragen dieser Tatsache Rechnung, wie beispielsweise die Einrichtung eines Aufbaustabes Fördercontrolling/Evaluation im (damaligen) Bundesministerium für Wirtschaft und Technologie im Jahr 2011. Von diesen ersten Ansätzen bis zu einer ausgeprägten Evaluationskultur ist aber noch ein weiter Weg, nicht zuletzt, weil in vielen Politikbereichen angezweifelt wird, ob solche Verfahren in diesen

[5] Für eine Beschreibung dieses Verfahrens vgl. Sachverständigenrat (2013).

Bereichen überhaupt anwendbar sind. In diesem Beitrag diskutieren wir, wie auf diesem Weg Fortschritte erzielt werden können. Wir beginnen mit einer kurzen Definition, was wir unter „Evaluation“ verstehen. Danach schildern wir, wie Evaluationen in Deutschland und in den USA genutzt werden und wie sie institutionell verankert sind. Im Anschluss daran gehen wir auf typische Argumente ein, die eine unzureichende Nutzung wissenschaftlicher Methoden erklären können, und diskutieren institutionelle Ansatzpunkte, um evidenzbasierte Forschung und Politikberatung auf eine breitere Basis zu stellen.

2 Was bedeutet Evaluation?

Unter „Evaluation“ verstehen verschiedene wissenschaftliche Disziplinen und wirtschaftspolitische Anwender teilweise sehr unterschiedliche Verfahren. In unserem Beitrag geht es um quantitative, also statistische und ökonometrische Verfahren zur Abschätzung von Kausaleffekten. Qualitative Evaluationsverfahren (http://www.degeval.de/) oder Evaluationen mit Makro-Daten haben je nach Fragestellung ihre Berechtigung und können komplementär zur Kausalanalyse auf der mikroökonomischen Ebene sein, werden hier aber nicht einbezogen.[6]

Somit beschäftigen wir uns mit Verfahren, die in der Forschung unter „New economics of program evaluation“ oder „treatment evaluation“ zusammengefasst werden, also experimentelle, quasi-experimentelle und nicht-experimentelle Verfahren (Angrist und Pischke, 2010, Bauer et al. 2009, Kugler et al. 2014). Diese Forschungsrichtung stellt die empirische Strategie, mit der Wirkungen von Interventionen identifiziert wird, in den Vordergrund. Das Grundproblem der kausalen Wirkungsanalyse ist es, dass nur beobachtet werden kann, wie es den Betroffenen einer Maßnahme nach Durchführung der Maßnahme geht, nicht aber, wie es ihnen ergangen wäre, hätte es die Maßnahme nicht gegeben. Da diese kontrafaktische Situation nicht direkt beobachtbar ist, arbeitet man mit dem Konzept sogenannter Vergleichsgruppen. Die Ergebnisse in der Vergleichsgruppe stellen das kontrafaktische Ergebnis für die Gruppe der Betroffenen dar. Dies setzt voraus, dass sich beide Gruppen in allen relevanten Eigenschaften außer der Betroffenheit von der Maßnahme gleichen.

Ein einfacher Vergleich zwischen allen Betroffenen und Nichtbetroffenen gibt meistens keinen Aufschluss über die Wirksamkeit der Maßnahme, weil der sogenannte Selektionseffekt nicht berücksichtigt wird. Wenn beispielsweise nur diejenigen in die Maßnahme

---

[6] Üblicherweise wird auch ein Unterschied zwischen einer „summativen“ Evaluation im Sinne einer Wirkungskontrolle ex post und einer „formativen“ Evaluation gemacht, wobei letztere dazu genutzt wird, um im Prozessablauf eines Programms steuernd einzugreifen. In diesem Beitrag geht es ausschließlich um die Bewertung ex post.

aufgenommen werden, die bereits ungünstige Ergebnisse aufweisen, oder wenn umgekehrt nur die die Schnellsten oder die Motiviertesten die Maßnahme erhalten, dann führt ein einfacher Vergleich zwischen Teilnehmenden und Nichtteilnehmenden zu verzerrten Ergebnissen. Eine solche Verzerrung kann insbesondere dann zuverlässig ausgeschlossen werden, wenn eine Maßnahme zufällig zugeteilt wird.

Die zufällige Zuweisung (die „Randomisierung“) von Personen oder anderen Betroffenen zu einer Maßnahme (einem „Treatment“) wird oft als der „Goldstandard“ in der Evaluationsforschung bezeichnet. Dieses Verfahren eines sogenannten „randomisierten Feldexperiments“ ist mittlerweile zum Standard bei vielen Projekten in Entwicklungsländern geworden; das können Maßnahmen zur Reduzierung von Malariafällen, zur Verbesserung der Trinkwasserqualität oder der Schulausbildung sein. Oft fordern die Geldgeber in Entwicklungshilfeprojekten, dass der entsprechende Nachweis für die Effektivität einer Maßnahme erbracht werden kann. Das Buch „Poor Economics“ (Banerjee und Duflo 2011) und die Arbeiten des Poverty Action Lab (J-PAL, http://www.povertyactionlab.org/) dokumentieren dies eindrucksvoll. Aber auch in anderen Politikbereichen wie der Bildungs- oder Arbeitsmarktforschung finden sich zahlreiche Anwendungen.

Allerdings sind randomisierte Experimente nicht in allen Fällen anwendbar. Diese Verfahren setzen voraus, dass es sich um neue Maßnahmen, nicht nur um die Fortführung bereits bestehender Programme handelt. Die Randomisierung erfordert auch, dass nicht alle Individuen oder Unternehmen, die eine Maßnahme nachfragen, auch gefördert werden, zumindest nicht sofort. Es muss sichergestellt sein, dass die Nichtgeförderten keine Ausweichmöglichkeiten nutzen können, um dennoch eine vergleichbare Förderung zu erhalten. Und schließlich sollten die Ergebnisse, die im Experiment erzielt wurden, auf das zu evaluierende Programm insgesamt übertragen werden können (sogenannte externe Validität des Experiments). Nicht alle diese Voraussetzungen sind in der Realität immer erfüllt oder erfüllbar. Hieraus wird oftmals voreilig gefolgert, die entsprechenden Methoden seien eben nicht anwendbar. Wie die Erfahrung in den USA zeigt, können durch eine gute Planung von Feldexperimenten solche Probleme aber durchaus vermieden werden (Arni 2012).

Wenn eine Randomisierung der Maßnahme nicht durchführbar ist, gibt es zudem eine Reihe weiterer Möglichkeiten für die quantitative Wirkungsanalyse. Verfahren wie das Instrumentvariablen-Verfahren, der Regressions-Diskontinuitäten-Ansatz oder das Differenz-von-Differenzen-Verfahren[7] führen unter bestimmten Voraussetzungen zu ebenso validen Ergebnissen wie ein Feldexperiment (DiNardo und Lee 2010). Sie verlangen allerdings eine sorgfältige Begründung, weshalb die vorgefundene Variation für den Zweck der Evaluation so gut wie zufällig ist. Beispielsweise kann durch die Ausnutzung „natürlicher“ Variation in der Vergabe der Maßnahme das randomisierte Feldexperiment zumindest approximiert

[7] Zu den Verfahren im Einzelnen siehe Kugler et al. (2014).

werden. So können administrativ verfügte Schwellenwerte wie eine zufällige Zuteilung der Maßnahme wirken. Ändern sich die Schwellenwerte für eine Förderung, können beispielsweise Unternehmen, die nach einer Änderung gerade nicht mehr für die Maßnahme in Frage kommen, mit solchen verglichen werden, die gerade noch in den Genuss der Förderung kommen.

Die vielfache Anwendung quasi-experimenteller Methoden und die verbesserten Qualitätsstandards in der Evaluationsforschung haben das Bewusstsein dafür geschärft, welche Annahmen für das jeweilige Verfahren kritisch sind und welche in bestimmten Kontexten als plausibel angenommen werden können. Nicht für jeden Zweck wird sich ein geeignetes Quasi-Experiment finden. Doch führt der wissenschaftliche Wettbewerb dazu, dass natürliche Experimente erkannt und für die Kausalanalyse genutzt werden. Das Potenzial für derartige Analysen ist längst nicht ausgeschöpft.

Quantitative Wirkungsanalysen bilden die Grundlage, um die gesamtwirtschaftliche oder fiskalische Kosten-Nutzen-Bilanz der jeweiligen Maßnahme aufzustellen. In solchen Effizienzanalysen werden zunächst auf der Basis der geschätzten Wirkungen die fiskalischen oder gesamtwirtschaftlichen Erträge errechnet; diese werden dann den Kosten gegenübergestellt. Der Schritt von der Wirkungs- zur Effizienzanalyse birgt zwar methodische Probleme, die keineswegs trivial sind (Heckman und Smith 1998). Diese Probleme allein können jedoch nicht erklären, warum die verfügbaren Methoden in Deutschland so selten zur Anwendung kommen.

3 Evidenzbasierte Politik am Beispiel der USA

3.1 Frühe Erfahrungen mit Feldexperimenten

Die Vereinigten Staaten können auf jahrzehntelange Erfahrung im Bereich evidenzbasierter Politik zurückschauen. Schon in den 1960er Jahren wurden die ersten, zum Teil sehr groß angelegten Feldexperimente durchgeführt, vornehmlich in den Bereichen Bildungs-, Arbeits-, Gesundheits- und Sozialpolitik.[8] Ein mittlerweile aus Lehrbüchern bekannter Klassiker ist das Projekt STAR, das in den späten 1980er Jahren im Bundesstaat Tennessee durchgeführt wurde. Im Rahmen dieses Experiments wurde erprobt, ob kleinere Klassengrößen die Leistungen der Schüler verbessern. Dafür wurden die Schüler durch einen Zufallsmechanismus einer größeren beziehungsweise einer kleineren Klasse zugeordnet (Stock und Watson, 2006: S. 390ff., Krueger 1999).

