ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Arni, Patrick

Working Paper

Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis

IZA Standpunkte, No. 52

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Arni, Patrick (2012) : Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis, IZA Standpunkte, No. 52, Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/91813

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

IZA

STANDPUNKTE

IZA Standpunkte Nr. 52

Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis

Patrick Arni

November 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis

Patrick Arni
IZA
DEEP, Universität Lausanne
CAFE, Aarhus University

IZA Standpunkte Nr. 52
November 2012

IZA

Postfach 7240
53072 Bonn

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

Die Schriftenreihe „IZA Standpunkte“ veröffentlicht politikrelevante Forschungsarbeiten und Diskussionsbeiträge von Wissenschaftlern und Netzwerkmitgliedern des IZA in deutscher Sprache. Für den Inhalt der publizierten Arbeiten sind allein die Autoren verantwortlich.

Das IZA ist ein privates, unabhängiges Wirtschaftsforschungsinstitut und betreibt nationale wie auch internationale Arbeitsmarktforschung. Als gemeinnützige GmbH wird es maßgeblich durch die Deutsche Post-Stiftung gefördert. Zu den zentralen Aufgaben des Instituts zählen die originäre und international wettbewerbsfähige Forschungstätigkeit auf allen Gebieten der Arbeitsökonomie, die Umsetzung wissenschaftlicher Erkenntnisse in praktikable wirtschaftspolitische Konzepte sowie die Vermittlung von Forschungsergebnissen und Konzepten an die interessierte Öffentlichkeit.

Im Interesse einer einheitlichen Textzirkulation werden Aktualisierungen einmal publizierter Arbeiten nicht an dieser Stelle vorgenommen, sondern sind gegebenenfalls nur über die Autoren selbst erhältlich.

IZA Standpunkte Nr. 52
November 2012

ZUSAMMENFASSUNG

Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis*

Wirkungsevaluationen stehen oft vor der Herausforderung, Kausalität zwischen der betrachteten neuen Politikmassnahme und den resultierenden Outcomes herzustellen. Mangelnde Vergleichbarkeit zwischen der Programmgruppe (neue Politik) und der Kontrollgruppe (Status Quo) macht oft eine kausale Interpretation der gefundenen Effekte schwierig (sind wirklich Programmeffekte oder eher Selektionseffekte für das Ergebnis verantwortlich?). Randomisierung – d.h. Zufallszuweisung in Programm- und Kontrollgruppe – sorgt für eine sehr hohe Vergleichbarkeit. In diesem Beitrag werden die Möglichkeiten der Nutzung von randomisierten Studien in der Evaluation von Pilotprojekten diskutiert. Erstens werden Gründe, die für Randomisierung sprechen, sowie Einschränkungen in der Anwendung der Methode diskutiert. Im zweiten Teil des Beitrages wird aufgezeigt, wo aktuell randomisierte Evaluationsstudien in Europa bereits eingesetzt werden – diese Beispiele demonstrieren die Einsatzmöglichkeiten und das Potenzial der Methode. Drittens wird auf die Praxis der Umsetzung und Planung von randomisierten Studien eingegangen: Eine Reihe von zentralen Punkten wird diskutiert, die bei der Implementierung von solchen kausalen Evaluationen im Auge behalten werden sollten.

JEL-Codes: C93, C81, H43, I38

Schlagworte: Kausale Evaluation, Randomisierung, Feldexperiment, Kontrollgruppe, Pilotprojekt

Kontaktadresse:

Patrick Arni
IZA
Postfach 7240
D-53072 Bonn
E-Mail: arni@iza.org

* Erscheint in: LeGes – Gesetzgebung und Evaluation, 2012/3.

Inhalt

1. Einleitung 2
2 Weshalb randomisieren? 4
2.1 Die Wichtigkeit der Kontrollgruppe 4
2.2 Bessere Vergleichbarkeit durch Randomisierung 5
2.3 Notwendige Rahmenbedingungen für Randomisierung 7
3 Was randomisieren? Aktuelle Beispiele aus Europa 9
3.1 Randomisierte Zuweisung in Programme 9
3.2 Randomisierung auf zwei Ebenen 10
3.3 Zufallszuteilung zu Anbietern 11
3.4 Randomisierung zusätzlich angebotener Information 12
3.5 Anwendungen in den Sozialversicherungen 13
3.6 Anwendungen im Bildungs- und Umweltbereich 15
4. Wie randomisieren? Planung und Praxis 17
5. Schlussbemerkungen 22
Literatur 23
Endnoten 25

1. Einleitung

Für die Auftraggebenden von Evaluationen steht häufig die Frage im Zentrum, ob die zu evaluierenden neuen Politikmassnahmen eine *kausale* Wirkung auf die interessierenden Zielgrössen (Outcomes) haben. Ist es effektiv die neue Massnahme für Sozialhilfebeziehende, die deren Erwerbschancen verbessert hat – oder waren es gewisse unbeobachtete Eigenschaften der Teilnehmenden an der Massnahme, die die gemessene Wirkung hervorgerufen haben? Ist es wirklich das neue Informationskonzept, das die Bürgerinnen und Bürger zu mehr stromsparendem Verhalten geführt hat – oder ist der eruierte Effekt einfach dadurch entstanden, dass die Regionen mit und ohne neuem Informationskonzept nicht wirklich vergleichbar waren? Für die Gestaltung von Politik ist es von zentraler Wichtigkeit, dass die Evaluation soweit wie möglich den kausalen Effekt einer Massnahme von den *Selektionseffekten* trennen kann. Letztere entstehen dadurch, wenn der Vergleich, der zur Wirkungsmessung herangezogen wird – Individuen mit neuer Massnahme vs. Individuen ohne neue Massnahme – verzerrt ist: Wenn sich die beiden verglichenen Gruppen an Individuen in beobachtbaren oder unbeobachtbaren Eigenschaften oder Voraussetzungen des Umfelds unterscheiden.

Dieser Beitrag stellt das Instrument der zufälligen Programmzuweisung – der *Randomisierung* – als Mittel für die Gestaltung von Evaluationen vor. Randomisierung ist, aus methodischer Sicht, die beste Möglichkeit, die *kausale Wirkung einer Massnahme zu identifizieren* und sie von Selektionseffekten zu trennen. Zufällige Programmzuweisung kommt in der Schweiz und in Europa bei der Evaluation von sozialpolitischen, arbeitsmarktlichen, umweltpolitischen o.ä. Programmen erst sehr zögerlich zum Einsatz – in der Tendenz jedoch klar steigend. Es lohnt sich, dieses Verfahren der kausalen Programm-Evaluation genauer zu beleuchten und zu diskutieren. Besonders auch deshalb, weil es für die Auftraggebenden bzw. die Politikgestaltenden einen Mehrwert bietet bezüglich der *Klarheit* der Interpretation der Ergebnisse (siehe Kapitel 2).

Solcherlei Vorteile sprechen dafür, ein *randomisiertes Projektdesign bei der Ausgestaltung von Pilotprojekten* in Betracht zu ziehen. Die Anwendung solcher Designs ist eigentlich keine Neuheit: Im Bereich der Medizin gehören sie schon seit Langem zum Alltag der Wirkungsmessung. Für die Evaluation der Wirkung neuer Medikamente oder Behandlungsweisen ist die zufallsgesteuerte Zuweisung der Probandinnen und Probanden in eine „Treatment-Gruppe“, der die neue Behandlung verordnet wird, und eine „Kontrollgruppe“, die nach bisheriger Behandlungsweise betreut wird, ein Standard-Verfahren. Auch für viele psychologische oder neurologische Fragestellungen gehören randomisierte Ansätze zum üblichen Methodenwerkzeug. Die Anwendung dieser Methoden auf die Einführung öffentlicher Politiken ist hingegen noch nicht so alltäglich.

Die methodischen Grundlagen für solche Anwendungen haben sich seit den 70er-Jahren massgeblich entwickelt: Der hier benutzte Begriff der *kausalen Programm-Evaluation* weist auf die breite statistische und ökonometrische Literatur hin, die unter den Stichworten „causal analysis“ (z.B. Angrist et al. 2009) und „program evaluation“ (z.B. DiNardo et al. 2011) in den letzten Jahrzehnten entstanden ist. Aufbauend auf dem statistischen Modell der Kausalität und der „potenziellen Outcomes“[1], das hauptsächlich auf Rubin (1974, 1977) zurückgeht, hat sich in dieser Literatur ein solider methodischer Rahmen entwickelt für die Anwendung randomisierter Designs in der Politikevaluation.

Dieser Rahmen wurde hauptsächlich anhand von Fragestellungen aus der Arbeitsmarktökonomie (z.B. List et al. 2011) entwickelt, ist aber grundsätzlich auf jeden Politikbereich anwendbar[2]. Voraussetzung ist eine klar umreissbare politische Massnahme (Intervention) oder ein Programm, dessen Wirkung gemessen werden soll. Der thematische Anwendungsbereich randomisierter Evaluationsstudien wurde in Europa in den letzten Jahren breiter: arbeitsmarktpolitische Programme, sozialpolitische Interventionen, Bildungsmassnahmen, Fragen der Diskriminierung, umweltpolitische Interventionen etc. (siehe aktuelle Beispiele in Kapitel 3). Allerdings ist die Zahl der umgesetzten Studien noch tief. Ausserhalb Europas ist die Nutzung von randomisierten Designs – oder synonym: von Feldexperimenten[3] – inzwischen in der Entwicklungsökonomie schon sehr verbreitet (siehe z.B. Duflo et al. 2007). Ebenso hat das Mittel der randomisierten Studie in Nordamerika auch bereits in nicht-akademischen Evaluations-Institutionen regelmässige Anwender gefunden[4].

Der in diesem Beitrag diskutierte Ansatz der Politikevaluation mittels Zufallszuweisung findet meist im Kontext von *Pilot- oder Modellprojekten* statt, wie die Beispiele in Kapitel 3 aufzeigen. Der Gründe hierfür sind praktischer und gesetzlicher Natur: Die zu evaluierenden Massnahmen, für die Randomisierung machbar ist, sind üblicherweise neue Programme oder Änderungen der Politikpraxis. Da deren Wirkung meist nicht schon im Vornherein klar ist – deshalb auch der Bedarf an Evaluation – macht deren schrittweise Einführung als Pilotprojekt Sinn. Der temporäre Test einer neuen Massnahme in einer Teilgruppe der potenziell von der Politik Betroffenen erlaubt eine Wirkungsmessung *vor* dem definitiven Entscheid zur Einführung oder Nichteinführung. Sollte die Evaluation negative Resultate zeigen, kann immer noch von einer definitiven Einführung abgesehen werden. Des Weitern ist der Status des Pilotprojekts in diversen Politikbereichen explizit gesetzlich vorgesehen, um den Test innovativer Massnahmen, die anfänglich noch nicht allen zuteil werden, zu ermöglichen (siehe Kapitel 2). Schliesslich ist aus methodischer Sicht zentral, dass Pilotprojekte erlauben, eine Teilgruppe an Personen mit der neuen Massnahme zu konfrontieren und eine andere nicht. Ein solcher Programm-Kontrollgruppen-Ansatz ist eine Grundvoraussetzung, um eine kausale Analyse der Wirkung der Massnahme zu ermöglichen. Dies wird in Kapitel 2 näher ausgeführt. Insofern sind Pilotprojekte also für randomisierte Studien besonders geeignet [5].

Dieser Beitrag fokussiert auf die Fragen der Praxis-Anwendung von randomisierten Evaluationen. Konkret diskutiert einige Aspekte zu den folgenden Fragen:

- *Weshalb randomisieren?:* Was spricht für das Mittel der zufälligen Programmzuweisung? Wo liegen die Begrenzungen und Problematiken?
- *Was randomisieren?*: Anhand von aktuellen Beispielen aus Europa soll sichtbar gemacht werden, wo zufällige Programmzuweisung bereits in einzelnen Fällen eingesetzt wird. Dies zeigt aus der Praxis-Sicht, welche Art von Interventionen mit randomisierten Designs evaluiert werden können.
- *Wie randomisieren?*: Es werden einige wichtige Punkte und Empfehlungen diskutiert, die in der Umsetzung von zufälligen Programmzuweisungen berücksichtigt werden sollten.

Diese drei Fragen werden nacheinander in den nächsten drei Kapiteln diskutiert. Kapitel 5 schliesst mit einer kurzen Konklusion ab.

2 Weshalb randomisieren?

2.1 Die Wichtigkeit der Kontrollgruppe

Wirkungsevaluationen stehen oft vor der Herausforderung, Kausalität zwischen der betrachteten Intervention und den resultierenden Outcomes herzustellen. Die grundlegende Voraussetzung hierfür, damit dies (annäherungsweise) möglich wird, ist das Vorhandensein einer *Kontrollgruppe*. Dies ist eine Personengruppe, in der die neue Intervention *nicht* zur Verfügung stand (Behandlung nach dem „Status Quo"). Durch den Vergleich dieser Kontrollgruppe mit der Personengruppe, in der die neue Massnahme Anwendung fand (Programm- oder Treatment-Gruppe), kann die Wirkung der neuen Intervention gemessen werden. Wird diese Kontrollgruppe nicht bereits bei der Gestaltung des zu evaluierenden Projekts mitgeplant, entsteht häufig das Problem, dass diese später (z.B. bei einer ex-post Evaluation) nicht so einfach geschaffen werden kann: Eine natürliche Kontrollgruppe, die per se schon gut vergleichbar wäre, existiert meist nicht.

