STANDPUNKTE

IZA Standpunkte Nr. 69

# Ökonometrische Methoden zur Evaluierung kausaler Effekte der Wirtschaftspolitik

Franziska Kugler
Guido Schwerdt
Ludger Wößmann

April 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

# Ökonometrische Methoden zur Evaluierung kausaler Effekte der Wirtschaftspolitik

**Franziska Kugler**
*ifo Institut an der Universität München*

**Guido Schwerdt**
*Universität Konstanz*
*und IZA*

**Ludger Wößmann**
*Ludwig-Maximilians-Universität München, ifo Institut*
*und IZA*

IZA Standpunkte Nr. 69
April 2014

IZA

Postfach 7240
53072 Bonn

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

IZA Standpunkte Nr. 69
April 2014

# ZUSAMMENFASSUNG

# Ökonometrische Methoden zur Evaluierung kausaler Effekte der Wirtschaftspolitik*

Die Öffentlichkeit hat ein Interesse zu wissen, ob politische Maßnahmen die mit ihnen verfolgten Ziele wirksam und wirtschaftlich erreichen. In empirischen Studien, die als Grundlage für evidenzbasierte Wirtschaftspolitik dienen können, werden häufig komplexe Methoden verwendet, um tatsächliche kausale Wirkungen von anderweitig verursachten Zusammenhängen zu unterscheiden. Der vorliegende Beitrag gibt einen nicht-technischen Überblick über Intuition und Anwendungsbeispiele des modernen wissenschaftlichen Instrumentariums zur Evaluierung kausaler Effekte. Die betrachteten ökonometrischen Methoden umfassen kontrolliert randomisierte Experimente, die Zufallsvergabe überzeichneter Programme, den Instrumentvariablen-Ansatz, den Regressions-Diskontinuitäten-Ansatz, den Differenzen-in-Differenzen-Ansatz sowie Panelmethoden mit fixen Effekten. Diese experimentellen und quasi-experimentellen Designs sollen der Wirtschaftspolitik helfen zu lernen, was funktioniert.

JEL-Codes: C1, C2, D04, J38, D78

Schlagworte: Evaluierung, ökonometrische Methoden, kausale Effekte, Wirtschaftspolitik, Experimente, quasi-experimentelle Methode, Instrumentvariablen-Ansatz, Regressions-Diskontinuitäten-Ansatz, Differenzen-in-Differenzen-Ansatz, Panelmethoden, fixe Effekte

Kontaktadresse:

Ludger Wößmann
ifo Institut
ifo Zentrum für Bildungs- und Innovationsökonomik
Poschingerstraße 5
81679 München
E-mail: woessmann@ifo.de

* Dieser Beitrag wurde für die Zeitschrift *Perspektiven der Wirtschaftspolitik* verfasst. Wir danken Martin Schlotter für die Zusammenarbeit an einem Vorgängerprojekt, Sascha Becker, Oliver Falck und Gabriel Felbermayr für nützliche Hinweise und Karen Horn und Karl-Heinz Paqué für redaktionelle Überarbeitungen.

# Ökonometrische Methoden zur Evaluierung kausaler Effekte der Wirtschaftspolitik

## 1. Evidenzbasierte Politik: Wie lernen wir, was funktioniert?

Die Notwendigkeit der „Evidenzbasierung" politischer Maßnahmen ist in aller Munde. Politiker, Verwaltungsangehörige, Wissenschaftler, Medien, die interessierte Öffentlichkeit – sie alle wollen wissen, ob politische Maßnahmen die mit ihnen verfolgten Ziele bestmöglich erreichen. Nicht nur in Zeiten knapper Kassen ist den Bürgern an einem wirksamen und wirtschaftlichen Einsatz der Steuergelder gelegen. Gleichwohl herrscht vielfach Unwissen über die tatsächlichen Wirkungen politischer Maßnahmen.[1] Dabei ist aus anderen Ländern bekannt, dass politische Maßnahmen durchaus nicht immer angemessene Wirkungen hervorbringen und dass es häufig zu Mitnahmeeffekten kommen kann, bei denen öffentliche Fördermittel Ausgaben ersetzen, die ansonsten privat getätigt würden. Außerdem ist die Untersuchung der Wirtschaftlichkeit politischer Maßnahmen auch gesetzlich vorgeschrieben: Basierend auf der schon im Grundgesetz (Artikel 114 Abs. 2) verankerten Prüfung der Wirtschaftlichkeit der Haushalts- und Wirtschaftsführung, schreibt die Bundeshaushaltsordnung (§ 7 Abs. 2) vor, dass für alle finanzwirksamen Maßnahmen angemessene Wirtschaftlichkeitsuntersuchungen durchzuführen sind. Als Instrument der Erfolgskontrolle sollen diese gemäß der einschlägigen Verwaltungsvorschrift feststellen, „ob und in welchem Ausmaß die angestrebten Ziele erreicht wurden, ob die Maßnahme ursächlich für die Zielerreichung war und ob die Maßnahme wirtschaftlich war" (Präsident des Bundesrechnungshofes 2013, S. 100).