[8] Zur Verbreitung randomisierter Feldexperimente in diesen Gebieten vergleiche den Digest of Social Experiments (Greenberg und Schroder 2004), der mehr als 240 experimentelle Evaluationen aufführt.

Eine besonders langfristig angelegte Untersuchung war die in den 1960ern initiierte Perry-Studie in Michigan, die die Wirksamkeit frühkindlicher Förderung untersuchen sollte.[9] Die geförderten Kinder im Vorschulalter stammten aus benachteiligten Familien und wurden zufällig in eine Behandlungs- und Kontrollgruppe eingeteilt und bis zum Alter von 40 Jahren regelmäßig nach ihren Lebensumständen befragt. Auf der Grundlage dieser Studie wurden umfangreiche Kosten-Nutzen-Analysen durchgeführt, die die Bildungs- und Arbeitsmarkteffekte, aber auch die Effekte verringerter Kriminalität zu quantifizieren versuchten.

Im Bereich des Gesundheitswesens setzte das RAND Health Experiment in den 70er Jahren Maßstäbe für spätere Studien. Im Rahmen dieser Untersuchung wurden Haushalten an sechs verschiedenen Standorten auf Zufallsbasis verschiedene Versicherungspläne angeboten, um herauszufinden, welchen Einfluss die Ausgestaltung der Krankenversicherung auf die Ausgaben für Gesundheitsdienstleistungen hat. Darauf aufbauend wurde in Oregon ab 2008 mit Hilfe eines Zufallsexperiments getestet, wie sich der Zugang zu einem erweiterten Versicherungsschutz auf die Nachfrage nach Gesundheitsleistungen auswirken würde.

Diese frühen Erfahrungen mit randomisierten Studien, zusammen mit einer hochentwickelten methodischen Forschung durch Pioniere der empirischen Wirtschaftsforschung wie James Heckman, Donald Rubin oder Burt Barnow, haben die Vereinigten Staaten zum Vorreiter auf dem Gebiet der evidenzbasierten Politik gemacht. In vielen Fällen hatten die Ergebnisse Konsequenzen für die konkrete Gestaltung wirtschaftspolitischer Maßnahmen. Anreize zur Aufnahme von Arbeit im „Aid to Families with Dependent Children“ (AFDC)-Programm, das in den 1980er Jahren experimentell evaluiert wurde, erwiesen sich hinsichtlich der künftigen Erwerbstätigkeit und Erwerbseinkommen als sehr wirksam; dies beeinflusste wiederum die Gesetzgebung des Bundes und der Bundesstaaten erheblich (Moffitt 2004). Die National Job Training Partnership Act (JTPA) Study ergab, dass die Wirksamkeit von Trainingsmaßnahmen für benachteiligte Jugendliche und Erwachsene deutlich geringer war als erhofft. Diese Erkenntnis führte zu deutlichen Budgetkürzungen (Smith 2009).

Einschränkend muss gesagt werden, dass sich die Anwendung dieser Wirkungsanalysen in den USA im Wesentlichen auf die genannten Politikbereiche beschränkt. So gibt es bisher kaum Beispiele der Anwendung im Bereich Landwirtschaft, der Verteidigung, oder der Förderung von kleinen und mittleren Unternehmen (KMU). Insofern gibt es insgesamt ein großes Maß an Heterogenität zwischen den einzelnen Politikbereichen und Regierungsinstitutionen.

---

[9] Siehe z.B. Schweinhart et al. (2005). Wie bei anderen Feldexperimenten wurden die Daten häufig in Sekundäranalysen ausgewertet, und die Schlussfolgerungen hinsichtlich der Qualität des Experimentes waren durchaus unterschiedlich (Hanushek und Lindseth 2009, Heckman et al. 2010).

3.2 Mittelvergabe und Wirkungsanalyse

In den vergangenen Jahren wurden Evaluation und Wirkungsforschung in den USA noch stärker im politischen Prozessen verankert, indem die Mittelvergabe von der Qualität der bereits erfolgten bzw. geplanten Evaluation abhängig gemacht wurden (Haskins and Baron 2011). Im Jahr 2001 wurde ein Program Assessment Rating Tool „PART“ eingeführt, mit Hilfe dessen Förderprogramme nach ihrer Effektivität bewertet werden können. Ziel ist es, ineffektive Programme von der weiteren Förderung ausnehmen zu können (http://strategisys.com/omb_part). Die Umsetzung von PART erfolgt durch das Office of Management and Budget (OMB), einer Institution, die aus dem Weißen Haus heraus die budgetären Prozesse in den USA steuert. Eine direkt vergleichbare Institution dieser Art existiert in Deutschland nicht.

Besonders forciert wurde das Ziel einer evidenzbasierten Wirkungsforschung im Zusammenhang mit den Fiskalprogrammen und den Programmen zur Bankenrettung in der Finanzkrise der Jahre 2008/2009. Auch hier kommt dem Office of Management and Budget eine zentrale Rolle zu. Es setzt die Standards für die Wirkungsanalysen. Ziel müsse es sein, so der ehemalige OMB-Direktor Peter Orszag, die Evaluationsstandards in die „DNA“ der Programme einzubauen (zitiert nach Liebman 2013). Bereits in der Planungsphase einer Maßnahme soll daher die spätere Evaluation berücksichtigt werden. Dies senkt die Kosten der Evaluation und zwingt die für die Umsetzung Zuständigen, die Evaluationskriterien bereits vorab klar zu definieren. Viele der neuen Programme machen deshalb eine rigorose Wirkungsanalyse zur Voraussetzung für die Bewilligung der Mittel.

Nach den Leitlinien des OMB sollte eine wettbewerbliche Mittelvergabe daran gebunden werden, wie überzeugend der Nachweis der Wirksamkeit einer Maßnahme ist. Der Großteil der Fördermittel soll an Projekte gehen, deren Wirksamkeit nachgewiesen ist, ein weiterer Teil der Mittel soll an Projekte vergeben werden, zu deren Wirksamkeit erste positive Evidenz vorliegt, und schließlich ist ein kleinerer Teil der Mittel für „innovative“ Maßnahmen vorgesehen, die erst in der Folge evaluiert werden.

Ein Beispiel für eine Mittelvergabe nach diesem Dreistufen-Standard ist das Investing in Innovation Fund (i3) Competitive Grant Program des Department of Education (Liebman 2013). Hier werden sogenannte Scale-up grants für Programme bereitgestellt, für die es bereits starke Evidenz über die Wirksamkeit aus randomisierten Studien und rigorosen quasi-experimentellen Studien gibt. Validation grants werden für Programme gewährt, für die es weniger belastbare Evidenz gibt, beispielsweise aufgrund einer zu geringen Stichprobengröße oder einer möglichen Verzerrung durch Selektion in eine bestimmte Maßnahme. Schließlich sollen mit Development grants bisher ungetestete Projekte gefördert werden, die ein hohes Potential haben und die im Zuge der Projektumsetzung weiter evaluiert werden sollen (Liebman 2013).

Die Umsetzung dieser Standards ist in den einzelnen US-Behörden unterschiedlich intensiv verbreitet. Besonders überzeugend ist die Evaluationspraxis im Department of Labor umgesetzt. Seit 2009 gibt es dort ein Chief Evaluation Office (CEO), das in die Planung von allen Evaluationen eingebunden ist und dafür sorgen soll, dass wirksame Maßnahmen identifiziert und weniger wirksame Maßnahmen verbessert werden. Für diese Aufgaben stehen dem Office Mittel zur Verfügung, die sich aus den Budgets der einzelnen Projekte speisen. Bisher flossen 0,5% der Mittel eines jeden Projektes an das CEO; ab dem Jahr 2013 soll dieser Anteil auf bis zu 1% ansteigen. Die an unabhängige Institute in Auftrag gegebenen Evaluationen sollen einem wissenschaftlichen Gutachterprozess unterliegen, die genutzten Daten öffentlich zugänglich gemacht werden.

Neben der relativ ausgedehnten Praxis der Wirkungsanalysen werden in den USA regelmäßig Effizienzanalysen durchgeführt; ihre Ergebnisse entscheiden mitunter darüber, ob Programme weitergeführt oder gestoppt werden. Eine Vorreiterfunktion spielt der Bundesstaat Washington, der mit dem Washington State Institute for Public Policy seit 1983 eine Forschungseinrichtung betreibt, die Wirkungsanalysen für den Gesetzgeber durchführt. Der Ansatz der quantitativen Evaluation wurde zunächst im Bereich des Justizsystems (z.B. für Maßnahmen zur Reintegration von Straftätern) angewendet und später auf weitere Bereiche ausgedehnt. In Publikationen werden die Kosten-Nutzen-Relationen der einzelnen Programme ausgewiesen (http://www.wsipp.wa.gov/rptfiles/12-04-1201.pdf). Dies ermöglicht es, Programme hinsichtlich ihrer fiskalischen Effizienz in eine Rangordnung zu bringen.

3.3 Qualitätssicherung durch Transparenz

Neben einer wissenschaftlichen Diskussion über die besten Verfahren und Methoden ist die Transparenz über die Durchführung der Wirkungsstudien Dreh- und Angelpunkt einer besseren evidenzbasierten Wirtschaftspolitik. Eine umfangreiche Dokumentation von Wirkungsstudien im Bildungsbereich in den USA stellt die Webseite „What Works Clearinghouse“ bereit. Hier werden Studien zu verschiedenen Bildungsprogrammen dokumentiert und methodisch bewertet. Anhand eines ausführlichen Kriterienkatalogs wird die methodische Qualität eingestuft. Danach entsprechen nur etwa 40% der begutachteten Studien hohen wissenschaftlichen Maßstäben. Diese wissenschaftlich fundierten Studien wiederum finden über verschiedene Maßnahmen hinweg in etwa der Hälfte der untersuchten Fälle positive oder potentiell positive Wirkungen – im Umkehrschluss heißt dies, dass viele der durchgeführten Maßnahmen die angestrebten Ziele nicht erreichen.