Eine nutzbare Kontrollgruppe muss dann über umständliche Verfahren erst definiert und konstruiert werden. Wenn hierfür nicht sehr umfangreiche Daten zur Verfügung stehen, ist die Vergleichbarkeit zwischen der Programmgruppe und der Kontrollgruppe trotzdem meist recht mangelhaft: Die (konstruierten) Gruppen können sich in *unbeobachteten* – und daher statistisch nicht kontrollierbaren – Eigenschaften der Individuen unterscheiden. Dies kommt daher, dass der Zugangsprozess und die Eigenschaften der im Nachhinein ausgewählten Kontrollgruppe nicht vollständig beobachtet und kontrolliert werden konnte. Werden beispielsweise Personen im Jahr vor der Einführung der neuen Massnahme als Kontrollgruppe herangezogen, ist es gut möglich, dass diese Gruppe andere Eigenschaften ausweist, weil sie anderen gesellschaftlichen und politischen Bedingungen ausgesetzt war. Wenn also der *Selektions*prozess in die (konstruierte) Kontrollgruppe und in die Programmgruppe nicht beobachtbar der gleiche ist, bleiben Vergleichsprobleme bestehen: Inwieweit geht die gemessene Wirkung effektiv auf *Programmeffekte* zurück, inwieweit auf nicht beobachtete *Selektionseffekte*? Solange diese Trennung unklar ist, ist eine kausale Interpretation der gefundenen Wirkungen schwierig.

In solchen Fällen, wo der Zuweisungsprozess zu Treatment- und Kontrollgruppe durch die Durchführenden des Pilotprojekts bzw. die Evaluierenden nicht kontrollierbar war, ist man darauf angewiesen, im Nachhinein mittels statistischer Modelle Selektionskorrekturen vorzunehmen (siehe z.B. Angrist et al. 2009 oder Blundell et al. 2009). Damit sind massgebliche Kosten verbunden: Zum Einen müssen oft zusätzliche Datenquellen gesucht werden. Sind diese Vergleichsdaten dann vorhanden, wird, zum Andern, meist ein hoher statistischer Preis bezahlt: Die Modelle, die Selektionskorrekturen erlauben, verlangen hohe Fallzahlen, um signifikante Resultate eruieren zu können. Bei kleineren Pilotprojekten entsteht dann oft die unglückliche Situation, dass keine statistisch abgesicherte Aussage mehr über den Programmeffekt gemacht werden kann, weil die gefundenen Effekte im statistischen Unsicherheitsbereich (grosse Irrtumswahrscheinlichkeit bzw. Konfidenzintervalle) liegen. Zudem müssen bei diesen Modellen mehr Annahmen getroffen und Parameter fixiert werden, die alle mit dem Risiko behaftet sind, dass sie potenziell nicht stimmen (Schneider et al. 2011). Die Modellannahmen werden je glaubwürdiger desto mehr Daten vorhanden sind – was wiederum mehr Kosten und Aufwand bedeutet.

Das Fazit aus dieser Betrachtung ist, dass viele Unsicherheiten in der Wirkungsmessung ausgeräumt werden können, wenn *bereits in der Gestaltung der zu evaluierenden Politikintervention die*

Kontrollgruppe mitgeplant wird. Bei Pilotprojekten lässt sich die Kontrollgruppe besonders gut mitplanen, da dort nicht alle betroffenen Personen sofort und flächendeckend mit der neuen Massnahme konfrontiert sind. Die Wichtigkeit der Planung einer gut vergleichbaren Kontrollgruppe ist jedoch generell, auch jenseits von Pilotversuchen, von zentraler Bedeutung für die kausale Evaluation. Die politischen Institutionen als Auftraggeber der Evaluation sollten also bei der Einführung von Pilotprojekten oder Politikreformen optimalerweise eine Vorgehensweise wählen, die es erlaubt, eine möglichst glaubwürdige – d.h. vergleichbare – Kontrollgruppe zu bilden. Ohne solche Kontrollgruppe lässt sich die zentrale Frage „wie hat die Intervention die Outcomes im Vergleich zum Status Quo verändert?“ nicht empirisch messen und beantworten. Wenn also eine empirische Messung der Programmwirkung politisch gewünscht – oder gesetzlich verlangt[6] – ist, muss *das Evaluationsdesign bereits Teil der Politikeinführung* sein.

2.2 Bessere Vergleichbarkeit durch Randomisierung

Welches sind nun konkret Designs von Kontrollgruppen-Ansätzen, die den Anspruch guter Vergleichbarkeit am besten erfüllen? Hierzu liefert die statistisch-ökonometrische Literatur zur Programmevaluation eine klare Antwort (siehe z.B. Angrist et al. 2009, Card et al. 2011, Imbens et al. 2008, List et al. 2011). Die optimalste bzw. glaubwürdigste Vergleichbarkeit liefert die *Zufallszuteilung in Programm- und Kontrollgruppe*: Durch diesen Prozess der Randomisierung wird sichergestellt, dass sich sowohl die beobachtbaren wie die unbeobachtbaren Eigenschaften der Individuen gleichmässig auf die beiden Gruppen verteilen. Das fundamentale Evaluationsproblem – dass man nie beide möglichen Ergebniszustände (Outcome mit Intervention und Outcome ohne Intervention) bei demselben Individuum gleichzeitig beobachtet – wird damit am besten gelöst, wie man anhand des „kausalen Modells“ von Rubin (1974, 1977) statistisch zeigen kann. Durch Zufallszuteilung in Programm- und Kontrollgruppe kann die *kausale Programmwirkung* am besten von verzerrenden Selektionseffekten getrennt werden[7].

Als zweitbeste Option für kontrollgruppen-basierte Politikevaluation erwähnt die Literatur (dito) die *quasi-experimentellen* Ansätze. Hierbei werden Situationen ausgenutzt, in denen Politikreformen unterschiedliche – aber vergleichbare – Gruppen schaffen von Betroffenen und weniger (oder nicht) Betroffenen; oder auch Situationen, in denen Reformen schrittweise eingeführt werden: Eine Altersgruppe ist, beispielsweise, betroffen von der neuen Politik, eine andere nicht. Eine Region hat ein neues Programm früher eingeführt, eine andere später. Durch die Einführung eines gewissen Stichtages (z.B. bei der Einschulung) wurden gewisse Gruppen zu einem gewissen Zeitpunkt von der neuen Politik betroffen, andere nicht. Viele weitere Beispiele sind hierzu denkbar. Die Grundannahme ist, dass die Festlegung, wo solche *Diskontinuitäten* der Einführung von Reformen auftreten, „administrativer“ oder gar zufälliger Natur war und somit unabhängig von der Wirkung der Politik ist (Exogenität). So ist beispielsweise die Festlegung der Altersgrenze, bis zu welcher gewisse Massnahmen angewandt werden, oft recht arbiträr und hat keinen direkten Bezug zur Auswirkung der Massnahme. Eine 54-jährige und eine 56-jährige Person würden wohl nicht fundamental anders auf die politische Massnahme reagieren (bei einer Altersgrenze von 55 Jahren, als Beispiel, für die Anwendung der Massnahme). Aufgrund solcher Diskontinuitäten können Aufteilungen in Programm- und Kontrollgruppen vorgenommen werden, die vergleichsweise glaubwürdig sind. Allerdings sind damit schon einige Annahmen verbunden.

Das prominenteste Modell, das für die Ausnutzung solcher diskontinuierlicher Situationen angewandt wird, ist der *Difference-in-differences* Ansatz: Hierbei werden die von der neuen Massnahme potenziell Betroffenen (z.B. jene über einem gewissen Alter) in die Programmgruppe eingeteilt und die anderen in die Kontrollgruppe; zudem wird „vor der Reform" und „nach der Reform" unterschieden. Um den Programmeffekt zu messen, wird nun die Doppel-Differenz im Outcome genommen: Es wird die Programm-Kontrollgruppen-Differenz nach der Reform von jener vor der Reform subtrahiert. Dies bedingt, neben der Exogenitäts-Annahme, eine weitere zentrale Annahme: Jene, dass keine „Trendbrüche" zwischen den beiden Gruppen auftreten (Parallelität der Trends). D.h. konkret: Wäre die Reform nicht aufgetreten, hätte sich der Unterschied im Verhalten zwischen der Programm- und der Kontrollgruppe in etwa gleichmässig weiterentwickelt. Wenn diese Annahme erfüllt ist, kann die Differenz zwischen Programm- und Kontrollgruppe von vor der Reform als Mass für den systematischen Unterschied zwischen den beiden Gruppen (Selektionseffekt) gesehen werden. Wird dieser Unterschied von der Programm-Kontrollgruppen-Differenz nach der Reform subtrahiert, bleibt noch jener Teil der Differenz übrig, der wahrscheinlich auf die Wirkung der Reform zurückzuführen ist.

Auch für diese quasi-experimentellen Ansätze – die dort sinnvoll sind, wo Randomisierung nicht möglich oder machbar ist – gilt dasselbe Fazit wie bereits oben erwähnt: Die Politikgestalterinnen und -gestalter können eine möglichst glaubwürdige Wirkungsmessung fördern, indem sie bei der Politikeinführung die Planung der Kontrollgruppe mitberücksichtigen. Dies wäre z.B. möglich, indem in der Politikeinführung eine Pilotphase zugelassen wird, wo noch nicht alle Personen an allen Orten zum selben Zeitpunkt mit der neuen Politik konfrontiert sind.

Diese Diskussion und Entwicklung von glaubwürdigen Ansätzen kausaler Evaluation von (Pilot-)Programmen beschränken sich nicht nur auf die methodische Literatur. In den USA fanden sie Eingang in die Gesetzgebung, konkret in die gesetzliche Regelung der Entwicklung und Evaluation des Bildungssystems: Seit 2002 gilt dort der „Education Sciences Reform Act"[8], der „scientifically valid education evaluation" u.a. als „an evaluation that employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible" definiert (Sec. 102, Abschnitt 19 D). Das daraufhin gegründete Institute of Education Sciences hat aus dieser gesetzlichen Grundlage einen Praxis-Guide (U.S. Department of Education 2003) für „education practitioners" entwickelt, der eine explizite Anleitung liefert, „how to evaluate whether an educational intervention is supported by rigorous evidence". Dort wird randomisierten Studien die höchste Qualität für die Generierung von „strong evidence" zugebilligt. Als nächste Qualitätsstufe („possible evidence") werden Studien betrachtet, die mit einem Programm-Kontrollgruppen-Design möglichst nahe an die Vergleichbarkeit von randomisierten Studien herankommen. Diese gesetzlichen und praktischen Vorgaben sind auch relevant für die Finanzierung von Bildungsevaluationen in den USA.

Die Nutzung eines randomisierten Designs ist dann besonders sinnvoll und naheliegend, wenn eine neue politische Massnahme zuerst als *Pilotprojekt* eingeführt wird. Der eigentliche Zweck eines Pilotprojektes besteht ja darin, eine Massnahme in einer beschränkten Population oder einem beschränkten Raum für eine gewisse Zeit auszutesten und zu evaluieren. Somit stellt sich automatisch die Frage, auf wen die neue Massnahme pilotweise angewandt wird und auf wen nicht. Die Antwort auf diese Frage kann die Zufallszuteilung sein. Eine solche kann zwei Problematiken lösen oder zumindest reduzieren: Erstens ist die randomisierte Zuteilung neutral, d.h. die Zuweisung in die Pilotmassnahme ist nicht interessengebunden oder unausgeglichen. Zweitens führt die

Zufallszuweisung zu höchst vergleichbaren Programm- und Kontrollgruppen, womit eine Grundvoraussetzung für eine Evaluation hoher Qualität und Glaubwürdigkeit gegeben ist.

Die Nutzung von Zufallszuweisung in Programm- und Kontrollgruppe, z.B. im Rahmen eines Pilotprojekts, hat noch einen weiteren Vorteil: Durch die sehr hohe Vergleichbarkeit der beiden Gruppen können deutlich *einfachere statistische Verfahren* angewandt werden, um die Wirkung der Massnahme zu eruieren. Der Totaleffekt könnte zum Beispiel durch einen einfachen Vergleich der Mittelwerte von Programm- und Kontrollgruppe statistisch sauber dargestellt werden. Solche einfacheren statistischen Methoden beinhalten ein deutlich kleineres Risiko von Schätzungenauigkeiten, im Vergleich zu komplexeren Methoden der Selektionskorrektur, die in nicht-randomisierten Situationen angewandt werden müssen (siehe weiter oben). Dies bedeutet, dass im randomisierten Fall bereits mit kleineren Populationen im Pilotprojekt statistisch signifikante Aussagen zu den Effekten der Massnahme gemacht werden können[9], als dies in der nicht-randomisierten Situation möglich ist. Der Einsatz von Zufallszuteilung kann also den Aufwand bzw. die Dimensionierung des Pilotprojekts tendenziell kleiner halten.

2.3 Notwendige Rahmenbedingungen für Randomisierung

Wo liegen die Einschränkungen in der Nutzung der Randomisierung in der Programmevaluation? Oft wird als ethisches Argument gegen die Zufallszuteilung die Ungleichbehandlung ins Feld geführt: Dass gewisse Personen damit von der neuen Massnahme profitierten und gewisse nicht. Dieses Argument ist für einen sorgfältig aufgesetzten randomisierten Pilotversuch so nicht stichhaltig – aus dreierlei Gründen: Erstens ist die Wirksamkeit und die effektive Wirkung der Massnahme *a priori* normalerweise *nicht* klar – deshalb besteht ja der Bedarf nach einem Pilotprojekt und einer Evaluation. Es kann deshalb a priori meist nicht gesagt werden, wer von der Massnahme profitiert und wer nicht. Zweitens hat die Situation, dass verschiedene Gruppen ungleich behandelt werden, nicht mit der Zufallszuteilung zu tun, sondern mit der Tatsache, dass ein Pilotprojekt lanciert wurde. Der Lancierung des Pilotprojekts geht ein politisch bewusster Entscheid dafür voraus, der hoffentlich auch die Grundsatzfrage, ob die neue Massnahme vertretbar ist, beinhaltet hat. Die Zufallszuteilung kann – umgekehrt – ein Argument sein, dass die Entscheidung, wo und für wen die neue Massnahme in der Pilotphase eingesetzt wird, nicht interessengesteuert ist, sondern eben zufallsgesteuert.