Trotz der zunehmenden Verfügbarkeit umfangreicher Mikrodaten, die für eine überzeugende Evaluierung zumeist essentiell sind, ist es keine leichte Aufgabe, die Ursächlichkeit von Effekten empirisch zu belegen. Eine kausale Interpretation muss jedoch möglich sein, wenn Evidenz zur Fundierung politischer Entscheidungen geeignet sein soll. Zur Evaluierung kausaler Effekte politischer Maßnahmen ist in den vergangenen Jahrzehnten ein wissenschaftliches Instrumentarium entwickelt worden, das darauf zielt, die tatsächliche Kausalität von anderweitig verursachten empirischen Zusammenhängen zu trennen. Es gilt mittlerweile als hinreichend „ausgereift" (Imbens und Wooldridge 2009, S. 76), um sich in der praktischen Anwendung als nützlich zu erweisen. Angrist und Pischke (2010) sprechen sogar von einer „Glaubwürdigkeitsrevolution in den empirischen Wirtschaftswissenschaften".

Da die entsprechenden Forschungsmethoden aufgrund ihrer technischen Komplexität nicht immer leicht verständlich sind, möchten wir im vorliegenden Beitrag einen nicht-technischen

[1] Zur Evaluierung wirtschaftspolitischer Fördermaßnahmen als Element einer evidenzbasierten Wirtschaftspolitik siehe das aktuelle Gutachten des Wissenschaftlichen Beirats beim Bundesministerium für Wirtschaft und Energie (2013).

Überblick über die von Ökonomen zunehmend in der empirisch-kausalanalytischen Wirkungsforschung angewendeten Methoden geben. Dabei sollen neben grundlegender Idee und Intuition auch deren Stärken und Schwächen zutage treten. Zur Vertiefung sei auf die einschlägige Fachliteratur verwiesen.[2]

Eine Reihe von Anwendungsbeispielen in Deutschland liefert die umfangreiche Evaluierung der aktiven Arbeitsmarktpolitik der vergangenen Jahre (Bundesministerium für Arbeit und Soziales und Institut für Arbeitsmarkt- und Berufsforschung 2011). Diese Evaluierung ist möglich geworden, weil es wesentliche Fortschritte in der Aufbereitung und Verfügbarkeit von Mikrodaten im Forschungsdatenzentrum der Bundesagentur für Arbeit gegeben hat (Allmendinger und Kohlmann 2005) und weil die zeitnahe Evaluationsforschung inzwischen auch explizit gesetzlich verankert ist (Brinkmann, Hujer und Koch 2006). Aufgrund der Evaluationsbefunde kam es zu einer grundlegenden Neuausrichtung der arbeitsmarktpolitischen Instrumente. Beispielsweise führten entsprechende Evaluierungsergebnisse dazu, dass Arbeitsbeschaffungsmaßnahmen abgeschafft und wirksame Instrumente weiterentwickelt wurden. Kürzlich wurden auch im Rahmen der Gesamtevaluierung der ehe- und familienbezogenen Leistungen verschiedene empirisch-kausalanalytische Wirkungsstudien durchgeführt.[3] Im vorliegenden Beitrag verweisen wir auf Beispielstudien aus der gesamten Breite wirtschaftspolitischer Maßnahmen, die neben Arbeitsmarkt und Familie auch Themen wie Industrie- und Technologiepolitik, Investitionsförderung, Regionalpolitik, Exportförderung, Gründung und Innovation, Steuern, Soziales, Bildung, Gesundheit, Umwelt und einiges mehr abdecken. Dabei handelt es sich um eine eher eklektische Auswahl von Beispielstudien aus dem In- und Ausland, welche die Breite der Anwendungsmöglichkeiten verdeutlichen soll und keinesfalls Anspruch auf umfassende Abdeckung oder auch nur Ausgewogenheit der ausgewählten Beispiele erheben kann.