Ähnlich veröffentlicht auch das Department of Justice Wirkungsanalysen aus dem Bereich der Kriminalitätsbekämpfung. Auf der Webseite „CrimeSolutions.gov“ werden die Ergebnisse von Wirkungsanalysen zu Maßnahmen aufgelistet, die beispielsweise auf die Wiedereingliederung straffällig gewordener Jugendlicher ausgerichtet sind.

Und auch das OMB fordert die umsetzenden Behörden auf, bereits durchgeführte Evaluationen im Internet öffentlich zugänglich zu machen. Dies ist vergleichbar mit einer Website zu klinischen Testreihen (HHS clinical trial and results data bank, ClinicalTrials.gov). Die Website soll Informationen über Evaluationen breit zugänglich machen und verhindern, dass negative Testergebnisse verschwiegen werden.

4 Evaluationskultur in Deutschland

Die Evaluation wirtschaftspolitischer Maßnahmen im Auftrag der für die Maßnahmen verantwortlichen Behörden ist in Deutschland noch stärker als in den USA auf einige wenige Politikbereiche begrenzt. Wirkungsanalysen werden von der Politik vor allem in den Bereichen Arbeitsmarkt- und Bildungspolitik und in der Familienpolitik in Auftrag gegeben. Laufende Programme werden jedoch ausschließlich quasiexperimentell oder nichtexperimentell evaluiert.

Anders als in den USA fehlen Studien, die einen Vergleich zu zufällig zusammengestellten Vergleichsgruppen ermöglichen. Eine Ausnahme bilden Evaluationen einzelner Pilotmaßnahmen. Bislang werden wenig quantitative Wirkungsanalysen in den Bereichen Forschungs- und Technologiepolitik, Steuerpolitik, Agrar- oder Entwicklungspolitik eingesetzt. Zwar gibt es in diesen Bereichen durchaus Studien zur Wirksamkeit wirtschaftspolitischer Maßnahmen, die auf die Initiative von Wissenschaftlern zurückgehen. Diese sind in ihren Möglichkeiten und Aussagen jedoch begrenzt, da der dafür notwendige Zugriff auf relevante (Mikro-)Daten ohne Unterstützung der Politik oft nicht möglich ist.

4.1 Wirkungsanalysen in der Arbeitsmarktpolitik

Am umfangreichsten sind Wirkungsanalysen in Deutschland bisher im Bereich der aktiven Arbeitsmarktpolitik, der Reform der sozialen Mindestsicherung sowie der Arbeitsmarktregulierung zum Einsatz gekommen. Dies ist eine relativ neue Entwicklung, denn lange Zeit waren quantitative Wirkungsanalysen aufgrund der mangelhaften Datenlage nicht möglich. Die bisherige Entwicklung lässt sich grob in vier Phasen gliedern:

In der ersten Phase bis etwa Ende der 1990er Jahre wurden quantitative Wirkungsanalysen nahezu ausschließlich auf Initiative der Wissenschaft durchgeführt. Forschung im Auftrag der Politik war grundsätzlich qualitativ ausgerichtet oder basierte auf vergleichsweise kleinen Befragungsdatensätzen wie dem Sozioökonomischen Panel (SOEP) oder dem Arbeitsmarktmonitor Ost. Die zu geringen Fallzahlen erlaubten keine aussagekräftigen Schlüsse auf die Wirksamkeit arbeitsmarktpolitischer Maßnahmen; schon gar nicht waren sie geeignet, differenzierte Aussage über einzelne Gruppen am Arbeitsmarkt oder bestimmte Maßnahmevarianten bereitzustellen (Fitzenberger und Speckesser 2000).

Die zweite Phase begann mit der Einführung des SGB III im Jahre 1998, durch die eine gesetzliche Verpflichtung zur Evaluation der aktiven Arbeitsmarktpolitik geschaffen wurde (heute § 282 SGB III). In Pilotstudien wurden Forscherteams mit der Erschließung administrativer Daten für die Evaluation von Fortbildung, Umschulung und Arbeitsbeschaffungsmaßnahmen beauftragt (Bender et al. 2005, Biewen et al. 2006).

Die Pilotprojekte der Phase zwei wurden wegweisend für Evaluationen, die in der dritten Phase (ab 2002) zu größeren Evaluationsprojekten führten. Von besonderer Bedeutung war dabei insbesondere die stärkere gesetzliche Verankerung der Wirkungsforschung im sogenannten Job-Aqtiv-Gesetz und in der Evaluation der Hartz-Gesetze (Heyer 2002). Zentral war hier die Bereitstellung von Geschäftsdaten der Bundesagentur für Arbeit an die Forschung durch das Institut für Arbeitsmarkt- und Berufsforschung. Aufbauend auf den Ergebnissen der genannten Pilotstudien wurden die Integrierten Erwerbsbiographien (IEB) als Standardformat für administrative Arbeitsmarktdaten geschaffen. Hierin sind Daten aus den administrativen Prozessen der Arbeitsvermittlung, Leistungsgewährung, Maßnahmenteilnahme und Sozialversicherung vereint. Diese Daten enthalten die für eine quantitative Evaluation von arbeitsmarktpolitischen Maßnahmen wesentlichen Informationen (Lechner und Wunsch 2013) und sind insofern als besser geeignet einzustufen denn vergleichbare administrative Daten in den USA (Smith 2009).

Die auf der Grundlage dieser Daten gewonnenen Ergebnisse quasi- und nichtexperimenteller Studien für Deutschland sind deshalb belastbarer als die vergleichbarer Studien in den USA, was zumindest in gewissem Umfang die geringere Verwendung von Feldexperimenten kompensiert. Die weitgehende Unwirksamkeit der Arbeitsbeschaffungsmaßnahmen (ABM) für die Wiedereingliederung in Arbeit wurde in diesen Studien etabliert und führte dazu, dass diese Maßnahmen nach und nach aus dem Instrumentarium der Arbeitsmarktpolitik verschwanden.

Seit Beginn der Phase vier (ab 2008) werden Evaluationen regelmäßig durchgeführt und vermehrt in der Gesetzgebung verankert – dies gilt beispielsweise für die gesetzliche Evaluation der bestehenden Mindestlöhne. Bisher bestehende Lücken in der Evaluation arbeitsmarktpolitischer Maßnahmen, beispielsweise hinsichtlich der Maßnahmen für benachteiligte Jugendliche (Eichhorst und Zimmermann 2007), wurden zumindest zum Teil geschlossen. Maßnahmen des Europäischen Sozialfonds wie z.B. der Kommunal-Kombi, ein Programm mit öffentlich geförderter Beschäftigung, werden regelmäßig auf den Prüfstand quantitativer Evaluationen gestellt.

Ursächlich für die zunehmende Verbreitung von quantitativen Wirkungsanalysen in der Arbeitsmarktpolitik zu Beginn der 2000er-Jahre dürften mehrere Entwicklungen gewesen sein. Zum einen war angesichts eher ernüchternder Ergebnisse der bis dahin durchgeführten Evaluationen und engerer Budgetspielräume eine Verbesserung der Zielgenauigkeit der Arbeitsmarktpolitik offensichtlich erforderlich. Dies wurde nicht nur von der Wissenschaft

angemahnt, sondern fand auch Unterstützung im Bundesministerium für Arbeit, dessen Mitarbeiter eine verstärkte Wirkungsforschung selbst auf die Agenda schrieben (Heyer, 2002). Die Notwendigkeit von Reformen am Arbeitsmarkt, die die „Agenda 2010“ und die Hartz-Reformen hervorbrachte, verstärkte ebenfalls die Nachfrage nach Evaluationen, auch wenn die Wirkungsforschung bereits zuvor gesetzlich verankert worden war. Schließlich befand sich die ehemalige Bundesanstalt für Arbeit in den Jahren nach 2002 in einem grundlegenden Wandlungsprozess zu einem modernen Dienstleister, und dazu gehörte auch eine verbesserte Evidenzbasierung ihrer Arbeit.

4.2 Was könnte verbessert werden?

Trotz großer Fortschritte gibt es jedoch nach wie vor wichtige Defizite in der Evaluation der Arbeitsmarktpolitik in Deutschland:

Erstens werden bislang nur selten Studien mit randomisierten Kontrollgruppen durchgeführt. Der Einsatz randomisierter Kontrollgruppenvergleiche in der Evaluation der Arbeitsmarktpolitik begann mit regional eng begrenzten Pilotprojekten (Schiel et al. 2006) und wurde mit Evaluationsprojekten zu einzelnen Maßnahmen fortgesetzt (Krug und Stephan, 2013). Experimentelle Verfahren werden jedoch in der Evaluation der Arbeitsmarktpolitik bislang nicht in größerem Umfang verwendet. Andere europäische Länder sind Deutschland in dieser Hinsicht voraus, beispielsweise Dänemark (Vikström et al. 2013) und Frankreich (Behagel et al. 2012).

Zweitens wird zwar die *Wirkung* arbeitsmarktpolitischer Interventionen vielfach untersucht, nicht jedoch die *Effizienz* der Maßnahmen. Vorhandene Effizienzanalysen sind häufig nicht überzeugend, da sie auf unzureichenden Daten beruhen; dies liegt vor allem an der unzureichenden Disaggregation von Kostendaten der Bundesagentur für Arbeit in der aktiven Arbeitsmarktpolitik (Heyer et al. 2011). Da frühe Evaluationsstudien in vielen Fällen überwiegend keine signifikanten Ergebnisse arbeitsmarktpolitischer Maßnahmen nachweisen konnten, war eine darauf aufsetzende Effizienzanalyse überflüssig. Mittlerweile zeichnen sich jedoch für viele (reformierte) Maßnahmen positive Wirkungen ab, so dass die Frage der Relation von Kosten und Nutzen zunehmend relevanter wird.