Drittens wurde der Status des Pilot- oder Modellprojekts gerade dafür geschaffen, dass eine neue, ergebnisoffene Massnahme zuerst getestet werden kann – ohne sie direkt breitflächig einzusetzen. Ein solcher Status ist beispielsweise für die Schweizer Arbeitslosenversicherung und die Invalidenversicherung extra gesetzlich vorgesehen: Die entsprechenden Autoritäten können unter gewissen Bedingungen „zeitlich befristete, vom Gesetz abweichende Pilotversuche zulassen“ (Art. 75a Abs. 1 Arbeitslosenversicherungsgesetz (AVIG), SR 837.0) bzw. „zum Zweck der Eingliederung befristete Pilotversuche bewilligen, die von den Bestimmungen dieses Gesetzes abweichen können“ (Art. 68quater Abs. 1 Bundesgesetz über die Invalidenversicherung (IVG), SR 831.20). Ein solcher temporärer Pilotprojektstatus ist aus ethischer wie auch politikgestalterischer Sicht sinnvoll: Er lässt einerseits das Ausprobieren und Evaluieren von Innovationen zu. Andererseits kann eine flächendeckende Direkteinführung einer ergebnisoffenen Massnahme ohne Evaluation potenziell auch flächendeckenden Schaden herbeiführen – wenn sich herausstellen sollte, dass die Massnahme

weitgehend negative, ungewünschte Auswirkungen zeigt[10]. Fundiert evaluierte Pilotprojekte können dies verhindern.

Die einschränkenden Rahmenbedingungen, unter denen eine Programm-Evaluation mittels Randomisierung sinnvoll sein kann, lassen sich u.a. mit folgenden Punkten umreissen: Erstens muss eine *klar abgrenzbare Intervention* definierbar sein, die mittels Randomisierung untersucht werden kann. Dies bedeutet einerseits, dass die Situation, wo die neue Massnahme eingesetzt bzw. zugewiesen wird, kontrollierbar sein muss. Nur dann lässt sich die Zuweisung auch randomisieren. Dies ist in Sozialversicherungen, sozialen und gesundheitlichen Institutionen oder dem Service Public häufig gegeben, aber nicht immer. Das nächste Kapitel diskutiert eine Serie von Beispielen randomisierbarer Interventionen. Andererseits bedeutet dies, dass komplexere Programme mit sehr vielen Teilaspekten nicht vollständig durch ein randomisiertes Design evaluiert werden können (z.B. Widmer et al. 2012). Denkbar ist in solchen Fällen, dass zentrale Elemente durch Randomisierung getestet werden und weitere Aspekte der Auswirkungen der neuen Politik mittels anderer geeigneter Methoden und Daten untersucht werden. Dies ist natürlich abhängig davon, wo auftraggeberseitig der Schwerpunkt des Erkenntnisinteresses liegt.

Zweite zu diskutierende Bedingung ist die Frage von *Aufwand und vorhandenen Ressourcen*. Randomisierte Evaluationen können zwar bei gleicher Grösse (oder Kleinheit) des Projekts präzisere Wirkungsmessungen liefern (siehe weiter oben), es entsteht aber in der Umsetzung ein gewisser Mehraufwand. Beispielsweise in der Organisation: Es braucht (Kontroll-)Massnahmen, die sicherstellen, dass die Durchführung der Zufallszuteilung und der Gestaltung von Programm- und Kontrollgruppen wie geplant vonstatten geht. Auch hier ist wiederum der Fokus des Erkenntnisinteresses der auftraggebenden Institution entscheidend, insbesondere wenn die Ressourcen für die Evaluation knapp bemessen sind und sehr viele Aspekte gleichzeitig untersucht werden sollen. Je nach Gewicht der einzelnen Evaluationsziele kann es mehr oder weniger Sinn machen, in die Umsetzung einer Randomisierung zu investieren oder nicht. Ein Vergleich von Kosten und Nutzen einzelner Elemente der geplanten Evaluation kann hier sicher hilfreich sein. Eine möglichst einfach implementierte und gut in bestehende Prozesse integrierte Form der Randomisierung generiert vergleichsweise wenig Zusatzkosten (siehe Kapitel 4).

Drittens sind Interventionen, bei denen *a priori klar ist, dass sie eine negative Wirkung aufweisen*, nicht geeignet für Randomisierung. So kann etwa eine Reform, die die Verstärkung der Sanktionen in der Arbeitslosen- oder Invalidenversicherung vorsieht, nicht mittels eines randomisierten Designs evaluiert werden. (Ein quasi-experimentelles Design wäre hingegen möglich, sofern die erwähnten Diskontinuitäten im Politikdesign vorhanden sind.). Viertens muss eine gewisse *Grundakzeptanz der neuen Massnahme* vorhanden sein – sowohl seitens der Teilnehmenden wie seitens der Zuweisenden. Ist die Akzeptanz zu tief, wird das Problem der Nichteinhaltung der Teilnahme- oder Zuweisungsregeln („non-compliance“) zu gross und die Randomisierung ist nicht mehr oder nur noch teilweise gegeben. Fünftens stellt sich bei randomisierten Studien oft eher die *Frage der Verallgemeinerbarkeit* (externe Validität) als bei nicht-randomisierten Studien. Diese Frage ist allerdings nicht durch die Randomisierung begründet, sondern durch die Grösse und die Repräsentativität der an der Studie teilnehmenden Population. Sie stellt sich somit grundsätzlich bei jedem Evaluationsvorhaben. Die Tatsache, dass aktuelle randomisierte Studien oft mit kleinen Populationen durchgeführt werden, verschärft die Diskussion um die Verallgemeinerbarkeit. Im Prinzip sprechen methodische und organisatorische Gründe nicht dagegen, Randomisierung auch in grösseren, repräsentativeren Populationen anzuwenden, wie aktuelle Beispiele aus Frankreich und

England im nächsten Kapitel aufzeigen. Kapitel 4 greift das Thema der Verallgemeinerung mit einigen Praxisempfehlungen wieder auf.

Sind die umrissenen Rahmenbedingungen gegeben und die Einschränkungen der Methode adäquat berücksichtigt, kann eine Evaluation mittels Randomisierung viel zur *Qualität der Ergebnisse* beitragen: Die Transparenz des Evaluationsdesigns ist hoch (Programm- und Kontrollgruppe sind sofort sichtbar), das Zustandekommen des Evaluationsergebnisses ist direkt nachvollziehbar (Vergleich Programm- vs. Kontrollgruppe), die Auswertungsmethodik kann vereinfacht werden, und die Diskussion um eine sinnvolle und glaubwürdige Kontrollgruppe ist bereits durch das Design des Pilotprojekts und die Zufallszuteilung geklärt.

3 Was randomisieren? Aktuelle Beispiele aus Europa

Nachdem im letzten Kapitel grundsätzliche Fragen zum Einsatz der Randomisierung für die kausale Evaluation von Programmen diskutiert wurden, soll in diesem Kapitel anhand von konkreten Beispielen aufgezeigt werden, wo Randomisierung in Evaluationen schon zum Einsatz kam und kommt. Im Folgenden wird auf eine Serie von aktuellen Beispielen von randomisierten Evaluationsstudien in Europa in den letzten Jahren eingegangen. Es wird explizit nicht in Anspruch genommen, dass die Auswahl der Beispiele repräsentativ oder umfassend wäre. Ziel ist, die konkreten Einsatzmöglichkeiten der Randomisierung zu demonstrieren – und damit auch eine Anregung zu liefern, diese Methodik breiter und innovativ auf neue Pilotprojekte und Einsatzbereiche anzuwenden. Ein wichtiges Auswahlkriterium der folgenden Beispiele war deren Aktualität; diverse der erwähnten Studien sind noch nicht abgeschlossen. Der inhaltliche Fokus liegt auf den Bereichen Sozialversicherungen, arbeitsmarktlichen Fragen und Bildung, mit wenigen Ausnahmen. Die folgenden Beispiele sind nach *Objekt der Randomisierung* angeordnet: Welche Aspekte von Politikmassnahmen sind randomisiert worden? Dies zeigt konzeptionell auf, wo konkret Randomisierung in der Praxis ansetzen kann. Des Weitern wird nach verschiedenen inhaltlichen Bereichen unterschieden.

3.1 Randomisierte Zuweisung in Programme

Ein „klassischer“ Einsatzbereich von Randomisierung ist die *Zufallszuweisung in Programme* der Aktiven Arbeitsmarktpolitik. Hierzu zählen Programmmassnahmen, die im Rahmen der Arbeitslosenversicherung den Stellensuchenden zur Unterstützung ihrer Arbeitsmarktfähigkeit zugewiesen werden: Weiterbildungs- und Beschäftigungsprogramme sowie beratende Unterstützung der Stellensuche (job search assistance). Solche „Aktivierungsprogramme“ wurden in den USA bereits schon früh – in den 80er- und 90er-Jahren – durch randomisierte Pilotversuche evaluiert[11]. In Europa fasste die randomisierte Evaluation solcher Programme erst später Fuss. Ein frühes Beispiel ist die Studie von Van den Berg et al. (2006), die in einem Pilotprojekt von 1998/99 mittels Randomisierung den intensivierten Einsatz von Beratung zur Stellensuche, kombiniert mit verstärktem Monitoring der Suchbemühungen, untersuchten. Sie fanden insignifikant negative Effekte auf die Arbeitslosendauer, aber auch auf den akzeptierten Lohn. Ein Problem stellte hier die relativ kleine Grösse des Projekts (knapp 400 Teilnehmende) dar: Die Effekte der Massnahmen hätten hoch sein müssen, um statistisch signifikant zu werden.

Mehr Verbreitung fand die Randomisierung in diesem Politikbereich seit Mitte der 2000er-Jahre. Eine Vorreiterrolle nahmen die skandinavischen Länder ein, insbesondere Dänemark. Dort werden seit einigen Jahren die massgeblichen Reformprojekte in der Arbeitsmarktpolitik systematisch mit randomisierten Designs getestet. Erste Publikationen sind Graversen et al. (2008) und Rosholm (2008). Sie untersuchen ein pilotweise eingeführtes Aktivierungsprogramm in Dänemark. Dessen Grösse ist ausreichend (ca. 4500 Teilnehmende), um Effekte der Teilnahme an der verstärkten Aktivierung – diese bestand aus einer Kombination von Beratung, Monitoring und Weiterbildung – sinnvoll messen zu können. Es stellte sich heraus, dass die randomisierte Programmgruppe eine um mehr als 2 Wochen kürzere Arbeitslosigkeitsdauer und einen deutlich höheren Anteil an Personen, die eine Stelle fanden, auswies. Ein Feldexperiment in Schweden (Hägglund 2009) mit einer ähnlichen Kombination an Massnahmen zeigte ebenfalls mehrheitlich positive Wirkungen in den zwei Outcomes. In den Niederlanden wurde 2012 vom zuständigen Ministerium eine Serie an Feldexperimenten in diversen holländischen Städten lanciert – als Reaktion auf eine Parlamentsmotion, die die Frage nach der Effektivität der arbeitsmarktlichen Wiedereingliederungsmassnahmen stellte. In mehr als einem Dutzend Teilprojekten[12] (von meist 500 bis gut 1000 Teilnehmenden) werden diverse arbeitsmarktliche Massnahmen in randomisierten Programm-Kontrollgruppen-Vergleichen getestet. Die Massnahmen reichen von intensiver Begleitung der Stellensuchenden über aktive Vermittlung von Stellenangeboten, verbunden mit Lohnkostensubventionen, bis zu rein elektronisch durchgeführter Beratung. Ergebnisse werden bis 2014 erwartet.

In der Schweiz kommt die Evaluation von aktivierender Arbeitsmarktpolitik mittels Zufallszuweisung in Programm- und Kontrollgruppe 2008 erstmals zum Einsatz. Im Kanton Aargau wurde eine intensivierte Unterstützungsstrategie für Stellensuchende ab 45 mittels eines randomisierten Designs getestet (Arni 2010, 2012a, b). Die Intervention umfasste 2-wöchentliche Beratung und vor allem ein intensives Coaching-Programm von rund 20 Arbeitstagen Länge. Trotz der kleinen Dimension des Pilotprojektes (ca. 350 Teilnehmende) resultiert ein statistisch signifikanter Effekt auf den Anteil der Personen, die eine Stelle fanden – dieser ist rund 10 Prozentpunkte höher in der Programmgruppe. Trotz der hohen Intensität der Intervention hat sich die Arbeitslosendauer der Individuen der Programmgruppe nicht verlängert im Vergleich zur Kontrollgruppe. Für die Evaluation dieses Feldexperimentes konnte nicht nur auf umfangreiche Daten aus der administrativen Datenbank der Arbeitslosenversicherung (ALV) zurückgegriffen werden, es wurden auch wiederholte Befragungen der Stellensuchenden und der Personalberatenden durchgeführt. Mit diesen kann die Wirkung der Intervention auf das Verhalten der Stellensuchenden nachgezeichnet werden. Es stellt sich einerseits heraus, dass die Personen der Programmgruppe offenbar effizienter gesucht haben – sie fanden mehr Stellen bei gleichem oder gar kleinerem Suchaufwand. Andererseits zeigt sich auch, dass die Ansprüche an den Lohn der gesuchten Stelle tendenziell gesenkt wurden (Arni 2012a,b).

3.2 Randomisierung auf zwei Ebenen

Jenseits der Randomisierung von Programmzuweisungen kann auch, zusätzlich, *auf der regionalen Ebene randomisiert* werden. Dies mag aus zweierlei Gründen sinnvoll sein: Zum Einen kann damit auch kausal analysiert werden, wie die Wirkung der Massnahme mit der Intensität ihrer Nutzung in der Population variiert. Zum Anderen kann so untersucht werden, ob Verdrängungs- oder Umschichtungseffekte eintreten, wenn eine Massnahme besonders intensiv eingesetzt wird.