Abschnitt 2 dieser Abhandlung dient der Einführung in das zentrale Problem der Evaluierung kausaler Effekte. Darauf aufbauend werden sechs Gruppen von Evaluationsmethoden vorgestellt.[4] Wir beginnen mit zwei Methoden, die auf expliziter Randomisierung basieren – kontrolliert randomisierte Experimente (Abschnitt 3.1) und zufällige Auslosungen bei über-

[2] Einschlägige einführende Lehrbücher bieten beispielsweise Stock und Watson (2011) und, in deutscher Sprache, Bauer, Fertig und Schmidt (2009). Für weiterführende methodische und technische Details vgl. Angrist und Pischke (2009), Angrist und Krueger (1999), DiNardo und Lee (2011), Imbens und Wooldridge (2009) und Manski (1995). Eine Erörterung der Möglichkeiten und Grenzen experimenteller und quasi-experimenteller Evaluierungsmethoden für die Politikanalyse bieten Heckman (2010), Imbens (2010) und Deaton (2010).

[3] Siehe http://www.bmfsfj.de/BMFSFJ/familie,did=195944.htm (14.4.2014).

[4] Grundstruktur und Intuitionsdarstellung basieren zum Teil auf Schlotter, Schwerdt und Wößmann (2011), die zahlreiche Anwendungsbeispiele aus der Bildungspolitik aufgreifen.

zeichneten Programmen (Abschnitt 3.2). Danach beschäftigen wir uns mit zwei Methoden, die darauf zielen, die experimentelle Umgebung mit nicht-experimentell erhobenen Daten nachzuahmen. In diesem Zusammenhang wird von natürlichen Experimenten oder quasi-experimentellen Verfahren gesprochen. Dazu gehören der Instrumentvariablen-Ansatz (Abschnitt 4.1) und der Regressions-Diskontinuitäten-Ansatz (Abschnitt 4.2). Zum Schluss werden Paneldaten-Methoden vorgestellt, mit denen das Ziel verfolgt wird, mögliche Verzerrungen bei Analysen von Beobachtungsdaten zu vermeiden – der Differenzen-in-Differenzen-Ansatz (Abschnitt 5.1) und Panelmethoden mit fixen Effekten (Abschnitt 5.2).

## 2. Die Herausforderung der Evaluierung kausaler Effekte

Die Evaluationsforschung zielt auf den Zusammenhang zwischen bestimmten Politikmaßnahmen und den damit verfolgten Zielen. Der Zweck von Evaluierungsstudien ist es herauszufinden, ob sich die „Beobachtungseinheiten“ – das können einzelne Personen, Haushalte, Unternehmen oder Landkreise sein – aufgrund der Teilnahme an der Politikmaßnahme oder der „Behandlung“ durch den politischen Eingriff im Hinblick auf die festgelegten Zielgrößen oder Ergebnisvariablen verbessern.

### *2.1 Das Thema: Von Korrelation zu Kausalität*

Mit Hilfe üblicher statistischer Methoden lässt sich recht einfach feststellen, ob zwischen zwei Begebenheiten ein Zusammenhang besteht, ob beispielsweise zwischen einer Politikmaßnahme und dem damit beabsichtigten Ziel ein Zusammenhang besteht. Eine wesentlich weiter gehende Frage ist jedoch, ob dieser statistische Zusammenhang – die Korrelation – auch als der kausale Effekt der Politikmaßnahme auf die Ergebnisgröße interpretiert werden kann. Von einem kausalen oder ursächlichen Effekt wird gesprochen, wenn sich die Ergebnisgröße allein aufgrund der Maßnahmenteilnahme verändert. Das Problem hierbei ist, dass es auch andere Gründe geben kann, die ursächlich für den Zusammenhang zwischen Maßnahmenteilnahme und Zielgröße sind.