Drittens fehlt es an einer systematischen Koordination von politischen Prozessen und Evaluationsstudien. Vielfach werden Evaluationen erst ausgeschrieben, wenn das zu überprüfende Programm bereits angelaufen ist. Dies macht eine Anwendung randomisierter Forschungsdesigns unmöglich. Darüber hinaus führt dies auch oft dazu, dass die Evaluierenden relevante Daten nicht nutzen können, weil versäumt wurde, das Einverständnis der Teilnehmenden dafür einzuholen.

Viertens ist nicht nur wichtig, dass Evaluationen durchgeführt werden, sondern dass Qualitätsstandards beachtet werden und auf Seiten der Evaluierenden und der Politik aus den

Erfahrungen systematisch gelernt wird. Eine Mindestanforderung ist, dass Zwischen- und Abschlussberichte publiziert werden, wie es derzeit in der Regel (aber nicht immer) der Fall ist. Ferner darf die wissenschaftliche Qualitätskontrolle, die zum großen Teil durch den wissenschaftlichen Begutachtungsprozess bei Zeitschriften gewährleistet wird, nicht durch Publikationsvorbehalte behindert werden. Replikationen von Evaluationsergebnissen als stärkste Form der Qualitätsprüfung (Abschnitt 4) finden bislang praktisch nicht statt.

Zweifel sind zudem angebracht, ob im politischen Prozess die richtigen Schlüsse aus den Ergebnissen der Evaluationen gezogen werden. Ein gutes Beispiel dafür ist die Förderung der Gründungstätigkeit von Arbeitslosen. Diese stellte sich im Rahmen der Hartz-Evaluation als eines der effektivsten Instrumente der Arbeitsmarktpolitik heraus (Eichhorst und Zimmermann 2007). Umso verwunderlicher ist es, dass die Förderung in der Folge sukzessive immer weiter eingeschränkt wurde.

Das Beispiel der deutschen Arbeitsmarktpolitik zeigt, dass in der Evaluierungspraxis deutliche Fortschritte erzielt wurden. Zugleich zeigt die dabei gewonnene Erfahrung, dass es mit der Durchführung von quantitativen Wirkungsanalysen allein noch nicht getan ist. Entscheidend ist die Einhaltung hoher Qualitätsstandards, das Ausschöpfen des ganzen Potentials der Evaluationsmethodik inklusive randomisierter Studien, die Durchführung von Effizienzanalysen und die systematische Koordination und Rückkopplung der politischen Prozesse mit den Evaluationsprozessen.

5 Hindernisse für eine stärkere Nutzung evidenzbasierter Wirtschaftspolitik

Die Frage, warum in Deutschland bisher nur in relativ geringem Umfang Wirkungsanalysen durchgeführt werden, wird typischerweise damit beantwortet, dass viele Politikmaßnahmen nicht für Evaluationen geeignet und die Kosten für quantitative Wirkungsanalysen unverhältnismäßig hoch seien. Diese Einwände müssen ernst genommen und im Einzelfall geprüft werden. Ein wichtiger Grund für den zögerlichen Einsatz von Wirkungsanalysen scheint aber auch ein politökonomischer zu sein: weder die verantwortlichen Politiker noch die betroffenen Wähler oder die mit der Durchführung betrauten Forscher scheinen ein echtes Interesse an verlässlichen Evaluationsergebnissen zu haben.

5.1 Sind die Anwendungsmöglichkeiten für Evaluationen begrenzt?

Quantitative Wirkungsanalysen von Politikmaßnahmen, die den methodischen Qualitätsstandards genügen, sind an Voraussetzungen geknüpft, die nicht immer leicht zu erfüllen sind (Heckman et al. 1999, Wooldridge and Imbens 2009). Die zu untersuchende Politikmaßnahme muss klar definiert und gegen andere Maßnahmen abgrenzbar sein. Es müssen zwei Gruppen von Personen oder Unternehmen unterscheidbar sein, eine Gruppe, die eine Politikmaßnahme erfährt (die Behandlungsgruppe) und eine Gruppe, die von der

Politikmaßnahme nicht betroffen ist (die Kontrollgruppe). Diese beiden Gruppen sollten sich in allen anderen Belangen nicht systematisch voneinander unterscheiden. Zwischen den Gruppen sollten keine Wechselwirkungen bestehen und die Teilnehmer nicht zwischen den Gruppen wechseln können. Schließlich müssen die Auswirkungen der Politikmaßnahme messbar sein, so dass sie mit den vorhandenen Daten und statistischen Verfahren signifikant nachgewiesen werden können.

5.1.1 Abgrenzung der Gruppen

Gerade die Abgrenzung von Personen, die durch die Politikmaßnahme betroffen sind, und solchen, für die das nicht der Fall ist, scheint auf den ersten Blick oft sehr schwierig. Und selbst wenn diese Abgrenzung möglich ist, kann nicht immer sichergestellt werden, dass sich die betrachteten Gruppen nicht systematisch voneinander unterscheiden, was eine kausale Wirkungsanalyse erschweren kann. Dass Evaluationen bisher vor allem in sehr spezifischen Politikbereichen durchgeführt werden, in denen diese Voraussetzungen vergleichsweise leichter zu erfüllen sind, ist deshalb kein Zufall.

Im Bereich der Entwicklungspolitik werden typischerweise Maßnahmen untersucht, die relativ klar definiert und deren Anwendung jeweils auf eine bestimmte Personengruppe beschränkt ist. Auch im Arbeitsmarkt- und Bildungsbereich zielen die meisten Maßnahmen und Projekte auf eine mehr oder weniger klar umrissene Gruppe von Teilnehmern. Dagegen betreffen zum Beispiel Änderungen im Einkommensteuertarif prinzipiell alle Steuerpflichtigen, so dass es keine Kontrollgruppe zu geben scheint, anhand derer das Szenario unveränderter Tarife nachgebildet werden kann. Tatsächlich gibt es jedoch auch hier Ansätze, um diese Unterscheidung treffen zu können. So sind z.B. Personen, deren Verdienst vor und nach der Änderung knapp unterhalb des steuerlichen Freibetrags lag, nicht direkt betroffen und können daher in einem „natürlichen Experiment“ als Kontrollgruppe dienen (Blundell et al. 1998).

Aber auch in Politikbereichen, die bisher nur wenig im Fokus von Evaluationsstudien standen, ist eine stringente Wirkungsanalyse möglich, insbesondere dann, wenn aus Budgetgründen nicht alle Personen oder Unternehmen gefördert werden können. Dies gilt beispielsweise im Bereich der Innovationsförderung im Unternehmenssektor. Wie Chatterji et al. (2013) in einer aktuellen Übersicht über die Evaluation von Clusterpolitik und Innovationsförderung zeigen, kann beispielsweise die Vergabe von Plätzen in einem geförderten Industrie- oder Gewerbegebiet nach dem Zufallsprinzip erfolgen, um so eine bessere Grundlage für die anschließende Wirkungsanalyse zu haben.

Schwieriger stellt sich die Evaluation von wirtschaftspolitischen Maßnahmen dar, die per Definition nur wenige Unternehmen betreffen. Im Bereich der Luftfahrtförderung würde man beispielsweise keine vergleichbare Gruppe von Unternehmen finden, die keine Förderung erhalten hat. Solche Ausnahmebereiche, in denen Evaluation der geschilderten Art nicht

möglich sind, wird es immer geben. Daraus kann aber kein Einwand gegen Evaluationen anderer Maßnahmen abgeleitet werden. Und schließlich sollte die Information, dass in bestimmten Ausnahmebereichen eine Überprüfung der Zielerreichung nicht möglich ist, auch in den politischen Entscheidungsprozess einfließen.

5.1.2 Ethische und rechtliche Aspekte

Vor allem in randomisierten Feldexperimenten sind ethische und rechtliche Bedenken zu berücksichtigen. So wird oft eingewendet, dass eine Vergabe von staatlichen Fördermitteln nach dem Zufallsprinzip gegen das Gleichheitsgebot verstoße. Ähnliche Einwände könnten grundsätzlich auch im Fall von Medikamententests vorgebracht werden. Dennoch ist Randomisierung im Fall von Medikamententests weit weniger umstritten. Denn wenn Medikamente getestet werden, ist in der Regel nicht offensichtlich, dass es sich um Begünstigungen handelt. Medikamente können positive, aber eben auch negative Wirkungen haben. Erst wenn die positive Wirkung eines Medikaments hinreichend nachgewiesen ist, gebietet es die Ethik, den Wirkstoff allen Patienten zur Verfügung zu stellen.

Bei wirtschaftspolitischen Maßnahmen erwarten jedoch viele, durch die Maßnahme begünstigt zu werden. Fördermaßnahmen für Unternehmensgründer beispielsweise kommen den betroffenen Personen direkt zugute. Verbietet es dann nicht der Gleichbehandlungsgrundsatz, die Kontrollgruppe von der Maßnahme auszuschließen? Die Frage ist nicht so eindeutig zu beantworten, wie es zunächst scheint. Denn auch ohne Randomisierung ist jemand, der eine Gründungsförderung nicht erhalten hat – sei es, weil ihm die Informationen fehlten, sei es, weil er die Kriterien nicht erfüllt hat – letztlich benachteiligt. Und ob der geförderte Gründer sich ein nachhaltiges Geschäftsmodell aufbauen kann oder sich nicht mit einer anderen beruflichen Entscheidung besser gestellt hätte, ist ebenfalls eine offene Frage.