Ein solches Randomisierungsschema wurde kürzlich in Frankreich umgesetzt. Crépon et al. (2012) dokumentieren ein grossflächiges Pilotprojekt, in welchem *zweistufig randomisiert* wurde: Die Massnahme – eine intensivierte Unterstützung von jungen Stellensuchenden mittels Zuweisungen und Beratung – kam, wie in den obigen Beispielen, in einer per Zufallszuweisung generierten Programmgruppe zum Einsatz, die Kontrollgruppe folgte dem Status Quo. Zusätzlich wurde aber, je nach Arbeitsmarktregion, auch der Anteil der Personen variiert, die Zugang zur intensivierten Massnahme hatten. Die 235 teilnehmenden Regionen wurden per Zufallszuteilung in Gruppen mit 0%, 25%, 50%, 75% oder 100% Teilnahme-Anteil an der neuen Massnahme eingeteilt. Als Resultat zeigt sich, dass die Massnahme über die kürzere Frist (bis 8 Monate nach Eintritt) einen signifikant positiven Einfluss auf die Wahrscheinlichkeit, eine Stelle zu finden, ausübt (plus 2,4 Prozentpunkte). Auf längere Frist (12 Monate) verschwindet jedoch die positive Wirkung auf den Stellensucherfolg wieder. Ebenso konnte kein Einfluss auf die Löhne festgestellt werden. Des Weitern zeigt sich offenbar ein Umschichtungseffekt (für Männer): Im Vergleich zu Regionen ohne Treatment (0% Teilnahme) ergeben sich für Personen der Kontrollgruppe in Regionen mit Treatment im Durchschnitt leicht tiefere Chancen, eine Stelle zu finden. Es wurden also zum Teil Stellen, die sonst an Personen der Kontrollgruppe gegangen wären, dank der neuen Massnahme an Personen der Programmgruppe umgeschichtet. Die Umschichtung erklärt allerdings nicht den gesamten positiven Effekt des Programmes.

Solche Erkenntnisse aus Untersuchungen mit zweistufiger Randomisierung sind für die Politikgestaltung wertvoll: Sie können einen quantitativen Hinweis geben, welcher Teilnahme-Anteil an der neuen Massnahme sinnvoll sein kann bei der definitiven Implementierung des Programms.

3.3 Zufallszuteilung zu Anbietern

Eine weitere nützliche Art der Randomisierung ist die *Zufallszuteilung zu Anbietern* von Programmen. Es ist kaum anzunehmen, dass die Wirkung von Programmmassnahmen nicht von deren Ausrichter abhängig ist – Qualität oder auch Nähe zu den Klientinnen und Klienten mögen hier Stichworte sein. Auch hierzu wurde in Frankreich soeben ein Feldexperiment abgeschlossen. Behaghel et al. (2012a) untersuchen ein Pilotprojekt, in dem intensivierte Beratung von Stellensuchenden randomisiert an private Anbieter ausgelagert wurde. Dabei wurden per Zufallszuteilung eine Kontrollgruppe (Status-Quo Intensität der Beratung in den öffentlichen Agenturen) und zwei Treatment-Gruppen gebildet: Das eine Treatment besteht in intensiverer Beratung in öffentlichen Agenturen der ALV, das andere in intensiverer Beratung durch Privatanbieter. Die höhere Intensität der Beratung steigert die Zugangsraten zu Stellen um 15 bis 35%. Interessanterweise ist diese Erfolgsquote bei öffentlichen Anbietern etwa doppelt so hoch wie bei privaten Anbietern (zumindest über einen Beobachtungszeitraum von 6 Monaten seit Eintritt).

Ebenso interessant ist, dass sich dieselbe Erkenntnis auch in Deutschland nachweisen lässt. Krug et al. (2012, vorläufig) liefern erste Ergebnisse einer randomisierten Studie mit Klientinnen und Klienten „mit besonderen Vermittlungshemmnissen“ in zwei deutschen Arbeitsagenturen. Sie kommen ebenso zum Schluss, dass die agentureigenen Beratungs- und Vermittlungsdienste erfolgreicher waren als die an private Vermittlungsdienstleister ausgelagerten. Personen, die von Ersteren beraten wurden, weisen ein Jahr nach Zuweisung signifikant weniger (akkumulierte) Tage an Arbeitslosigkeit auf. Ein Teil des Effekts geht allerdings darauf zurück, dass sich beim öffentlichen Anbieter mehr Personen ohne Stelle aus der Arbeitslosigkeit abmelden.

Diese Studien zeigen, dass in den letzten Jahren auch erste randomisierte Evaluationen im Arbeitsmarktpolitik-Bereich in den grossen Ökonomien Kontinentaleuropas, Frankreich und Deutschland[13], möglich wurden.

3.4 Randomisierung zusätzlich angebotener Information

Nicht nur Anbieter, Programme und Regionen lassen sich randomisieren, auch die *Randomisierung von zusätzlich angebotenen Informationen* ist hilfreich zur Erprobung von entsprechenden Pilotmassnahmen. Hierzu sind aktuell mehrere Projekte in der Schweiz und in Deutschland in der Umsetzung.

In der Schweiz werden momentan zwei Feldexperimente diesen Typs im Rahmen der dritten Evaluationswelle der Arbeitslosenversicherung durchgeführt. Das eine Projekt[14] testet im Kanton Waadt mittels eines randomisierten Designs aus, ob ein ergänzendes Informationsmodul zur Frage, wie am besten das eigene soziale Netzwerk für die Stellensuche eingesetzt werden kann, eine positive Wirkung auf den Sucherfolg der Personen hat. Das Modul von 20 Minuten Länge wurde in die anfängliche Informationsveranstaltung eingebaut, die für alle neu in die Arbeitslosigkeit eintretenden Stellensuchenden obligatorisch ist. Die Randomisierung erfolgte hier über das Datum der Veranstaltung: An gewissen Daten wurde das zusätzliche Modul eingebaut, an anderen nicht.

Beim zweiten Pilotprojekt[15] wird die zusätzliche Information den Personalberatenden in den RAV (Regionalen Arbeitsvermittlungszentren) angeboten. Diese stehen vor der Herausforderung, möglichst schnell die Arbeitsmarktchancen der stellensuchenden Person am Anfang ihrer Arbeitslosigkeit einzuschätzen. Diese Einschätzung bildet die Grundlage für eine optimale Beratungsstrategie und Nutzung von arbeitsmarktlichen Massnahmen. Im Projekt wird den Personalberatenden ein zusätzliches Instrument – der „Jobchancen-Barometer“ – als Ergänzung angeboten, das eine Einschätzung der Arbeitsmarktchancen aufgrund vergangener vergleichbarer Fälle und von Eigenschaften der stellensuchenden Person liefert. Hier erfolgt die Randomisierung in der Pilotphase nach Fall: Bei gewissen Stellensuchenden wird die Einschätzung angeboten, bei anderen nicht. Ergebnisse zu diesen Projekten sind im Laufe der Jahre 2013 und 2014 zu erwarten. Beide Projekte werden mittels umfangreicher administrativer Daten sowie mittels zusätzlicher Befragungen ausgewertet.

Auch in Deutschland wird aktuell an zwei Feldexperimenten mit einem „Info-Treatment“ gearbeitet. In einem Projekt[16] wird mittels eines randomisierten Designs untersucht, ob ein verstärktes Marketing des Instruments „Entgeltsicherung für ältere Arbeitnehmer“ dessen Nutzungszahlen und gegebenenfalls auch den Arbeitsmarkterfolg älterer Arbeitsloser steigern kann. Diese Entgeltsicherung ist eine befristete Lohnsubvention speziell für ältere Arbeitslose. Das Ziel der Entgeltsicherung ist es, Lohneinbussen teilweise auszugleichen, wenn ältere Arbeitslose eine geringer entlohnte Beschäftigung aufnehmen. An zufällig ausgewählte Förderberechtigte der Entgeltsicherung wurde entweder eine neu erstellte Informationsbroschüre über den „Kombilohn“ oder das offizielle Merkblatt zur Entgeltsicherung versandt. Die Kontrollgruppe bilden jene förderberechtigten Stellensuchenden, die keine der beiden Broschüren erhalten haben. Im zweiten Pilotprojekt[17] geht es darum, die Stellensuchenden mittels einer Broschüre verstärkt über die tatsächlichen Konsequenzen von Arbeitslosigkeit und die Wirksamkeit von aktiver Arbeitssuche zu informieren. Oft werden solche Aspekte von neu Stellensuchenden unterschätzt. Zufällig ausgewählten Arbeitssuchenden (rund 30‘000 Personen) wurde diese Broschüre zugeschickt. Als Kontrollgruppe dienen Stichproben von

Stellensuchenden, die in derselben Zeit arbeitslos wurden. Beide Studien nutzen neben den administrativen Daten eine ergänzende Nachbefragung, um Motive und Reaktionen der arbeitssuchenden Personen auf die „Info-Treatments“ sichtbar zu machen. Ergebnisse sind 2013 zu erwarten.

Eine spezielle Art der Randomisierung von Information ist die *randomisierte Kontrolle von Information in Bewerbungen*. Dieser Ansatz wird bereits seit den 80er-Jahren benutzt, um verschiedene Formen von Diskriminierung aufgrund von Angaben in Bewerbungen zu untersuchen. Hier sollen zwei aktuelle Beispiele aus Frankreich und Belgien erwähnt werden. In einer grossflächigen Studie in Frankreich haben Behaghel et al. (2012b) untersucht, ob die Anonymisierung von Angaben im Lebenslauf das Anstellungsverhalten der Firmen verändert hat. Dazu wurden Fälle von Arbeitsstellen betrachtet, in denen die Arbeitsagenturen (als Dienstleistung für die Firmen) die Stellen ausschrieben und eine erste Selektion von Bewerbenden aussuchten. Innerhalb dieser Selektion wurde mit einem Zufallsgenerator entschieden, ob Grundinformationen zur Person – Name, Adresse, Geschlecht, Photo, Alter, Zivilstand und Anzahl Kinder – weggelassen wurden oder nicht. Dann wurde diese Selektion an Bewerbungsdossiers zur Entscheidung an die Firma weitergereicht. Es zeigt sich, dass in der anonymisierten Prozedur vergleichsweise mehr Frauen zu Bewerbungsgesprächen eingeladen wurden als in der nicht-anonymisierten[18]. In einer aktuellen Studie (Beart et al. 2012) in Flandern, Belgien, wurden fiktive Bewerbungen auf 382 Ausschreibungen von Stellen für junge Bewerberinnen und Bewerber ohne Berufserfahrung generiert. Per Zufallszuteilung wurde entweder ein flämischer oder ein türkischer Name eingesetzt. Als Ergebnis zeigt sich, dass Bewerbungen mit türkisch tönenden Namen signifikant weniger häufig zu Gesprächen eingeladen wurden als jene mit einheimischem Namen. In Arbeitsmärkten mit Personalengpass zeigt sich kein derartiger Diskriminierungseffekt.

Diese im Englischen als „correspondence testing“ bekannte Randomisierungsmethode kann auch für Evaluationen jenseits von Fragen der Diskriminierung benutzt werden. So haben Falk et al. (2005) diese in einer kleineren Studie in der Schweiz auf die Analyse der Wirkung von Computerkursen angewandt. Zusammen mit den Stellensuchenden, die im Rahmen ihrer Arbeitslosigkeit einen Computerkurs besucht hatten, wurden Bewerbungen entweder mit oder ohne das Kurszertifikat versandt. Es stellte sich heraus, dass das Zertifikat einen insignifikant negativen Effekt auf die Wahrscheinlichkeit, für ein Bewerbungsgespräch eingeladen zu werden, ausübten. Die Autoren interpretieren das Resultat dahingehend, dass diese Grundlagen-Computerkurse von potenziellen Arbeitnehmenden als Signal betrachtet wurden, dass die betreffende Person über wenig Computer-Kenntnisse verfügt.

3.5 Anwendungen in den Sozialversicherungen

Nicht nur die Evaluation von arbeitsmarktlichen Fragen kann mit randomisierten Ansätzen angegangen werden, auch andere sozialpolitischen Institutionen können damit untersucht werden. Die *Randomisierung von Massnahmen in der Invalidenversicherung* (IV) ist ein neues Einsatzgebiet, zu dem noch kaum Studien bestehen. Eine sich noch in Arbeit befindliche Studie (Kauer et al. 2012, vorläufig) untersucht in der Schweiz die Auswirkungen von Lohnsubventionen der IV bei Anstellungen von Personen mit Behinderung. Auch hier wurde der Ansatz des „correspondence testing“ benutzt: Per Zufallszuteilung wurde teilweise erwähnt, dass die bewerbende Person eine Lohnsubvention der IV mitbringt, teilweise nicht. Die Wirkung dieser Information wird wiederum

anhand der Zahl der Einladungen zu Bewerbungsgesprächen gemessen. Die erste Welle dieser Untersuchung, die Abgänger aus geschützten Ausbildungseinrichtungen betrachtet, kommt zum Schluss, dass die Lohnunterstützung keine positive Wirkung auf die Bewerbungschancen der Personen zeigt – manchmal, besonders im Fall von Spontanbewerbungen, kann sie sogar kontraproduktiv wirken.

Engström et al. (2012) dokumentieren ein grosses randomisiertes Feldexperiment in Schweden (ca. 13‘500 Teilnehmende), in dem krankgeschriebene Personen in der Programmgruppe früher und priorisiert einer Intervention zugewiesen wurden. Diese Intervention der schwedischen Sozialversicherungen im Jahr 2007 bestand darin, früher als bisher abzuklären, welche Massnahmen der beruflichen Reintegration die betroffene Person benötigte, und diese umzusetzen. Entgegen den Erwartungen findet die Studie, dass die Individuen der Programmgruppe in den 15 Monaten nach der Intervention mehr Krankheitsabwesenheiten zeigten und ebenso eine grössere Wahrscheinlichkeit, IV-Rentenbeziehende zu werden. Die Autoren führen als mögliche Erklärung an, dass die frühere Intervention dazu beigetragen haben könnte, dass der schlechte Gesundheitszustand der Personen verstärkt sichtbar gemacht wurde, mit negativen Auswirkungen auf die Arbeitssuche.