Wenn zum Beispiel Unternehmen mit vergleichsweise schlechtem Umsatz- und Beschäftigungserfolg durch eine politische Maßnahme gefördert werden, dürfte sich über verschiedene Unternehmen hinweg ein negativer Zusammenhang zwischen Umsatz- und Beschäftigungserfolg und der Teilnahme an der Fördermaßnahme finden. Aber der Grund für diesen negativen Zusammenhang liegt darin, dass das Unternehmen an der Maßnahme teilnimmt, weil es sich in einer schlechten Lage befindet – nicht andersherum. Dies ist ein Beispiel für

„umgekehrte Kausalität“, bei der die interessierende Ergebnisvariable in Wirklichkeit einen kausalen Effekt auf die interessierende Behandlungsvariable hat.

Ein weiterer Fall, in dem die Maßnahmenteilnahme nicht ursächlich für den beobachteten Zusammenhang ist, ist die Problematik „ausgelassener Variablen“. In diesem Fall beeinflusst eine nicht im Modell berücksichtigte Variable sowohl die Behandlungsvariable als auch die Ergebnisvariable. Wenn beispielsweise einige Unternehmer erfinderischer und dynamischer – oder kurz: erfolgreicher – sind als andere, obwohl die Firmen ansonsten bisher vergleichbar waren, dann sind sie vermutlich auch eher dabei erfolgreich, sich um die Teilnahme an einer Fördermaßnahme zu bewerben. Sie fördern ihren Unternehmenserfolg aber auch in anderer Hinsicht erfolgreicher, sei es durch Produkt- und Prozessinnovationen, bessere Personalpolitik oder umsichtigere strategische Entscheidungen. Wenn diese Entscheidungen den künftigen Unternehmenserfolg verbessern, kommt es zu einem positiven Zusammenhang zwischen Maßnahmenteilnahme und Unternehmenserfolg – selbst dann, wenn sich die Teilnahme an der Maßnahme an sich überhaupt nicht kausal auf den Unternehmenserfolg auswirkt. Die ausgelassene Variable des erfolgreichen Unternehmertums ist für den Zusammenhang verantwortlich. Wenn eine solche „unbeobachtete Heterogenität“ zwischen von einer Maßnahme Betroffenen und Nicht-Betroffenen im Hinblick auf Eigenschaften besteht, die sowohl mit der Behandlungs- als auch mit der Ergebnisvariable zusammenhängen, wird die Identifikation von kausalen Effekten erschwert.[5]

Immer dann, wenn andere Gründe bestehen, die die Korrelation zwischen den zwei interessierenden Größen – der Maßnahmenteilnahme und der Ergebnisgröße – entstehen lassen, kann der gesamte Zusammenhang nicht als kausaler Effekt der Behandlung auf das Ergebnis interpretiert werden. Ökonomen sprechen hier von „Endogenitätsproblemen“. Die Teilnahme an der Maßnahme bzw. die Behandlung durch die Intervention kann in dem betrachteten Modell nicht als „exogen“, also „von außen kommend“ angesehen werden, sondern hängt endogen von den im Modell relevanten Variablen selbst ab. Diese endogene Bestimmung der Maßnahmenteilnahme hängt entweder von der Ergebnisgröße selbst ab oder wird zusammen mit der Ergebnisvariable durch einen weiteren Faktor bestimmt. Aufgrund des Endogeni-

[5] Weitere mögliche Ursachen von Endogenität, die sich ebenfalls unter dem Oberbegriff ausgelassene Variablen oder unbeobachtete Heterogenität betrachten lassen, sind Selbst-Selektion (Beobachtungssubjekte mit verschiedenen Eigenschaften können selbst entscheiden, ob sie an der Maßnahme teilnehmen) und Simultanität (die Maßnahmenteilnahme und die Ergebnisgröße sind Entscheidungsvariablen, die gemeinsam bestimmt werden). Ökonometrisch betrachtet können auch Messfehler in der Behandlungsvariable als Endogenitätsproblem aufgefasst werden, weil dadurch ein Zusammenhang zwischen Behandlungsvariable und Ergebnisvariable entsteht. Dieser Zusammenhang verzerrt die Schätzung im Allgemeinen (im Falle eines klassischen Messfehlers) dahingehend, dass kein Effekt identifiziert werden kann, obwohl er vorhanden ist.

tätsproblems spiegeln Schätzungen des Zusammenhangs zwischen Behandlungs- und Ergebnisvariable, die auf reinen Korrelationen basieren, den tatsächlichen kausalen Effekt der Maßnahmenteilnahme auf die Zielgröße nur verzerrt wider.