Tatsächlich ist aufgrund einer Budgetbeschränkung häufig eine Rationierung ohnehin unumgänglich. Es muss nach bestimmten Auswahlkriterien oder rein auf Grund von administrativen Ermessensspielräumen entschieden werden, wer die knappen Mittel erhält. Wenn aber ohnehin ausgewählt werden muss, dann ist das Zufallsprinzip oft noch die fairste Art der Rationierung. Vor allem in Entwicklungsländern wird eine Verteilung nach dem Zufallsprinzip bei grundsätzlich knappen und damit rationierten Ressourcen geschätzt, weil sie ein erheblicher Fortschritt gegenüber der sonst üblichen Korruption bei Verteilungsverfahren ist und jedem zumindest die gleiche faire Chance auf Förderung einräumt.

Eine Vergabe nach dem Zufallsprinzip muss auch nicht bedeuten, dass die Teilnehmer der Kontrollgruppe auf immer von der Maßnahme ausgeschlossen werden müssen. Für die Wirkungsanalyse ist es schon ausreichend, wenn die Maßnahme zeitlich versetzt eingeführt wird. Die kurzfristige Benachteiligung einiger ist dann gegen den Erkenntnisgewinn für alle

abzuwägen. Denkbar ist auch, verschiedene Varianten von Maßnahmen experimentell gegeneinander zu testen.

Wichtig ist es in jedem Fall, darüber aufzuklären, in welchen Bereichen randomisierte Feldexperimente sinnvoll möglich sind und in welchen Bereichen andere Verfahren der Evaluationsforschung zum Einsatz kommen sollten. Hier hat die Wissenschaft eine wichtige Informationsfunktion, die sich auf die Aufklärung über die Möglichkeiten, aber auch über die Grenzen von Evaluationsmethoden bezieht.

5.2 Stehen Kosten und Erträge einer Evaluation in einem angemessenen Verhältnis?

Auch wenn Maßnahmen mit wissenschaftlichen Methoden evaluierbar sind, wird oft eingewendet, eine Evaluation lohne sich nicht, weil die Kosten der Evaluation ihren Nutzen überstiegen. Um eine wissenschaftlichen Standards entsprechende Evaluation durchzuführen, müssen zwei Voraussetzungen erfüllt sein. Zum einen müssen Daten in einem bestimmten Umfang und in einem bestimmten Format erhoben und aufbereitet werden. Evaluationsverfahren gelten als „datenhungrig“. Informationen müssen nicht nur von den betroffenen Unternehmen oder Personen selbst, sondern auch von einer entsprechenden Vergleichsgruppe erhoben werden. Gerade das spricht dafür, eine Evaluation frühzeitig anzulegen – denn müssen Informationen über eine Kontrollgruppe im Nachhinein erhoben werden, kann das umso teurer werden.

Zum anderen verursacht die Durchführung der entsprechenden Untersuchungen selbst Kosten. Bei knappen Projektbudgets, so die Befürchtung, seien dafür keine ausreichenden Mittel verfügbar. Hier gilt es die Kosten aus der Evaluation und die Kosten aus ineffektiven Maßnahmen gegeneinander abzuwägen. Problematisch ist dabei allerdings, dass Kosten und Erträge einer Evaluation zu unterschiedlichen Zeitpunkten anfallen. Kurzfristig müssen Ressourcen für die Evaluation aufgebracht werden, mögliche Erfolge im Sinne einer besseren und zielgerichteten Politik werden aber erst sehr viel später realisiert. Beispielsweise entscheidet sich erst nach mehreren Jahren, welchen Erfolg eine bestimmten Bildungs- oder Weiterbildungsmaßnahme hatte. Und schließlich könnten Kosten und Erträge in föderalen Systemen auch auf unterschiedlichen Ebenen anfallen: Maßnahmen zur Verbesserung der Bildung und Arbeitsmarktintegration auf regionaler oder lokaler Ebene können z.B. langfristig positive Effekte auf das Steueraufkommen auf föderaler Ebene haben.

Aus wohlfahrtsökonomischer Sicht ist nur die Frage erheblich, ob die Evaluationskosten den -nutzen insgesamt übersteigen. Kosten und Nutzen einer Evaluation sind nicht leicht abschätzbar. Jedoch fallen die Kosten der Evaluation erheblich geringer aus, wenn die Evaluation mit Einführung der Maßnahme mitgeplant und die dafür notwendigen Informationen miterhoben werden. Dies ist wesentlich kostengünstiger als eine nachträgliche Erhebung der erforderlichen Daten. Wichtig ist hierbei, dass die mit der Durchführung des Projekts beauftragten Personen über das Design der Evaluation informiert werden. Aus den

USA wird das Beispiel berichtet, bei der Einführung von *smart meters* seien Daten über die Vergleichsgruppe der Nichtgeförderten unzureichend erhoben worden – weil man es nicht für notwendig erachtet habe.

Ein wesentlicher Faktor, der die Kosten von Evaluationen senkt, ist die Nutzung von administrativen Daten, von Daten aus Geschäftsprozessen von Verwaltungen also. Gerade durch die in den letzten Jahren forcierte Einführung von Forschungsdatenzentren wurden auch in diesem Bereich wichtige Schritte gemacht, um die nötigen Voraussetzungen im Bereich der Dateninfrastruktur zu schaffen (Rat für Sozial- und Wirtschaftsdaten 2010, Wissenschaftsrat 2012). Denn so fallen nur die Kosten für Aufbereitung und Bereitstellung der Daten für die Evaluierenden zusätzlich an. Laufende Kosten bestehen dann nur noch im Zuschnitt der Daten für das jeweilige Projekt. Dies dürfte in fast allen Fällen um Größenordnungen günstiger sein als die Durchführung von Befragungen im erforderlichen Umfang. Wiederum können die USA als Beispiel dienen, hier werden Zugänge zu administrativen Daten der Regierung unter www.data.gov bereitgestellt.

Nicht immer ist die Möglichkeit gegeben, administrative Daten zu nutzen, weil sie vielfach nicht die erforderlichen Informationen erhalten; in diesem Fall kann aber über eine Verknüpfung von Befragungs- und Prozessdaten nachgedacht werden. Das Potenzial wird in Deutschland noch nicht ausgeschöpft. Beispiele sind Daten aus dem Bildungssystem, insbesondere Einzeldaten aus der Schulstatistik oder Daten von Förderprogrammen, die derzeit häufig noch in Form von Daten über die Gesamtzahl der Geförderten und nicht in Form von Einzeldaten von Personen oder Unternehmen geführt werden. Dazu müssen die Möglichkeiten, diese Daten mit anderen Datenbeständen zu verknüpfen, erweitert werden. Der föderale Staatsaufbau in Deutschland erschwert häufig die Schaffung guter Datengrundlagen. Von hoher Bedeutung wäre eine bessere Abstimmung der Länderbehörden, insbesondere bei den Datenschutzbeauftragten.

Und nicht zuletzt muss auch die Frage nach den Kosten des Status Quo gestellt werden. Derzeit werden vor allem qualitative Evaluationen auf der Grundlage von Fallbeobachtungen und nichtstandardisierten Befragungen vorgenommen. Diese Evaluationen sind ebenfalls teuer, können aber keine Aussagen zur kausalen Wirkung von Maßnahmen liefern. Und selbst bei standardisierten Befragungen fallen erhebliche Kosten an, zumindest wenn die Befragungen jeweils nach wissenschaftlichen Standards durchgeführt werden. Im Vergleich dazu nimmt sich der Aufwand für quantitative Evaluationen oft eher bescheiden aus. Ein Verweis auf die Kosten erscheint deshalb nicht als schlagkräftiges Argument dafür, qualitativen Evaluationen den Vorzug zu geben.

Ein Vergleich von Fördervolumina und Evaluationskosten zeigt, dass letztere oft verschwindend gering sind. Die bislang teuerste Evaluation in Deutschland, die Evaluation der Hartz I-III-Gesetze, kostete 10,3 Millionen Euro. Für aktive Arbeitsmarktpolitik wurden in Deutschland allein im Jahr 2002, bevor die Evaluation der Hartz I-III-Gesetze durchgeführt

wurde, 22,1 Milliarden Euro ausgegeben (Eichhorst und Zimmermann 2007), also mehr das 2000-fache.

5.3 Wer hat ein Interesse an Evaluationen?

Bei der Entscheidung über die Durchführung von Evaluationen sind grundsätzlich drei Gruppen betroffen: Politiker und gegebenenfalls Verwaltungsbehörden, die für die Politikmaßnahmen verantwortlich zeichnen und die Durchführung der Wirkungsanalysen in Auftrag geben, Bürger, die von den Politikmaßnahmen betroffen sind und als Steuerzahler für die Finanzierung dieser Maßnahmen aufkommen müssen, und die Forschungsinstitute bzw. die dort beschäftigten Wissenschaftler, die in der Regel die Evaluierungen durchführen. Ob Wirkungsanalysen in Auftrag gegeben und in welcher Qualität sie angeboten werden, hängt wesentlich von dem Zusammenspiel der Interessen dieser Gruppen ab.

Auf Seiten der Politik könnte der Anreiz, Evaluationen in Auftrag zu geben, nach einem Regierungswechsel besonders hoch sein, da dann die Programme der Vorgängerregierung auf den Prüfstand gestellt werden können. Als Beispiel wird hier oft die Einführung von Evaluationsstandards unter Präsident Bush in den USA genannt, der die von der Vorgängerregierung eingeführten sozialpolitischen Programme kritisch überprüfen wollte (Baron und Haskins 2011). Auch das im Koalitionsvertrag 2009 verankerte Anliegen, eine Wirkungskontrolle der bestehenden gesetzlichen Mindestlöhne in Deutschland hinsichtlich der Beschäftigungsfolgen durchzuführen, kann so interpretiert werden.