Auch im Bereich *Sozialhilfe und Wiedereingliederung* in den Arbeitsmarkt können randomisierte Ansätze hilfreich sein, um transparente Evaluationsergebnisse zu produzieren. Umfassend, was die Grösse anbetrifft, ist diesbezüglich das ERA Feldexperiment, das von 2003 bis 2008 in Grossbritannien durchgeführt wurde (Hendra et al. 2011). ERA steht für „Employment Retention and Advancement“. Das Ziel des grossflächigen Pilotprojekts (16‘000 freiwillig Teilnehmende) besteht darin, sozialhilfeabhängige, schwer vermittelbare Personen nachhaltig wieder in nicht unterstützter Vollzeitarbeit zu etablieren. Das Projekt fokussiert auf alleinerziehende Eltern, die arbeitslos oder „working poor“ sind, sowie auf langzeitarbeitslose Personen ab 25 Jahren Alter. Die randomisiert zugewiesene Intervention besteht darin, dass diese Personen während 33 Monaten Zugang hatten zu zusätzlicher Beratung in den „Jobcentres Plus“. Diese Beratung zielte darauf ab, den Programmgruppen-Teilnehmenden dabei zu helfen, einmal gefundene Jobs zu behalten oder stabilere oder besser bezahlte Stellen zu erreichen. Des Weitern bekamen die Personen der Treatment-Gruppe dreimal pro Jahr einen Geldbonus (über zwei Jahre), falls sie in einer Vollzeitanstellung waren. Schliesslich wurden auch Unterstützungen für Weiterbildung bezahlt. Die alleinerziehenden Eltern der Programmgruppe profitierten über die kürzere Frist, was die Löhne und Vollzeitbeschäftigung angeht, wie die Resultate zeigen. Über die längere Frist (die Personen wurden 5 Jahre beobachtet) jedoch verschwand der positive Wiedereingliederungseffekt, besonders nach dem Ende der Unterstützung (nach 33 Monaten). In einer Kosten-Nutzen Analyse schnitt nur die Subgruppe der besser ausgebildeten alleinerziehenden Eltern positiv ab. Erfreulicher sind die Resultate für die langzeitarbeitslosen Programmgruppen-Teilnehmenden: ERA erhöhte den Anteil der Personen in Beschäftigung leicht und nachhaltig, substanziell und nachhaltig (über die 5 Jahre) waren die Zunahmen im Einkommen. Hinsichtlich Kosteneffektivität war ERA für diese Gruppe sowohl für die Teilnehmenden wie für das Budget der sozialen Institutionen finanziell positiv.

Mehr randomisierte Studien zu den Anreizwirkungen von Massnahmen in der IV wie auch im Sozialhilfe-Bereich wären wünschbar, insbesondere in Kontinentaleuropa. Steigende Bezügerzahlen wie auch zunehmende Ausgaben und Budgetdruck haben die politische Diskussion um Reformen in diesen Sozialwerken in den letzten Jahren intensiviert. Gleichzeitig liegt noch kaum empirische Evidenz aus systematischen Evaluationen vor, in denen sich die getesteten Massnahmen als effektiv

und nachhaltig erwiesen. Diese Entwicklungen dürften die Nachfrage nach systematischen Erprobungen neuer Massnahmen in diesen Bereichen steigern.

3.6 Anwendungen im Bildungs- und Umweltbereich

Die Anwendung von Feldexperimenten ist nicht nur im Bereich Sozialversicherungen und Arbeitsmarkt mach- und wünschbar. Randomisierte kausale Evaluationen sind grundsätzlich in allen Bereichen denkbar, in denen klar abgrenzbare Programmmassnahmen ergriffen werden. Zur Illustration seien hier beispielsweise drei aktuelle *randomisierte Studien aus dem Bildungs- sowie aus dem Umweltbereich* erwähnt. Ein Feldexperiment im Bereich der Weiterbildung wurde mit Teilnehmenden der Schweizerischen Arbeitskräfteerhebung (SAKE) von 2005 und 2006 durchgeführt (Schwerdt et al. 2012). Den zufällig bestimmten Mitgliedern der Programmgruppe, die zwischen 20 und 60 Jahre alt sein mussten, wurde ein Bildungsgutschein abgegeben, zusammen mit einem Brief, der ihnen erklärte, dass sie Teil eines Projektes zur Unterstützung von „lebenslangem Lernen“ seien. Die Gutscheine wiesen einen Wert von 200, 750 oder 1500 CHF auf und konnten bis Mitte 2006 für beliebig wählbare Weiterbildungskurse eingesetzt werden. Nur gerade rund 18 % der gut 2400 Programmgruppen-Mitglieder lösten den Gutschein effektiv ein und besuchten einen Kurs. Die ausgewählten Kurse dauerten im Durchschnitt 42 Stunden und deckten in ihrer grossen Mehrheit arbeitsmarktrelevante Themen ab. Die Intervention zeigte keinen signifikanten Effekt auf die Outcomes, die mit der SAKE gemessen wurden – weder auf Einkommen, Beschäftigung noch auf die Aufnahme von Weiterbildung im Jahr nach der Intervention. Während die Gutscheine am ehesten bei Personen mit höherer Bildung Weiterbildungsaktivitäten auslösten, hätte sich die Wirkung aufs Einkommen am ehesten bei Personen mit Lehrabschluss positiv gezeigt. Bei diesen Personen wurde aber deutlich weniger Weiterbildungsbeteiligung ausgelöst. Die unfokussierte Ausgabe von Bildungsgutscheinen scheint in diesem Falle also nicht ein effizientes Mittel zur Steigerung des lebenslangen Lernens zu sein.

Lebenslanges Lernen beginnt zur Stunde Null: Die frühest-möglichen Bildungsmassnahmen sind Interventionen in den ersten drei Lebensjahren zur Unterstützung der frühkindlichen Entwicklung. Diese Massnahmen, die unter den Begriffen der Frühförderung (Schweiz) oder der Frühen Hilfen (Deutschland) von sozialen Institutionen angeboten werden, sind sowohl auf die Kleinkinder selbst wie auch auf deren Eltern ausgerichtet. Sie konzentrieren sich auf Familien „in belastenden Lebenslagen mit geringen Bewältigungsressourcen“ und zielen darauf ab, das Risiko zu vermindern, dass die Kinder bereits mit schlechteren Entwicklungschancen ins Leben starten (NZFH 2010, Heckman 2010). Entsprechend konzentriert sich das „Modellprojekt Pro Kind“ in Deutschland auf Frühe Hilfen für Familien in belastenden Lebenslagen (Lutz et al. 2010). Eine solche wird im Projekt einerseits durch den Bezug von Arbeitslosengeld oder Sozialhilfe definiert sowie andererseits durch einen weiteren „persönlichen oder sozialen Belastungsfaktor“ (z.B. Minderjährigkeit, alleinerziehend, soziale Isolation oder fehlende Ausbildung). Erstgebärende Frauen in solchen Lebenslagen, die in der randomisierten Programmgruppe teilnehmen, profitieren von 52 Hausbesuchen von Hebammen oder Sozialpädagoginnen, die während der Schwangerschaft beginnen und am zweiten Geburtstag des Kindes enden. Die Familienbegleiterinnen unterstützen dabei die Teilnehmerinnen in Bereichen wie der Schaffung einer gesundheitsförderlichen Umgebung, der Entwicklung der Elternrolle, des sozialen Umfelds und der eigenen Lebensperspektive sowie der Nutzung von Gesundheitsversorgung und sozialen Diensten. Je rund 370 Frauen in der Kontroll- und der Programmgruppe nehmen am

Pilotprojekt teil. Erste Ergebnisse zeigen auf, dass die gesundheitliche Entwicklung der Kinder in der Programmgruppe positiv beeinflusst wird bzw. Entwicklungsverzögerungen reduziert werden.

Um die (potenzielle) Breite des Ansatzes der Randomisierung aufzuzeigen, soll hier die Diskussion von Anwendungen mit einem aktuellen Beispiel aus einem ganz anderen Bereich abgeschlossen werden. In der Schweiz wurde 2010 ein randomisiertes Projekt[19] lanciert, das verschiedene mögliche Anreize zum Stromsparen evaluiert. Unter den rund 6000 freiwillig an der Studie teilnehmenden Personen wurden vier Treatment-Gruppen und eine Kontrollgruppe gebildet mittels Randomisierung. Die erste Programmgruppe bekam ein Smart Metering System, d.h. bei diesen Personen wurde zuhause ein „intelligentes" Strommesssystem installiert. Die zweite Programmgruppe erhielt individuelle Stromberatung. Der dritten und vierten Programmgruppe wurden zwei unterschiedliche Versionen von Briefen versandt, wo sie mit dem Stromkonsum von vergleichbaren Haushalten konfrontiert wurden. Um sinnvolle Vergleichsreferenzen schaffen zu können, wurden die Teilnehmenden zuerst nach ihrem Stromverbrauch (vor der Intervention) sortiert und dann randomisiert in die Gruppen verteilt. Interessierende Outcomes dieser Evaluationsstudie sind einerseits der Stromverbrauch und andererseits verschiedene Masse zum Wissen bezüglich Stromverbrauch und zu Einsparpotenzialen sowie Einstellungen zu Umweltthemen. Im Laufe von 2013 sind erste Ergebnisse zu erwarten, inwieweit sich Smart Metering, Stromberatung oder soziale Vergleiche auf das Stromnutzungsverhalten der teilnehmenden Personen ausgewirkt haben.

3.7 Fazit

Die Vielfalt an randomisierten Pilotprojekten, die in den letzten Jahren in Europa lanciert worden sind, zeigt das *Potenzial der Randomisierung* für die kausale Evaluation von politischen Massnahmen ansatzweise auf. Dieses Potenzial ist bei Weitem noch nicht ausgeschöpft. Je nachdem, welche Evaluationsfragen im Zentrum stehen, können gewisse Aspekte von Pilotprojekten – Programmteilnahme, Information, Anbieter, Dauer bzw. Zeitpunkt etc. – randomisiert aufgesetzt werden. Besteht der Wille in öffentlichen und privaten Institutionen, innovative Massnahmen und Reformen im Rahmen von Pilotversuchen oder Modellprojekten auszutesten, kann mit *relativ kleinem Zusatzaufwand* die Pilotstudie randomisiert aufgesetzt werden. Die einzige wesentliche Anpassung, die in diesem Fall erfolgt, ist, dass die Frage, wer in der Pilotphase mit der neuen Massnahme konfrontiert wird, über Zufallszuteilung geregelt ist anstatt über ein anderes festzulegendes Kriterium. Der *Nutzen* dieser Anpassung ist massgeblich, kann doch dadurch die Qualität und Klarheit der Messung der kausalen Programmwirkung klar erhöht werden.

Institutionen unterschiedlichster Politikbereiche sind aufgrund des hohen Tempos gesellschaftlicher Veränderungen und des hohen Budgetdrucks permanent mit der Herausforderung konfrontiert, die Auswahl, Gezieltheit und Wirksamkeit ihrer politischen Instrumente zu optimieren. Als Grundlage für solche Optimierungen ist fundiertes Wissen zu kausalen Programmwirkungen von hohem Wert. Gut umgesetzte randomisierte Studien können solches Wissen in hoher Qualität liefern. Thematisch sind der Anwendung von Randomisierung kaum Grenzen gesetzt, sofern das Objekt der Evaluation eine klar umreissbare Massnahme oder Intervention ist. Verschiedenste Anwendungsbereiche jenseits der oben diskutierten sind denkbar. Mögliche Beispiele wären etwa randomisierte Erprobungen von neuen Präventionsmassnahmen oder Informationsstrategien etc.

4. Wie randomisieren? Planung und Praxis

Im letzten Teil dieses Beitrags soll im Folgenden auf einige Punkte eingegangen werden, die in der Praxis bei der Planung und Umsetzung von randomisierten Evaluationen wichtig zu berücksichtigen sind. Wenn diese (und vergleichbare) Themen bereits im Voraus in der Planung des (Pilot-)Projektes bedacht und diskutiert werden, können potenziell einige „Fallen" vermieden werden – Aspekte, bei denen sonst später Kritik oder Verzerrungen der Ergebnisse der kausalen Evaluation auftreten könnten. Die folgenden Empfehlungen sind nach der Ablaufchronologie eines zu evaluierenden (Pilot-)Projektes gegliedert.

Evaluation bereits in der Projektgestaltung mitberücksichtigen. Dieser Punkt ist absolut zentral, wenn es um die Umsetzung von randomisierten Studien und kausalen Evaluationen generell geht. Die Glaubwürdigkeit und Aussagekraft der Evaluation steht und fällt mit dem *Design* des Pilotprojektes. Wie die Diskussion in Kapitel 2 und die folgenden Punkte im Detail aufzeigen, ist es notwendig, dass die Programm-und Kontrollgruppen-Definitionen, die Randomisierung und alle damit verbundenen organisatorischen Massnahmen im Projektdesign mitberücksichtigt werden. Dies bedingt einen gewissen – v.a. konzeptionellen – Zusatzaufwand zum Zeitpunkt der Projektplanung. Der Zusatznutzen, der dadurch generiert wird – durch Transparenz und Glaubwürdigkeit des Evaluationsdesigns und damit der späteren Ergebnisse, durch methodisch einfachere Auswertungen etc. – rechtfertigen den höheren Anfangsaufwand. Es handelt sich also um eine anfängliche Investition, die sich später auszahlt.

Bestimmung der Mindestgrösse der Teilnehmenden-Population. Diese Frage stellt sich prinzipiell bei jedem Projekt, auch ohne Randomisierung. Im Falle von randomisierten Studien kommt die Frage tendenziell früher auf, da die Randomisierung in organisatorische Prozesse eingebaut werden muss (siehe nächsten Punkt), für deren Umsetzung die Grösse des Projekts relevant sein kann. Je grösser das Projekt ist, desto einfacher sollte die Randomisierungsmethode sein. Bei randomisierten Projekten ist die Berechnung der notwendigen Mindestgrösse der Teilnehmenden-Population vergleichsweise einfacher als bei nicht-randomisierten Projekten, da meist einfachere statistische Methoden zur Berechnung des Effekts der Intervention auf die Zielgrössen zum Einsatz kommen. Für diese Grössenbestimmung muss zuerst folgende Grundsatzfrage geklärt werden: Welche Grösse des Effekts der geplanten Intervention ist zu erwarten (oder gewünscht)? Diese Frage kann auf verschiedene Arten geklärt werden: Aus theoretischen Überlegungen, aus früheren vergleichbaren Praxiserfahrungen, oder die auftraggebende Institution gibt eine Zielgrösse vor. Ebenso kann eine Kosten-Überlegung sinnvoll sein: Wie gross muss die positive Wirkung der Intervention sein, damit sie die Kosten der Intervention rechtfertigt? Ist diese anzuvisierende Effektgrösse festgelegt, kann mittels statistischer Berechnungen (sogenannten „power calculations", siehe Endnote 9 als Beispiel) hochgerechnet werden, wie gross die Teilnehmenden-Population mindestens sein muss, um die geplante Ziel-Effektgrösse statistisch signifikant messen zu können. Wird diese Minimalzahl in der Projektumsetzung unterschritten, ist das Risiko relativ gross, dass die spätere Wirkungsmessung mit dem unbefriedigenden Ergebnis endet, dass keine statistisch abgesicherten Programm-Effekte ausgewiesen werden können. Es empfiehlt sich, die Planung der Teilnehmenden-Population auf eine grössere Anzahl auszurichten als die errechnete Mindestgrösse. Häufig führen unvorhergesehene Veränderungen im Umfeld oder Ausfälle von Teilnehmenden zu kleineren Populationszahlen als ursprünglich geplant.

Wahl eines möglichst einfachen Randomisierungs-Prozesses. Die Zufallszuteilung muss im Feld möglichst einfach und praxisnah von der Hand gehen, damit die Implementierung solcher Projekte erfolgreich ist. Die mitarbeitenden Personen im Alltagsgeschäft der betrachteten Institution sollten möglichst wenig damit belastet werden. Eine gute Möglichkeit sind deshalb Computer-Zufallsgeneratoren, die entweder automatisch oder bedient durch die Evaluationspersonen die Randomisierung vornehmen. Im in Kapitel 3 erwähnten Projekt von Arni (2010, 2012a,b) etwa kam ein solcher Zufallsgenerator zum Einsatz, um eine Zufallsliste zu generieren, wo die teilnehmenden stellensuchenden Personen dann in der Reihe ihres Eintritts eingetragen wurden. Die Zufallsliste ist eine simple Zufallsabfolge von P's (Zuordnung zu Programmgruppe) und K's (Zuordnung zu Kontrollgruppe). Informations-Treatments können oft direkt am Computer zufallsverteilt werden. So geschehen z.B. mit dem Versand der Infobroschüren in den erwähnten deutschen Projekten, oder auch im Projekt „Jobchancen-Barometer" der dritten Evaluationswelle der ALV in der Schweiz (dort entscheidet der Zufallsgenerator über die Verfügbarkeit der Barometer-Prognose). Auch administrative Daten, von denen bekannt ist, dass sie zufällig bzw. ungesteuert entstehen, können zur einfachen Randomisierung benutzt werden: Zum Beispiel der Tag oder die Woche des Programmstarts bei sehr häufig durchgeführten Massnahmen (wird z.B. im Projekt SOCNET der dritten Evaluationswelle genutzt) oder der Tag des Geburtstages der Person. Andere zufällig und möglichst gleichmässig verteilte Daten oder Ereignisse sind denkbar als Basis für die Randomisierung.

Strikte Abfolge und Zeitplanung der Interventionen. Um zeitlich bedingte Verzerrungen im Programm-Kontrollgruppen-Vergleich zu vermindern, ist eine möglichst strikt fixierte Abfolge und Terminierung der verschiedenen Schritte der Projektinterventionen wichtig. Dies beginnt bereits ganz am Anfang: Die Gruppe der Teilnehmenden am Projekt muss *vor* der Randomisierung definiert und selektiert werden. Die Projektpopulation kann, je nach Situation, durch den institutionellen Kontext definiert werden (z.B. die Eintretenden in die ALV zu einem gewissen Zeitpunkt) oder z.B. durch Selektion von Freiwilligen aus einer vorher angefragten Grundpopulation von potenziell in Frage kommenden Personen[20]. Erst *nach* dieser Auswahl kommt die Zufallszuteilung in Programm- und Kontrollgruppe zum Einsatz. Ziel ist, dass die Zufallszuteilung nachher nicht mehr durch nicht-zufällige Nachselektionen oder aus dem Projekt Ausgeschlossene beeinträchtigt wird. Falls die Intervention danach mehrere aufeinanderfolgende Schritte enthält (z.B. Information, dann Beratung, dann Zahlung eines Geldbonus), sollten die Abfolge sowie die Terminierung der Schritte für alle Personen der(selben) Programmgruppe möglichst *gleich* sein. Je besser dies gegeben ist, desto besser kann man die Wirkung eines *spezifischen* Interventionsschrittes anhand seiner Terminierung identifizieren in der Evaluation. Wenn zum Beispiel die spezielle Beratung der teilnehmenden Personen immer nach sechs Monaten erfolgt, weiss man, dass die frühestmögliche Wirkung auf die Outcomes dieser Personen diesem Zeitpunkt eintreten kann; dies lässt sich in der Auswertung entsprechend berücksichtigen[21]. Sind einzelne Teilschritte von Interventionen immer ähnlich terminiert, lässt sich damit auch das Risiko vermindern, dass sich die einzelnen interessierenden Effekte überlappen und nicht mehr auseinander zu halten sind.

Kontrolle externer Einflüsse (interne Validität). Dies ist in der Praxis wohl der schwierigste Punkt. Im Alltagseinsatz ist es offensichtlich nicht möglich, alle externen Einflüsse, die die Ergebnisse der Studie mitprägen können, auszuschliessen (dies unterscheidet Feld- von Laborexperimenten). Allerdings vermindert die Randomisierung dieses Problematik massgeblich: Dank der Zufallszuteilung wird der Programm-Kontrollgruppen-Vergleich nicht verzerrt durch allgemein wirkende Ereignisse (wie etwa Saisonalität, ein Medienereignis, eine andere Politikreform etc.); deren Wirkung ist dann gleich

verteilt auf die beiden Gruppen. Anders sieht es aus, wenn diese externen Einflüsse durch die Programm-Interventionen verstärken oder abschwächen: Dann ist der gemessene Programm-Effekt ein Gemisch aus der eigentlichen Wirkung der Intervention plus der ungleichen Auswirkung der externen Einflüsse. Daher ist auch bei randomisierten Studien darauf zu achten, dass externe Einflüsse, die potenziell mit der Intervention zusammenspielen könnten, möglichst klein gehalten oder kontrolliert werden können. So ist es zum Beispiel wünschenswert, wenn in der Zeit des Pilotversuchs nicht noch andere Politikreformen oder Strategieveränderungen in der betrachteten Institution ablaufen. Eine zeitlich relativ schnelle Durchführung der Interventionen (d.h. relativ viele Teilnehmende in kurzer Zeit), kann hilfreich sein, zeitliche Volatilitäten relativ klein zu halten. Die Personen, die die Programm- und Kontrollgruppe betreuen in der Institution, sollten optimalerweise dieselben oder sehr gut vergleichbar sein (ausser, wenn die unterschiedliche Betreuung Teil des Treatments ist). Dies sind nur einige Aspekte, die es im Auge zu behalten gilt. Weitere relevante externe Einflüsse mit Verzerrungspotenzial werden in den nächsten zwei Punkten erwähnt.

Zentrale Rolle der Information. Die Kommunikation und das Wissen, welche projektbezogenen Informationen zu den Teilnehmenden gelangen, sollte optimalerweise durch die Evaluatorin oder den Evaluatoren beobachtbar und mitbestimmbar sein. Ungleicher Informationsstand zwischen Programm- und Kontrollgruppe kann die Messung der Programmwirkung verzerren. Bezüglich der Menge der bekannt zu gebenden Information zum Projekt existiert ein gewisses Dilemma: Wollte man den theoretischen Standard der „Doppelverblindung“ – d.h. Durchführende und Betroffene der Intervention sind sich dessen nicht bewusst – anwenden, müsste man die Information zum Projekt auf Null minimieren. Dies ist in der Praxis nicht machbar und meist auch nicht wünschbar. Evaluation ist typischerweise mit Fragebogen verbunden, schon dies alleine verhindert eine Nullinformation. Auf der anderen Seite sollten allerdings ungleiche Informationsstände (ausser bei einem Info-Treatment) sowie „Framing“ verhindert werden. Letzteres kann beispielsweise dadurch entstehen, wenn das Projektziel zu explizit an die Teilnehmenden kommuniziert wird: Besteht zum Beispiel das Ziel eines Projektes darin, Personen vor einem Fall in die Sozialhilfe zu schützen, könnten sich die so informierten Teilnehmenden stigmatisiert fühlen: Die Institution „stempelt“ sie als potenzielle Sozialhilfebezüger „ab“. Dies kann zu Frustration führen, was natürlich auf keiner Seite gewünscht ist. In der Kommunikation scheint daher oft ein Mittelweg sinnvoll: Die einheitliche, frühzeitige Information aller Teilnehmenden (Programm- und Kontrollgruppe) über die Existenz und grobe inhaltliche Ausrichtung des Pilotprojektes, ohne die expliziten Forschungsfragen öffentlich zu machen. Dies verhindert ungleiche Informationsstände, die Bildung von Gerüchten und Halbwahrheiten und ungewünschte Interpretationen, soweit möglich.

Brutto-Analyse vs. Detail-Analyse. Bei einer zu evaluierenden Intervention mit mehreren Teilen bzw. Stufen mögen die Auftraggebenden oft an zwei politikrelevanten Fragen besonders interessiert sein: Erstens, wie hat sich die Intervention als Ganzes auf die interessierenden Zielgrössen ausgewirkt? Zweitens, welche Elemente der Intervention waren besonders wirksam und damit für primär für die Gesamtwirkung verantwortlich? Die erste Frage – die Brutto-Analyse zur Eruierung des Gesamteffekts – lässt sich dank der anfänglichen Randomisierung methodisch sehr einfach und sauber beantworten: Vergleich von verschiedenen statistischen Grössen der Outcomes der Programmgruppe mit denselben Grössen der Kontrollgruppe. Die Detail-Analyse eines spezifischen Interventionsschrittes, der deutlich später als der Randomisierungszeitpunkt stattfand, wird statistisch hingegen aufwändiger: Einerseits stellen sich dann Fragen, wann genau das Outcome auf diesen Interventionsschritt gemessen werden soll. Wie oben beschrieben, ist ein fixiertes Timing der

Schritte hier zentral zur Identifikation von Teileffekten. Dennoch löst dies nicht alle Probleme. Vorangehende Schritte der Intervention können beispielsweise weiterhin Wirkung entfalten. Solche Überlappungen von Teileffekten können nur mit gewissen Annahmen (z.B. über die Persistenz und Dauer von gewissen Teileffekten) statistisch auseinander gehalten werden.

Trotz diesen Einschränkungen ist die Schätzung von Treatment-Effekten für die verschiedenen Phasen der Intervention von grossem Interesse für die Politikgestaltung: Eine solche Analyse bringt Aussagen zur *Entwicklung* der Wirkung der Intervention über die Zeit. Eine weitere Herausforderung bei einer solchen Analyse ist die sogenannte *dynamische Selektion*: Oft „verschwindet" bei wiederholten Erhebungen über die Zeit ein gewisser Teil der anfänglichen Teilnehmendenpopulation. Dies kann in der Natur der Sache liegen (z.B. Abgänge aus Arbeitslosigkeit, Sozialhilfe, IV werden oft nicht mehr beobachtet in den vorhandenen Datenquellen) oder auch aktiv durch fehlende Teilnahme der Beteiligten bei wiederholten Befragungen („panel attrition") auftreten. Da Abgänge aus der Teilnehmendenpopulation oder Nichtteilnahme an Befragungen oft mit den Zielgrössen der Studie zusammenhängen, können solche Prozesse potenziell die Wirkungsmessung verzerren. Dies trifft dann ein, wenn die dynamische Selektion die Programm- und die Kontrollgruppe *unterschiedlich* stark betrifft. Dann geht ein Teil der mit der anfänglichen Randomisierung hergestellten Vergleichbarkeit verloren. In diesem Fall werden statistische Selektionskorrekturen notwendig. Deskriptive Vergleiche der Eigenschaften der beiden Gruppen an diversen Zeitpunkten können einen Hinweis geben, inwieweit die Programm- und Kontrollgruppe in der später noch vorhandenen Teilpopulation noch vergleichbar sind. Dies betrifft allerdings nur die beobachtbaren Eigenschaften; unbeobachtete Eigenschaften könnten trotzdem ungleich verteilt sein. Letzteres kann nur durch die Nutzung von entsprechenden statistischen Modellen und Annahmen analysiert werden. Es muss hier festgehalten werden, dass Probleme der dynamischen Selektion in nicht-randomisierten Studien deutlich häufiger vorhanden sind; dort muss die Selektion von Anfang an korrigiert werden. Dank Randomisierung reduziert sich das Selektionsproblem in seinem Ausmass und beschränkt sich auf spätere Beobachtungszeitpunkte.