Aber auch ein Politiker, der sich daran macht, die im Wahlprogramm versprochenen Politikmaßnahmen umzusetzen, sollte ein Interesse daran haben, durch Wirkungsanalysen zu ermitteln, wie wirksam die eingesetzten Maßnahmen tatsächlich sind. Wie überzeugt er davon ist, dass das durchgeführte Projekt einer Evaluierung standhält, kann ein Politiker besonders glaubhaft signalisieren, wenn er mit Einführung der Maßnahme die Evaluation gleich mit in Auftrag gibt. Schließlich kann eine gut angelegte Evaluation es wahrscheinlicher machen, dass funktionierende Programme politische Machtwechsel überdauern, wie das Beispiel PROGRESA in Mexiko zeigt (Gertler 2004).[10]

Trotzdem ist nicht auszuschließen, dass Politiker eine Wirkungsanalyse scheuen, wenn sie befürchten, dass durch mögliche negative Ergebnisse ihre Wiederwahlchancen beeinträchtigt werden. Solange im öffentlichen Raum keine echte Evaluationskultur etabliert ist, die eine offene Diskussion über Vor- und Nachteile bestimmter Maßnahmen zur Erreichung eines politischen Ziels zulässt, wird die Politik möglicherweise sogar davor zurückschrecken,

[10] PROGRESA steht für ein gesundheits- und bildungspolitisches Programm (Programa de Educación, Salud y Alimentación), das in Mexiko durchgeführt wurde.

konkrete, überprüfbare Ziele ihres Handelns zu nennen. Diese wären aber die Voraussetzung für eine Wirkungsanalyse und Überprüfbarkeit von Maßnahmen.

Auf Seiten der Wähler bzw. Steuerzahler könnte ein besonders starkes Interesse vermutet werden, die Wirksamkeit der von ihnen finanzierten Maßnahmen auf den Prüfstand zu stellen. Allerdings sind die Gruppen, die von einer Maßnahme profitieren, oft sehr gut organisiert oder politisch relevant. Von der Solarförderung profitieren beispielsweise weite Teile der Industrie und viele Hausbesitzer. Stromkunden und Steuerzahler im Allgemeinen, die die hohen Kosten zu tragen haben, sind hingegen weniger gut organisiert.

Deshalb ist es wichtig, dass die Entscheidungen über Evaluationen nicht im Einzelfall und damit diskretionär von denjenigen, die die Maßnahme zu verantworten haben, gefällt werden. Vielmehr muss die Evaluation von Politikmaßnahmen ab einer bestimmten Fördersumme verpflichtend gemacht werden. Dabei sollte zugleich die Durchführbarkeit der Evaluation sichergestellt werden, damit tatsächlich Informationen über die Wirksamkeit der Maßnahmen gewonnen werden. Die Wähler sollten die Offenlegung dieser Wirkungsanalysen einfordern. Die Medien könnten diesen Meinungsbildungsprozess unterstützen.

Um den politischen Widerstand gegen Wirkungsanalysen zu überwinden, kann zudem ein Fokus auf vergleichende Evaluationen verschiedener Maßnahmen sinnvoll sein. Dies hilft herauszufinden, wie die Wirksamkeit einer Maßnahme gesteigert werden kann, ohne das politische Ziel selbst in Frage zu stellen. Die Anreize für solche Wirkungsanalysen können außerdem erhöht werden, wenn die Mittelvergabe an einen methodisch überzeugenden Nachweis der Wirkung der Maßnahme geknüpft wird. Das oben zitierte Beispiel der Setzung von Standards in den USA („PART“) kann als Beispiel dienen.

Hilfreich ist auch die Popularisierung von bestimmten Evaluierungskonzepten. Die Idee von randomisierten Experimenten ist vergleichsweise einfach zu kommunizieren. Die meisten wissen, was in einer medizinischen Studie eine Kontrollgruppe und was ein Placebo ist. Gerade die leichte Verständlichkeit hat vermutlich in den USA, aber auch in Entwicklungsländern dazu beigetragen, Unterstützung für Evaluationen zu gewinnen. Schwieriger ist es, nichtexperimentelle Evaluationen, die auf ökonometrischen Verfahren beruhen, verständlich zu machen. Aber auch in diesen Fällen kann man sich um Prägnanz bemühen. So hat die – in den meisten Fällen nicht ganz korrekte – Redeweise von einem „statistischen Zwilling“ (für eine vergleichbare Kontrollperson) das Matching-Verfahren popularisiert.

Mit der Durchführung der Evaluationsstudien werden in der Regel Forschungsinstitute betraut. Die dort beschäftigten Forscher und die von ihnen eingesetzten Methoden spielen daher eine besonders wichtige Rolle für die Qualität der Wirkungsanalysen. Eine entscheidende Voraussetzung für gute Wirkungsanalysen ist dabei eine gute Datenbasis. Nur sie gewährleistet eine methodisch überzeugende Analyse. Gleichzeitig sichert sie das wissenschaftliche Interesse der Forscher, da mit der Qualität der Datenbasis auch die

Publikationsmöglichkeiten deutlich steigen. Eine Offenlegung der Datenbasis zu Replikationszwecken stellt außerdem sicher, dass andere Forscher die Ergebnisse überprüfen können und so der wissenschaftliche Wettbewerb für die notwendige Qualitätssicherung sorgen kann. Allerdings sind im derzeitigen Publikationsbetrieb die Anreize, Evaluationen zu replizieren, nicht hinreichend groß. Wird ein Ergebnis bestätigt, ist das wissenschaftlich nicht interessant. Treten Abweichungen auf, muss sich der Forscher gegen eine etablierte Meinung durchsetzen – was oft die Veröffentlichungschancen nicht erhöht. Es handelt sich hierbei um ein generelles Phänomen, das aber bei Replikationsstudien besonders virulent ist. Daher ist die Wissenschaft gefordert, Evaluationen, Replikationsstudien und „Nicht-„Ergebnissen breiteren Raum zu gewähren (Economist 2013).

6 Institutionelle Rahmenbedingungen für eine stärker evidenzbasierte Wirtschaftspolitik

Viele Faktoren haben einen Einfluss darauf, ob eine wirtschaftspolitische Maßnahme evaluiert und in einem offenen Prozess mit unterschiedlichen Ansätzen zur Erreichung eines gegebenen politischen Ziels experimentiert wird. Die institutionellen Rahmenbedingungen spielen eine wichtige Rolle, aber auch die Bereitschaft der Politik und der Forschung, gemeinsam nach Lösungen zu suchen. Oft sind es aber auch schlicht einzelne Personen, die die Weichen in Richtung einer besseren Wirkungsforschung stellen können.

Im Folgenden wollen wir einige Ansätze skizzieren, die dazu beitragen könnten, mehr Evidenz in die deutsche Wirtschaftspolitik zu bringen. Wichtig ist zum einen ein Bekenntnis zu einer besseren empirischen Fundierung politischer Entscheidungen auf zentraler Ebene. Wichtig sind aber auch kleine Pilotprojekte und die Entwicklung neuer Methoden und Ansätze auf Projektebene. Um Skeptiker und Kritiker zu überzeugen, kann sich die Evaluation von Projekten als besonders hilfreich erweisen, die aktuell nicht im Kreuzfeuer der politischen Diskussion stehen.

6.1 Agenda-Setzung und Qualitätssicherung in der Evaluationsforschung

In den USA ist auf zentraler Ebene Initiative ergriffen worden. Es wurde eine klare Agenda für kausale Wirkungsanalysen gesetzt und Maßnahmen zur Qualitätssicherung in der Evaluationspraxis verankert. Durch die Reservierung eines festen Etats eines jeden Projektbudgets für Evaluationen wurde die Finanzierbarkeit der Wirkungsanalysen sichergestellt.

In Deutschland existiert keine dem oben erwähnten OMB vergleichbare Institution, die von zentraler Ebene aus einen so unmittelbaren Einfluss auf budgetäre Prozesse hat. Möglicherweise könnten in Deutschland aber das Bundeskanzleramt oder das Bundesfinanzministerium eine ähnliche Rolle als Agenda-Setter und für die Vorgabe von

Qualitätsstandards spielen. Der Bundesrechnungshof hingegen hat mit der Prüfung der Haushalts- und Wirtschaftsführung des Bundes andere Aufgaben.

Unabhängig von der konkreten Ansiedelung einer Institution, die eine Qualitätssicherung bei Evaluationen vorantreibt, ist eine gesetzliche Verankerung von Evaluationen für Politikmaßnahmen ab einer bestimmten Größe entscheidend, wobei die Bedingungen für eine aussagekräftige Evaluation gegeben sein müssen. In Ländern wie Schweden und der Schweiz hat Evaluation sogar Verfassungsrang. In Artikel 170 der Schweizer Verfassung heißt es „Die Bundesversammlung sorgt dafür, dass die Massnahmen des Bundes auf ihre Wirksamkeit überprüft werden.“ Zudem sollten alle neuen Programme mit einer sunset clause versehen werden, wie beispielsweise in Großbritannien üblich. Die Fortführung des Programms sollte dann vom Nachweis einer erfolgreichen Evaluierung abhängig gemacht werden. Und schließlich ist eine verbindliche Berichtspflicht über Qualitätssicherung notwendig.

6.2 Institutionelle Verankerung in den politischen Prozessen

Ein umfassendes Bekenntnis für Qualitätssicherung in der Evaluation sollte von einer konkreten institutionellen Verankerung in den einzelnen politischen Prozesses begleitet werden. Ein Chief Evaluation Office sollte, ähnlich wie im amerikanischen Department of Labor, an der Vergabe und Qualitätssicherung der Evaluationen beteiligt und für die Schulung der Mitarbeiter zuständig sein.