Effektive Beteiligung an der Intervention. Wichtig ist, dass sich die Evaluatorinnen und Evaluatoren bewusst ist, dass in randomisierten Studien normalerweise die *Zuweisung* zur Intervention (zum Treatment) randomisiert ist. Ob dann alle Personen der Programmgruppe die Intervention auch wirklich durchlaufen, ist nicht gesichert. Ein extremes Beispiel ist diesbezüglich die in Kapitel 3 diskutierte Studie von Schwerdt et al. (2012): Die Bildungsgutscheine wurden nur von einem Bruchteil der Programmgruppe wirklich benutzt. Andere Formen der Nichtaufnahme wären z.B. Situationen, wo die Teilnehmenden das Treatment nicht mögen (oder es als nutzlos betrachten) und deshalb versuchen, dieses bewusst zu vermeiden („non-compliance") – beispielsweise, indem sie sich krank melden. Ist die Treatment-Aufnahme-Quote nicht 100%, ergeben sich grundsätzlich zwei Formen der Analyse und damit verbundenen Politikfragen (s.a. Angrist et al. 2009): Die erste ist die Analyse nach Treatment-Absicht (*ITT, intention-to-treat*): Alle Teilnehmenden der Programmgruppe werden im Vergleich zur Kontrollgruppe betrachtet, unabhängig davon, ob sie effektiv die Intervention durchliefen oder nicht. Diese sogenannte ITT-Analyse entspricht dem statistisch einfachen Gruppenvergleich. Die Politikfrage, die damit beantwortet wird, ist folgende: „Falls die Nutzung der neuen Massnahme so intensiv ist wie in der Population der Programmgruppe, wie hoch ist dann die Wirkung der Massnahme für diese ganze Population?" Die zweite Art der Analyse ist jene der Wirkung der Treatment-Teilnahme. Dazu werden jene Personen der Programmgruppe mit der Kontrollgruppe verglichen, die die Intervention effektiv durchlaufen haben[22]. Hier wird folgende

Politikfrage behandelt: „Welches ist die Wirkung der *Teilnahme* an der neuen Massnahme auf die interessierenden Zielgrössen?“. Beide Fragen sind wichtig und instruktiv für die Politikgestaltung. Sie sollten jedoch in der Darstellung der Evaluationsergebnisse bewusst auseinandergehalten und in den richtigen Kontext gesetzt werden.

Verallgemeinerbarkeit (externe Validität). Wenn die Ergebnisse der Evaluation einer randomisierten Studie vorliegen, lohnt es sich, deren externe Validität bzw. Verallgemeinerbarkeit zu diskutieren. Da randomisierte Pilotprojekte oft in einer vergleichsweise kleinen Population durchgeführt werden, ist diese Frage von Relevanz. Zu fragen gilt es u.a., wie „speziell“ die Population im Pilotprojekt war im Vergleich zur gesamten Population von Personen, die bei einer definitiven Implementierung der neuen Massnahme betroffen wären. Ebenso sollte diskutiert werden, ob die Einführung der neuen Massnahme sich indirekt auf Nichtteilnehmer des Projekts ausserhalb der Pilotregionen ausgewirkt haben könnte („spill over“-Effekte oder „contagion“, s.a. White 2006)[23]. Drittens gilt es zu diskutieren, ob die Ergebnisse der Pilotstudie bei Wiederholung zu einem anderen Zeitpunkt (andere Saison, andere Konjunkturlage, anderes politisches Umfeld etc.) noch in etwa dieselben wären. Diese Fragen stellen sich natürlich auch in nicht-randomisierten Studien.

Etwas häufiger relevant für randomisierte als für nicht-randomisierte Projekte sind Fragen, die unter dem Begriff des „Hawthorne-Effekts“ zusammengefasst werden können. Dieser besagt, dass Personen, die sich bewusst beobachtet fühlen, ihr Verhalten anpassen können (sich z.B. mehr anstrengen). Die anfängliche Information der Teilnehmenden (siehe weiter oben) sowie die Tatsache, dass meistens begleitend zur Studie Befragungen durchgeführt werden, können allenfalls einen Hawthorne-Effekt auslösen[24]. Da Information und Befragungen die Programm- *und* die Kontrollgruppe betreffen, wird das Ergebnis des randomisierten Programm-Kontrollgruppen-Vergleiches durch ihn nicht verzerrt. Die allfällige Präsenz eines Hawthorne-Effekts muss jedoch in der Diskussion der externen Validität berücksichtigt werden: Da in der dauerhaften Implementierung der neuen Massnahme später der Hawthorne-Effekt wegfällt, wird die schliesslich beobachtete Wirkung der neuen Massnahme auf einem höheren oder tieferen Niveau des Outcomes sichtbar, als es im Pilotprojekt der Fall war. Das Risiko eines Hawthorne-Effektes könnte minimiert werden, indem die Befragungen nicht begleitend sondern retrospektiv durchgeführt würden und die anfängliche Information auf ein Minimum beschränkt würde. Dies würde jedoch auf Kosten der Qualität der Befragungen gehen (retrospektiv ist meist ungenauer als begleitend) und die weiter oben diskutierten Problematiken ungleicher und ungenügender Information verstärken. All diese Aspekte sollten berücksichtigt oder zumindest diskutiert werden, wenn es darum geht, eine Prognose abzugeben, inwieweit die im Pilotprojekt eruierten Wirkungen sich auch zeigen würden bei einer allgemeinen, breiflächigen Einführung der neuen Massnahme.

Anwendungsbereich der diskutierten Fragen. Die Diskussion dieser Planungs- und Praxisfragen zur Durchführung randomisierter Studien soll mit folgender wichtiger Bemerkung abgeschlossen werden: Die meisten der soeben diskutierten Punkte gelten auch für die Implementierung nicht-randomisierter Evaluationen. Viele dieser Punkte sind im nicht-randomisierten Fall meist schwieriger (d.h. statistisch komplexer) zu handhaben: Etwa die Frage der Kontrolle externer Einflüsse oder von dynamischer Selektion, oder auch jene der Identifikation von Teileffekten von mehrstufigen Massnahmen. Der Unterschied liegt darin, dass man im Falle von randomisierten Evaluationen nicht darum herum kommt, sich mit diesen Grundfragen der sauberen Durchführung von Studien früh auseinanderzusetzen. Erst im Nachhinein geplante Evaluationen sind bei randomisierten Studien nicht möglich.

5. Schlussbemerkungen

Dieser Beitrag hat zum Ziel, das Instrument der Randomisierung (Zufallszuweisung) für die Nutzung in Evaluationen von Pilotprojekten zu diskutieren. Er zeigt auf, weshalb und unter welchen Bedingungen eine vermehrte Nutzung von Randomisierung in der Programm-Evaluation sinnvoll sein kann, diskutiert aktuelle Anwendungsbeispiele aus Europa und weist auf eine Reihe wichtiger Punkte hin, die bei der Umsetzung von randomisierten Studien bedacht werden sollten.

Um möglichst viel an Klarheit und Aussagekraft aus einer Programm-Evaluation herausholen zu können, ist es von zentraler Bedeutung, dass diese *bereits in der Gestaltung des Pilotprojektes mitgeplant wird*. Auftraggebende und Evaluierende sind also in dieser frühen Phase der Projektplanung besonders gefordert. Dies bedingt, dass die Evaluierenden eines Pilotprojektes früh genug bestimmt werden sollten: Sie sollten in der für die spätere Auswertung zentralen Phase des Projektdesigns – in der Planung, wie konkret die Pilotmassnahme implementiert wird – bereits mit einbezogen werden. Dadurch können sie bereits im Voraus auf Punkte der Projektgestaltung hinweisen, die zentral sein werden, um ein klares Evaluationsdesign zu ermöglichen. Ebenso kann so im Vornherein diskutiert werden, ob es realistisch ist, mit den zur Verfügung gestellten Ressourcen und der geplanten Projektgrösse alle von der auftraggebenden Seite gewünschten Aspekte evaluieren zu können. Häufig kann es sinnvoll sein, gemeinsam eine Priorisierung festzulegen, damit die vorhandenen Ressourcen besonders auf jene von der Auftraggeberseite als besonders zentral betrachteten Zielgrössen zu fokussieren. Als Fazit dieser Überlegungen könnte festgehalten werden: Die Qualität der Evaluation entscheidet sich bereits massgeblich zum Zeitpunkt der Projektgestaltung.

Für Auftraggebende und Evaluierende kann es generell hilfreich sein, bei der Projektgestaltung *in Experimenten zu denken*. Die zentrale Frage ist: Welches Politikexperiment gilt es genau durchzuführen, um die Evaluationsfrage, die von Interesse ist, möglichst direkt beantworten zu können? (s.a. Angrist et al. 2009) Wer muss mit wem verglichen werden, damit sich die Wirkung der neuen Politikmassnahme glaubwürdig messen lässt? Die gedankliche Aufstellung des Politikexperiments und die Definition von Programm- und Kontrollgruppe sind auch dann von Nutzen, wenn eine randomisierte Umsetzung des Politikexperiments in der Praxis nicht möglich sein sollte. Diese konzeptionellen Gedanken schärfen die Definition und Planung der Evaluation und damit die Klarheit der Interpretation der späteren Ergebnisse. Zudem ist es auch in Fällen, wo Randomisierung nicht umsetzbar ist, von zentraler Bedeutung, dass eine *Kontrollgruppe* bei der Projektgestaltung mitgeplant wird.

Die Randomisierung der Zuweisung in Programm- und Kontrollgruppe bringt massgebliche Gewinne an Vergleichbarkeit und damit an Klarheit in der Identifikation der Wirkung der Pilotmassnahme.

Literatur

Angrist, Joshua / Pischke, Jörn-Steffen, 2009, Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

Arni, Patrick, 2010, Intensivberatung und -coaching für ältere Stellensuchende – ein Weg zu verbesserten Arbeitsmarktchancen? Ergebnisse der systematischen Wirkungsevaluation, *Schlussbericht* im Auftrag des Amts für Wirtschaft und Arbeit (AWA) des Kantons Aargau, Universität Lausanne.

Arni, Patrick, 2012a, Do Older Job Seekers Find More Jobs When Being Coached? A Field Experiment on Search Behavior, mimeo, IZA Bonn.

Arni, Patrick, 2012b, What's in the Blackbox? A Field Experiment on the Effect of Labor Market Policy on Search Behavior & Beliefs, mimeo, IZA Bonn.

Ashenfelter, Orley, 1987, The Case for Evaluating Training Programs with Randomized Trials, *Economics of Education Review*, 6, S. 333-338.

Baert, Stijn / Cockx, Bart / Gheyle, Niels / Vandamme, Cora, 2012, Do employers discriminate less if vacancies are difficult to fill? Evidence from a field experiment, mimeo, Ghent University, Belgien.

Behaghel, Luc / Crépon, Bruno / Gurgand, Marc, 2012a, Private and Public Provision of Counseling to Job-Seekers: Evidence from a Large Controlled Experiment, mimeo, Paris School of Economics.

Behaghel, Luc / Crépon, Bruno / Le Barbanchon, Thomas, 2012b, Do Anonymous Resumes Make the Field More Even? Evidence from a Randomized Field Experiment, mimeo, Paris School of Economics.

Blundell, Richard / Costa Dias, Monica, 2009, Alternative Approaches to Evaluation in Empirical Microeconomics, *Journal of Human Resources*, 44, S. 565-640.

Card, David / Ibarrarán, Pablo / Villa, Juan Miguel, 2011, Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide, IZA Discussion Paper No. 6085.

Crépon, Bruno / Duflo, Esther / Gurgand, Marc / Rathelot, Roland / Zamora, Philippe, 2012, Do labor market policies have displacement effect? Evidence from a clustered randomized experiment, mimeo, CREST, Paris.

DiNardo, John / Lee, David S., 2011, Program Evaluation and Research Designs, in: Orley Ashenfelter und David Card (Hrsg.), *Handbook of Labor Economics*, North Holland, Elsevier Science, Band 4A, S. 463-536.

Duflo, Esther / Glennerster, Rachel / Kremer, Michael, 2007, Using Randomization in Development Economics Research: A Toolkit, in: T. Paul Schults und John Strauss (Hrsg.), *Handbook of Development Economics*, North Holland, Elsevier Science, 4, S. 3895-62.

Engström, Per/ Hägglund, Pathric / Johansson, Per, 2012, Early interventions and disability insurance: experience from a field experiment, IFAU Working Paper 2012:9, Uppsala, Schweden.

Falk, Armin / Lalive, Rafael / Zweimüller, Josef, 2005, The Success of Job Applications: A New Approach to Program Evaluation, *Labour Economics*, 12 (6), S. 739-748.

Graversen, Brian K. / van Ours, Jan, 2008, How to Help Unemployed Find Jobs Quickly: Experimental Evidence from a Mandatory Activation Program, *Journal of Public Economics*, 92, S. 2020-2035.

Hägglund, Pathric, 2009, Experimental evidence from intensified placement efforts among unemployed in Sweden, IFAU Working Paper 2009:16, Uppsala, Schweden.

Heckman, J. J., 2010, Effective Child Development Strategies, verfasst für: S. Barnett und E. Zigler (Hrsg.), Debates and Issues in Preschool Education (im Erscheinen).

Hendra, R./ Riccio, J.A./ Dorsett, R./ Greenberg, D.H./ Knight, G./ Phillips, J./ Robins, P.K./ Vegeris, S. / Walter, J. mit Hill, A./ Ray, K./ Smith, J., 2011, Breaking the low-pay, no-pay cycle: Final evidence from the UK Employment Retention and Advancement (ERA) demonstration, *Department for Work and Pensions Research Report no. 765*, Corporate Document Services, Sheffield, UK.

Imbens, Guido / Wooldridge, Jeffrey, 2008, Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47, S. 5–86.

Kauer, Lukas / Deuchert, Eva, 2012, Wage subsidies for people with a disability: Helping or hindering? Evidence from a field experiment, mimeo (vorläufige Version), Universität St. Gallen.

Krug, Gerhard / Stephan, Gesine, 2012, Is contracting-out intensified placement services more effective than in-house production? Evidence from a randomized field experiment, mimeo (vorläufige Version), IAB, Nürnberg.

List, John A. / Rasul, Imran, 2011, Field Experiments in Labor Economics, in: Orley Ashenfelter und David Card (Hrsg.), *Handbook of Labor Economics*, North Holland, Elsevier Science, Band 4A, S. 103-228.

Lutz, Peter F. / Sandner, Malte, 2010, Zur Effizienz Früher Hilfen: Forschungsdesign und erste Ergebnisse eines randomisierten kontrollierten Experiments, *Vierteljahreshefte zur Wirtschaftsforschung*, 79 (3), S. 79-97.

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

Müntnich, Michael / Schewe, Torben / Stephan, Gesine, 2010, Durch Zufall zum Erkenntnisgewinn – Emu trifft Pinguin, IAB Forum 2/2010, Nürnberg.

NZFH (Hrsg.), 2010, Frühe Hilfen – Modellprojekte in den Ländern, Nationales Zentrum Frühe Hilfen, Köln.

Rosholm, Michael, 2008, Experimental Evidence on the Nature of the Danish Employment Miracle, IZA Discussion Paper No. 3620.

Rubin, Donald, 1974, Estimating the Causal Effects of Treatments in Randomized and Non-Randomized Studies, *Journal of Educational Psychology*, 66, S. 688-701.

Rubin, Donald, 1977, Assignment to a Treatment Group on the Basis of a Covariate, *Journal of Educational Statistics*, 2, S. 1-26.

Schneider, Hilmar / Uhlendorff, Arne / Zimmermann, Klaus F., 2011, Ökonometrie vs. Projektdesign: Lehren aus der Evaluation eines Modellprojekts zur Umsetzung des Workfare-Konzepts, IZA Discussion Paper No. 5599.

Schwerdt, Guido / Messer, Dolores / Woessmann, Ludger / Wolter, Stefan C., 2012, Effects of Adult Education Vouchers on the Labor Market: Evidence from a Randomized Field Experiment, *Journal of Public Economics*, 2012, 96 (7-8), S. 569-583.

U.S. Department of Education / Institute of Education Sciences / National Center for Education Evaluation and Regional Assistance, 2003, Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide, Washington, D.C.

Van den Berg, Gerard J. / van der Klaauw, Bas, 2006, Counseling and monitoring of unemployed workers: theory and evidence from a controlled social experiment, *International Economic Review*, 47, S. 895–936.

White, Howard, 2006, Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank, World Bank, Washington, D.C.

Widmer, Thomas / de Rocchi, Thomas, 2012, Evaluation – Grundlagen, Ansätze und Anwendungen, Rüegger Verlag, Zürich.

Endnoten:

[1] Dieser Ansatz beschreibt die Wirkungsmessung im Rahmen von potenziellen Outcomes: D.h., einerseits, wie das Ergebnis bzw. die Zielgrösse ausgesehen hätte, wenn die Politikmassnahme angewandt worden wäre (treated outcome); und, andererseits, wie das Ergebnis ausgesehen hätte, wenn die Massnahme nicht zum Einsatz gekommen wäre (non-treated outcome). Die Differenz der beiden potenziellen Outcomes entspricht der Wirkung der Massnahme (treatment effect). In der Realität werden nie beide potenziellen Outcomes gleichzeitig beobachtet. Dieses fundamentale Problem der Evaluation muss durch geeignete Evaluationsdesigns und statistische Methoden (so gut wie möglich) gelöst werden. Der aus statistischer Sicht sauberste Ansatz hierzu ist der Einsatz von Randomisierung. Kapitel 2 diskutiert näher, unter welchen Bedingungen dieser möglich ist.

[2] Ein anderer Strang sozialwissenschaftlicher Literatur diskutiert die Frage der Ermittlung der Wirkung von spezifischen Programm-Interventionen unter dem Begriff „impact evaluation". Verschiedene Definitionen dieses Begriffs werden diskutiert. Ein aktueller Bericht der Weltbank (White 2006) definiert ihn als die „counterfactual analysis of the impact of an intervention on final welfare outcomes". Diese Definition ist eng verwandt mit dem im Haupttext verwendeten Begriff der kausalen Programm-Evaluation. Die konzeptionelle Begriffsdiskussion liegt nicht im Fokus dieses Beitrages und wird daher nicht weiter detailliert.

[3] In der englischsprachigen ökonomisch-statistischen Literatur werden die Begriffe *„randomized controlled trial"* und *„field experiment"* meist synonym benutzt.

[4] Eine wichtige Plattform für entwicklungsökonomische Anwendungen und methodische Fragen ist das Poverty Action Lab: http://www.povertyactionlab.org/. Zwei wichtige nicht-akademische non-profit Kompetenzzentren in der Anwendung von randomisierten Designs sind die MDRC (http://www.mdrc.org/) in den USA und die SRDC (http://www.srdc.org/) in Kanada.

[5] Der Pilotprojekt-Status ist aber, aus methodischer Sicht, keine Notwendigkeit: Randomisierung bzw. Feldexperimente, wie sie in den nächsten Kapiteln weiter beleuchtet werden, können auch ohne die explizite Aufsetzung eines Pilotversuches durchgeführt werden. Notwendig sind lediglich die Schaffung einer Kontrollgruppe sowie der Möglichkeit der Zufallszuweisung. Entsprechend sind auch die in Kapitel 4 diskutierten methodischen Praxisfragen relevant sowohl für Fälle mit wie auch für solche ohne Vorhandensein eines Pilotprojekts.

[6] In der Schweiz schreibt die Bundesverfassung (Art. 170) vor, dass „Massnahmen des Bundes auf ihre Wirksamkeit überprüft werden" sollen. Dies wird in vielen Politikbereichen noch spezifisch gesetzlich festgehalten, so zum Beispiel bei der Arbeitslosenversicherung: siehe Art. 59a und 73a AVIG und Art. 122b AVIV. Eine umfassende Übersicht von Evaluationsklauseln im Bundesrecht findet sich unter http://www.bj.admin.ch/content/bj/de/home/themen/staat_und_buerger/evaluation/materialien_/uebersicht.html .

[7] Wie in Endnote 1 ausgeführt, zeigt das Rubin-Modell anhand potenzieller Outcomes auf, welche statistische Bedingungen nötig sind für eine möglichst unverzerrte Wirkungsmessung. Ziel ist die Messung des Programmeffekts (treatment effect) unter Bedingungen, wo der Selektionseffekt (siehe Haupttext) möglichst ausgeschlossen ist. Das Selektionsproblem ist gelöst, wenn die Zuweisung zu einer Intervention *unabhängig* ist von den potenziellen Outcomes (siehe z.B. Angrist et al. 2009). D.h. konkret, dass die gemessene Wirkung einer Intervention – beispielsweise der Zuweisung zu einem Bildungsprogramm – nicht beeinflusst ist von den Eigenschaften der Personen, die die Intervention aufnehmen. Die gemessene Wirkung der Intervention ist dieselbe, ob nun die Person(engruppe) A oder die Person(engruppe) B dem Bildungsprogramm zugewiesen werden (und die jeweils anderen Personen der Kontrollgruppe). Bei Zufallszuweisung ist dies der Fall und das Selektionsproblem damit gelöst: Es wird nur die kausale Programmwirkung gemessen, Selektionseffekte treten nicht auf.

[8] Public Law 107-279, Nov. 5, 2002, H.R. 3801

[9] Z.B. Card et al. (2011), Table 1, zeigen auf, von welcher Grösse die Population in Programm- oder Kontrollgruppe sein müsste, damit ein Effekt einer gewissen Grösse statistisch signifikant wird.

[10] Hier kann wiederum die Medizin als Vergleichsbeispiel herangezogen werden: Die Medikamentenzulassungsstellen würden keine Medizin zulassen, die nicht vorher in einem kontrollierten, beschränkten Rahmen getestet wurde. Auch da kann den Probandinnen und Probanden nicht mit Sicherheit gesagt werden, ob ihnen ihre Teilnahme am Versuch nützt oder schadet.

[11] Ein frühes Beispiel aus dem Bereich Weiterbildung ist Ashenfelter (1987); Meyer (1995) fasst eine Serie von Feldexperimenten in den frühen 90er-Jahren zusammen. Aktuell führen in den USA häufig Institutionen wie die erwähnte MDRC Randomisierungen in diesem Bereich durch.

[12] Van der Klaauw, Bas und SEOR/Regioplan: Experimentelle Untersuchungen der Effektivität arbeitsmarktlicher Reintegrationsmassnahmen (Originaltitel in Niederländisch), Projektbeginn: 2012, durchführende Institutionen: VU University Amsterdam und SEOR/Regioplan, Auftraggeber: Niederländisches Ministerium für Soziales und Arbeit.

[13] Müntnich et al. (2010) stellen das erwähnte Feldexperiment vor und diskutieren anhanddessen die Rolle von randomisierten Pilotprojekten für die deutsche Arbeitsmarktpolitik-Evaluation und kommen zum Schluss: „Gezielt eingesetzte Modellversuche mit einer zufallsgesteuerten Personenzuweisung können ergänzend wichtige Informationen zur Wirksamkeit arbeitsmarktpolitischer Massnahmen bereitstellen. (...) 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."

[14] Arni, Patrick, Bonoli, Giuliano, Lalive, Rafael und Daniel Oesch: L'impact des réseaux sociaux sur le retour à l'emploi des chômeurs (SOCNET), Projektbeginn: 2011, durchführende Institutionen: Universität Lausanne und IZA, Projektträger: Nationaler Forschungsschwerpunkt (SNF) LIVES (www.lives-nccr.ch) sowie 3. Evaluationswelle der Arbeitlosenversicherung, Staatssekretariat für Wirtschaft Seco, Bern.

[15] Arni, Patrick, Wunsch, Conny: Die Rolle der Erwartungshaltungen in der Stellensuche und der RAV-Beratung, Projektbeginn: 2011, durchführende Institutionen: IZA und VU University Amsterdam, Auftraggeber: 3. Evaluationswelle der Arbeitlosenversicherung, Staatssekretariat für Wirtschaft Seco, Bern

[16] Stephan, Gesine, Homrighausen, Pia und Gerard J. van den Berg: Info-Treatment Entgeltsicherung für ältere Arbeitnehmer, Projektbeginn: 2011, durchführende Institutionen: IAB und Universität Mannheim, Auftraggeber: IAB, Nürnberg, Projektdetails: http://www.iab.de/138/section.aspx/Projektdetails/k120702323 (Okt. 2012).

[17] Altmann, Steffen, Falk, Armin, Koch, Susanne und Florian Zimmermann: Sanfte Politikmassnahmen zur Erhöhung der Suchanstrengung von Arbeitslosen, Projektbeginn: 2010, durchführende Institutionen: IZA, Universität Bonn und IAB, Auftraggeber: IAB, Nürnberg.

[18] Die Studie findet auch, dass Bewerberinnen und Bewerber mit Migrationshintergrund oder mit Adresse in einer sozial benachteiligten Nachbarschaft im anonymisierten Fall weniger oft zu Bewerbungsgesprächen eingeladen wurden als im nicht-anonymisierten Fall. Die Gründe für dieses unerwartete Ergebnis werden aktuell noch untersucht.

[19] Degen, Kathrin, Efferson, Charles, Goette, Lorenz und Rafael Lalive: Smart Metering, Expert Advice, and Social Comparison: What Matters for Energy Consumption?, Projektbeginn: 2010, durchführende Institutionen: Universitäten Lausanne und Zürich, Auftraggeber: ewz Zürich und Bundesamt für Energie, Bern.

[20] In beiden Fällen muss auch die Frage gestellt werden, ob die so selektierte Teilnehmerpopulation „repräsentativ" ist für die Population von Interesse des Auftraggebers. Diese Frage ist Teil der Diskussion der externen Validität (siehe weiter unten).

[21] Im Prinzip ist auch denkbar, dass die beteiligten Personen die Auswirkung eines gewissen Interventionsschrittes antizipieren, sodass messbare Effekte schon vor dem effektiven Eintritt des Interventionsschrittes auftreten. Diese Thematik der Antizipationswirkungen von Programmen ist jenseits des Fokus dieses Beitrages und wird daher nicht weiter diskutiert.

[22] Ob dieser Vergleich ohne oder mit Selektionskorrektur vorgenommen werden muss, hängt von der Natur des Treatment-Aufnahme-Prozesses ab: Falls die Entscheidung, an der Treatment-Intervention teilzunehmen, unabhängig ist von Faktoren, die auch das Outcome der Person mitbestimmen, dann führt der erwähnte Vergleich zu unverzerrten Ergebnissen. Dies wäre z.B. der Fall, wenn die Treatment-Nichtaufnahme mit einem Wetterereignis zu tun hat. Zu korrigierende Selektion kann hingegen auftreten, falls die Entscheidung zur Treatment-Aufnahme von Eigenschaften getrieben ist, die auch das Outcome beeinflussen und die nicht direkt durch beobachtbare Informationen statistisch kontrolliert werden können. Eine solche Eigenschaft könnte z.B. unbeobachtete Motivation sein.

[23] Die in Kapitel 3 anhand der Studie von Crépon et al. (2012) vorgestellte zweistufige Randomisierung, bei der zusätzlich noch eine Zufallszuteilung auf Ebene der Regionen stattfindet, wäre ein Evaluations-Design, das „spill over"-Effekte explizit identifizieren und untersuchen könnte.

[24] Auch eine nicht-randomisierte Studie kann einen Hawthorne-Effekt auslösen, wenn sie mit begleitenden Befragungen arbeitet. Der Hawthorne-Effekt kann nur ausgeschlossen werden im Falle einer ex-post Evaluation mit retrospektiver Befragung.