Erste Entwicklungen in diese Richtung sind derzeit im Bundesministerium für Wirtschaft und Energie zu beobachten, wo 2011 ein Aufbaustab Fördercontrolling/Evaluation eingerichtet wurde, der frühzeitig in die Ausschreibung und Vergabe von Evaluationen durch die Fachreferate und in die Berichtsabnahme eingebunden werden muss. Auch im Bundesministerium für Bildung und Forschung wird die Einrichtung einer „Kompetenzstelle Evaluation“ angeregt, die für grundsätzliche und strategische Fragen zum Thema Evaluation zuständig sein soll.

In der konkreten Umsetzung der Evaluationen sollten die diskretionären Spielräume der politischen Entscheidungsträger über die Durchführung und die Art einer Evaluation möglichst weit reduziert werden. Denn im politischen Prozess wird es sonst immer wieder Anreize geben, sich einer rigorosen Evaluation zu widersetzen. Eine Evaluation könnte die Maßnahme, für die eine bestimmte ministeriale Stelle verantwortlich ist, in Frage stellen. Durch eine Anpassung der internen Anreizsysteme dahingehend, dass nicht die Fortführung der Maßnahme per se relevant ist, sondern die Fortführung und die Entwicklung effektiver Maßnahmen, können zudem mögliche Widerstände in der Verwaltung reduziert werden.

Personalkapazitäten in den Ministerien können den Ausschlag dafür geben, welchen Stellenwert Wirkungskontrolle in den Ministerien hat und welche Agenda für Evaluationen gesetzt wird. Voraussetzung hierfür sind Rekrutierungsprozesse, die entsprechend

ausgebildete Hochschulabsolventen in die Ministerien bringen. Als hilfreich hat es sich in anderen Ländern erwiesen, wenn es einen engen personellen Austausch zwischen Wissenschaftlern, die über einen bestimmten Zeitraum hinweg in der Administration arbeiten, einerseits und der Politik andererseits gibt.

Schließlich sollten in den Ministerien oder ihren nachgeordneten Behörden Forschungsdatenzentren eingerichtet werden, um die für die Evaluationen notwendigen Informationen über die Fördermaßnahmen transparent bereit zu stellen und für Replikationsstudien zugänglich zu machen.

Neben diesen organisatorischen und personalpolitischen Erwägungen, die die einzelnen Ministerien betreffen, sind drei weitere Faktoren entscheidend. Erstens sollten, um die Kosten einer Evaluation möglichst gering zu halten, Evaluationen grundsätzlich gleichzeitig mit der Konzeption einer Maßnahme geplant werden. Zweitens sollte über eine Veröffentlichung der durchgeführten Evaluationen Transparenz hergestellt werden. Drittens sollten die verwendeten Daten – natürlich unter Einhaltung relevanter Anforderungen des Datenschutzes – auch weiteren nicht beteiligten Forschern zur Verfügung gestellt werden, um eine Replizierbarkeit der Studien zu gewährleisten.

6.3 Einbindung von Wissenschaft und Forschung

Nicht alle Ministerien und ausführenden Behörden werden in ausreichendem Umfang Personal vorhalten können, um Evaluationen selbst durchführen zu können. Und in der Tat stünde eine rein interne Evaluation im Widerspruch zu dem hier geforderten offenen Prozess wissenschaftlich fundierter Bewertungen wirtschaftspolitischer Maßnahmen. Daher ist die geeignete Einbindung von Wissenschaft und Forschung zentral. So können Infrastrukturen, die in Universitäten und Forschungsinstituten bereits vorgehalten werden, möglichst effektiv genutzt werden.

An der Schnittstelle zwischen Politik und Forschung ist eine wettbewerbliche Vergabe von Mitteln für Evaluationen entscheidend. Es sollte vermieden werden, dass einzelne Evaluatoren auf einen bestimmten Auftraggeber finanziell angewiesen sind und daher die erforderliche Distanz zu den zugrundeliegenden politischen Prozessen fehlt. Welche Bieter dann letztlich den Zuschlag für eine Evaluation bekommen, muss Ergebnis des Marktprozesses sein. In Deutschland gibt es mit den außeruniversitären Forschungsinstituten und den mit ihnen kooperierenden Hochschulen einen relativ großen Markt an möglichen Anbietern an Evaluationsdienstleistungen. In den USA sind die Marktstrukturen andere: hier sind es auch private Institute wie MDRC (Manpower Demonstration Research Corporation) und Mathematica, die sich eine entsprechende Reputation für wissenschaftlich fundierte Evaluationen aufgebaut haben.

Unabhängig davon, welche Institutionen letztlich mit einer Evaluation betraut werden, müssen die Prozesse so ausgestaltet sein, dass keine Gefälligkeitsgutachten produziert werden, die nur „beweisen“, was die Politik ohnehin schon wusste. Unabhängigkeit der Auftragnehmer ist daher ein zentrales Kriterium. In vielen europäischen Ländern sind die Zahl der relevanten Institute und das Ausschreibungsvolumen zu klein, als dass sich wirklich ein Markt etablieren könnte. Ein stärkerer internationaler Wettbewerb wäre deshalb vorteilhaft, wird aber durch Sprachprobleme erschwert. Eine Replizierbarkeit der Evaluation ist eine zentrale Möglichkeit, Gefälligkeitsgutachten zu vermeiden. Dafür muss, wie oben ausgeführt, die notwendige Datenbasis geschaffen und für Forscher geöffnet werden.

Eine engere Einbindung von Wissenschaft und Forschung bedeutet dabei nicht, dass wirtschaftspolitische Zielsetzungen unkritisch von der Wissenschaft übernommen werden. Wissenschaftliche Evaluation von politischen Maßnahmen ist vielmehr eine positive Analyse, die nicht bedeutet, die Ziele der Politik zu akzeptieren. Sie ergänzt, vermindert aber nicht die Notwendigkeit einer normativen Diskussion zwischen Wissenschaft und Politik über die zu Grunde liegenden politischen Ziele.

6.4 Verbesserung von Lehre und Forschung

Wissenschaftliche Methoden entwickeln sich ständig weiter, neue politische Ziele und Maßnahmen erfordern die Entwicklung neuer Evaluationsansätze. Um einen kontinuierlichen Dialog zwischen Forschung und Politik sowie die Weiterbildung der für Evaluationen zuständigen Mitarbeiter zu ermöglichen, ist die Schaffung eines gemeinsamen Forums sinnvoll. In den USA wird eine solche Plattform beispielsweise über die Association of Public Policy and Management (APPAM) bereitgestellt.

In der universitären Lehre kann die Vermittlung der notwendigen Methodenkenntnisse noch stärker als bisher in die Lehrpläne eingebunden werden; das Spektrum an einschlägigen Lehrbüchern ist groß (Angrist und Pischke 2009, Bauer et al. 2009, Stock und Watson 2006). In der akademischen Forschung ist ein Umdenken erforderlich, das die positive Rolle von Replikationsstudien betont (Stevensohn und Wolfers 2013), sowohl für die Beratung als auch als Element der Qualitätssicherung in der Wissenschaft selbst. Replikationsstudien könnten besser in die Ausbildung integriert und beispielsweise als Qualifizierungsarbeit vergeben werden.

7 Fazit

In den vergangenen Jahren sind in Deutschland erhebliche Fortschritte bei der Bereitstellung von Daten der amtlichen Statistik und insbesondere von Einzeldaten gemacht worden. Die neue Bundesregierung will diesen Trend mit einer Verbesserung der Nutzung von Daten der amtlichen Statistik und von „Big Data“ weiter beschleunigen. Generell wurden die

wissenschaftlichen Methoden, mit denen die Wirkung wirtschaftspolitischer Maßnahmen untersucht werden können, erheblich verbessert. An unseren Universitäten sind diese Methoden vielfach zum Standardrepertoire der empirischen Ausbildung geworden. Kurzum: grundsätzlich steht die nötige Infrastruktur bereit, mit deren Hilfe wir besseren Aufschluss darüber erlangen können, welche Maßnahmen wirken – und welche nicht. Gerade in Zeiten enger Konsolidierungsauflagen für die öffentlichen Haushalte ist eine bessere Überprüfung bestehender Ausgaben zentral.

Jedoch werden moderne Evaluationsverfahren in Deutschland bislang viel zu wenig genutzt. In diesem Beitrag zeigen wir, dass sowohl Politik als auch Wissenschaft dazu beitragen können, die vorhandenen Potenziale besser zu nutzen. Evaluationen müssen, anders als oft vermutet, nicht teuer sein. Ihr Ziel ist es ja letztlich, die vorhandenen Mittel effektiver und effizienter einzusetzen. Und: je früher Evaluationen bei der Umsetzung einer bestimmten Maßnahmen mit eingeplant werden. umso geringer sind die Kosten.

Auf Seiten der Politik können, wie das Beispiel der USA zeigt, eine klare Agenda für evidenzbasierte Forschung und einheitliche Standards für Qualitätssicherung in der Beratung wichtige Schritte sein. Dies gelingt umso besser, je stärker Wirkungsforschung institutionell verankert ist. Die Etablierung eines Chief Evaluation Office, der bessere Austausch zwischen Politik und Wissenschaft, und eine wettbewerbliche Vergabe von Evaluationsaufträgen könnten wichtige Beiträge leisten. Zudem sollten alle neuen Programme mit einer sunset clause versehen werden., Das heißt, dass nach einem vorab festgelegten Zeitraum geprüft werden sollte, ob die Programme erfolgreich waren. Und nicht zuletzt könnte über neu eingerichtete Forschungsdatenzentren in den Ministerien der Zugang zu relevanten Informationen verbessert werden.

Auf Seiten der Wissenschaft kann eine bessere Evaluationskultur vor allem dadurch etabliert werden, dass man Replikationsstudien, die eine kritische Überprüfung von Evaluationen ermöglichen, fördert. Vielfach scheitern zudem Evaluationen nicht an fehlenden wissenschaftlichen Methoden und Möglichkeiten, sondern an einem Austausch darüber, welche Methoden auf eine bestimmte Maßnahme passen. Hierzu ist ein enger Dialog zwischen Politik und Wissenschaft nötig, der beispielsweise durch regelmäßige Workshops und Weiterbildungsmaßnahmen gefördert werden kann.

Auf Seiten der Wähler und Steuerzahler schließlich sollte die Frage nach Evidenz für die Wirksamkeit politischer Maßnahmen viel nachdrücklicher gestellt werden. Die gesellschaftliche Diskussion über Themen wie Betreuungsgeld, Ganztagsschulen, Rente mit 63 wird, so hat man oft den Eindruck, vor allem unter dem Einfluss organisierter Interessen und nach den Regeln des politischen Kalküls ausgetragen, während wenig nach empirischen Befunden zu Wirksamkeit und Effizienz gefragt wird. Ein aufgeklärter gesellschaftlicher Dialog setzt jedoch voraus, dass man die Augen nicht vor den Fakten verschließt, sondern im

Gegenteil darauf bestehen, diese Fakten zu ermitteln und in die politische Entscheidungsfindung einzubringen. Wir täten gut daran, mehr Empirie zu wagen.

8 Literatur

Angrist, J. D. und J.-S. Pischke (2009). *Mostly Harmless Econometrics. An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. und J.-S. Pischke (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24 (2): 3-30.

Arni, P. (2012). Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis. *LeGes – Gesetzgebung und Evaluation* 23 (3): 355-386.

Banerjee, A., und E. Duflo (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs.

Bauer, T.K., M. Fertig und C.M. Schmidt (2009). *Empirische Wirtschaftsforschung: Eine Einführung*. Berlin: Springer.

Behagel, L., B. Crépon und M. Gurgand (2012). Private and Public Provision of Counseling to Job-Seekers: Evidence from a Large Controlled Experiment, IZA Discussion Paper No. 6518.

Bender, S., A. Bergemann, B. Fitzenberger, M. Lechner, R. Miquel, S. Speckesser, C. Wunsch (2005). *Über die Wirksamkeit von FuU-Maßnahmen, Ein Evaluationsversuch mit prozessproduzierten Daten aus dem IAB.* Beiträge zur Arbeitsmarkt- und Berufsforschung, 289, Nürnberg.

Bernhard, S., K. Hohmeyer, E. Jozwiak, S. Koch, T. Kruppe, G. Stephan, J. Wolff (2009). Aktive Arbeitsmarktpolitik in Deutschland und ihre Wirkungen, in: Institut für Arbeitsmarkt- und Berufsforschung (Hrsg.), Handbuch Arbeitsmarkt 2009, Bertelsmann, 149-201.

Biewen, M., B. Fitzenberger, A. Osikominu, R. Völter, M. Waller (2006). Beschäftigungseffekte ausgewählter Maßnahmen der beruflichen Weiterbildung in Deutschland: Eine Bestandsaufnahme. *Zeitschrift für Arbeitsmarktforschung*, 39 (3-4), 365-390.

Bräuninger, M., J. Michaelis und M. Sode (2013), 10 Jahre Hartz-Reformen, HWWI Policy Paper 73.

Bundesministerium für Arbeit und Soziales und Institut für Arbeitsmarkt- und Berufsforschung (2011). Sachstandsbericht der Evaluation der Instrumente. Berlin: Bundesministerium für Arbeit und Soziales.

Chatterji, A., E. Glaeser und W. Kerr (2013). Clusters of Entrepreneurship and Innovation. NBER Chapters, in: *Innovation Policy and the Economy*, Volume 14 National Bureau of Economic Research, Inc.

DiNardo, J.; D.S. Lee (2010). Program Evaluation and Research Designs. NBER Working Paper No. 16016.

Eichhorst, W. und K.F. Zimmermann (2007). Dann waren's nur noch vier… Wie viele (und welche) Maßnahmen der aktiven Arbeitsmarktpolitik brauchen wir noch? – Eine Bilanz nach der Evaluation der Hartz-Reformen, IZA Discussion Paper No. 2605.

Europäische Kommission (2013). Green Paper: Long-Term Financing of the European Economy. COM(2013) 150 final. Brüssel.

Fitzenberger, B.; S. Speckesser (2000). Zur wissenschaftlichen Evaluation der aktiven Arbeitsmarktpolitik in Deutschland: Ein Überblick. ZEW Discussion Papers No. 00-06.

Fritsch, M., und M. Wyrwich (2012). The Long Persistence of Regional Entrepreneurship Culture: Germany 1925-2005. Jena Economic Research Papers 2012-036, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

Greenberg, D., und M. Shroder (2004). *Digest of Social Experiments*, 3rd Edition. Washington, DC: Urban Institute Press.

Gutberlet, T. (2012). Cheap Coal, Market Access, and Industry Location in Germany 1846 to 1882. University of Arizona.

Hanushek, E., Lindseth, A. A., 2009. Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in America's Public Schools. Princeton University Press, Princeton, NJ.

Haskins, R., und J. Baron (2011). Building the Connection between Policy and Evidence – the Obama evidence-Based initiatives. NESTA.

Heckman, J.J., und J.A. Smith (1998). Evaluating the Welfare State. NBER Working Paper No. 6542.

Heckman, J.J.; S.H. Moon; R.Pinto; P.A. Savelyev; A. Yavitz (2009). The rate of return to the HighScope Perry Preschool Program. Journal of Public Economics, 94(1-2), 114-128.

Heyer, G. (2002). Rahmenzielsetzungen und Wirkungsforschung im Job-AQTIV-Gesetz. Vortrag auf dem Workshop Evaluation, Nürnberg, 9.11.2001. IAB-Werkstattbericht Nr. 2/2002, Nürnberg.

Heyer, G., S. Koch, G. Stephan, J. Wolff (2011). Evaluation der aktiven Arbeitsmarktpolitik, Ein Sachstandsbericht für die Instrumentenreform 2011. IAB-Discussion Paper 17/2011.

Jacobi, L. und J. Kluve, J. (2007): Before and After the Hartz Reforms: The Performance of Ac-tive Labour Market Policy in Germany, *Zeitschrift für Arbeitsmarktforschung* 40(1), S. 45–64.

Krug, G. und G. Stephan (2013): Is the Contracting-Out of Intensive Placement Services More Effective than Provision by the PES? Evidence from a Randomized Field Experiment. IZA Discussion Paper No. 7403.

Kugler, F., G. Schwerdt und L. Wößmann (2014), Ökonometrische Methoden zur Evaluierung kausaler Effekte der Wirtschaftspolitik, *Perspektiven der Wirtschaftspolitik*, in dieser Ausgabe.

Lechner, M. und C. Wunsch (2013). Sensitivity of matching-based program evaluations to the availability of control variables, *Labour Economics*, 21(C), 111-121.

Manski, Charles F. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Moffit, R.A. (2004). The Role of Randomized Field Trials in Social Science Research. *American Behavioral Scientist*, 47 (5), 506-540.

Nationaler Normenkontrollrat (2013). Kostentransparenz verbessert – Entlastung forcieren. Jahresbericht 2013. Juli. Berlin.

Rat für Sozial- und Wirtschaftsdaten (2010). Kriterien des Rates für Sozial- und Wirtschaftsdaten (RatSWD) für die Einrichtung der Forschungsdaten-Infrastruktur. http://ratswd.de/dl/doc/RatSWD_FDZKriterien_0.PDF (30.6.2013).

Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung (2013). Jahresgutachten 2013/14 "Gegen eine rückwärtsgewandte Wirtschaftspolitik". Wiesbaden.

Schiel, S., R. Cramer, R. Gilberg, D. Hess, H. Schröder (2006). Evaluation des arbeitsmarktpolitischen Programms FAIR. Stand der Begleitforschung zum Ende der Programmlaufzeit. IAB Forschungsbericht Nr. 7/2006, Nürnberg.

Schmidt, C.M. (2009). Wirtschaftswissenschaft und Politikberatung in Deutschland – Bedeutung, Möglichkeiten und Grenzen der Kausalanalyse. In: *Wirtschaftspolitik im Zeichen europäischer Integration: Festschrift für Wim Kösters anlässlich seines 65. Geburtstages*, hrsg. von A. Belke, H.-H. Kotz, S. Paul und C.M. Schmidt. Berlin: Duncker & Humblot.

Schweinhart, L.J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2004). The High/Scope Perry Preschool Study Through Age 40: Summary, Conclusions, and Frequently Asked Questions. Ypsilanti, MI: High/Scope Press.

Smith, J. (2009). What Can the ESF Learn from US Evaluations of Active Labor Market Programs? Department of Economics. University of Michigan.

Stock, J.H. und M.W. Watson (2006). Introduction to Econometrics, second edition. Addison Wesley.

The Economist (19.10.2013). How science goes wrong. London.

Vikström, J., M. Rosholm und M. Svarer (2013). The Relative Efficiency of Active Labour Market Policies: Evidence from a Social Experiment and Non-Parametric Methods. Labour Economics, 24, 58-67.

Wissenschaftlicher Beirat des BMWi (2013). Evaluierung wirtschaftspolitischer Fördermaßnahmen als Element einer evidenzbasierten Wirtschaftspolitik. Berlin.

Wissenschaftsrat (2012). Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Berlin.

Wolfers, J. und B. Stevenson (2013). Six Ways to Separate Lies From Statistics. http://users.nber.org/~jwolfers/popular.php, 1. Mai 2013.