### *2.2 Grenzen herkömmlicher Methoden zur Kontrolle beobachteter Einflüsse*

In herkömmlichen Ansätzen versucht man, Endogenitätsprobleme zu beheben, indem die oben genannten alternativen Ursachen der Korrelation beobachtet und die Unterschiede in der Ergebnisgröße, die auf diese anderen beobachteten Faktoren zurückzuführen sind, herausgerechnet werden. So geht man in multivariaten Modellen vor, die den Effekt vieler Variablen auf die Ergebnisgröße gleichzeitig zu schätzen erlauben, wie die klassische Methode der Kleinsten Quadrate. Diese Methoden ermöglichen es, den Zusammenhang zwischen Behandlungs- und Ergebnisvariable zu schätzen – nachdem deren Zusammenhänge mit anderen beobachteten Variablen herausgerechnet wurden. Wenn sich in dem Fördermaßnahmen-Beispiel von oben die ausgelassenen Variablen vollständig beobachten lassen, so können wir diese einfach in das multivariate Modell aufnehmen. Die störende Beeinflussung wird damit ausgeschaltet. In den meisten Fällen können wir die ausgelassene Variable jedoch nicht vollständig beobachten. So wäre in unserem Beispiel eine perfekte Messung der unternehmerischen Fähigkeiten kaum möglich. Solange ein Teil der ausgelassenen Variable unbeobachtet bleibt, ist der geschätzte bedingte Zusammenhang nicht notwendigerweise für eine kausale Interpretation geeignet.

Es wird deutlich, dass die klassischen Methoden immer mit großer Vorsicht zu interpretieren sind, wenn relevante Faktoren, die mit der Maßnahmenteilnahme und der Zielgröße zusammenhängen, unbeobachtet bleiben. In den vergangenen Jahrzehnten hat sich in der Literatur, die Politikmaßnahmen evaluiert, immer deutlicher herausgestellt, dass es tatsächlich unzählige wichtige Faktoren gibt, die in den empirischen Modellen unbeobachtet bleiben. In den Studien wird dann oft der vergebliche Versuch unternommen, alle relevanten Störfaktoren zu berücksichtigen. Man denke zum Beispiel an Faktoren wie die Managementkompetenz, bestimmte Investitionsentscheidungen, motivierende Anreize für Mitarbeiter, Entscheidungen eng konkurrierender Unternehmen, Veränderungen in der Weltmarktnachfrage nach spezifischen Produkten und vieles mehr. Selbst wenn es gelingt, beobachtbare Maßzahlen bestimmter Dimensionen dieser Faktoren zu erhalten, können andere – oft wichtige – Dimensionen unbeobachtet bleiben. Zudem kann selbst die Berücksichtigung aller beobachtbaren Faktoren nicht das Problem der Endogenität lösen, wenn diese ihren Ursprung in einer umgekehrten Kausalität hat, wenn also die Zielgröße die Maßnahmenteilnahme verursacht.

Dieselben Einwände, die gegen die herkömmlichen Methoden vorgebracht werden, gelten auch für eine andere Technik, die in jüngster Zeit immer populärer geworden ist, nämlich Matching-Verfahren. Die zentrale Idee dabei ist es, zusammenpassende Paare von Maßnahmenteilnehmern und Nicht-Teilnehmern zu finden, die einander (vor der Maßnahmenteilnahme) so weit wie möglich in beobachtbaren Eigenschaften ähneln. Unter bestimmten Annahmen kann diese Methode auch eine mögliche Verzerrung der Schätzung des Maßnahmeneffektes reduzieren. Solange jedoch relevante Eigenschaften unbeobachtet bleiben, kann die Verzerrung nicht vollständig beseitigt werden. In diesem Sinne können Matching-Verfahren nicht das Problem der Endogenität lösen und sind genauso von Verzerrungen aufgrund unbeobachteter Faktoren betroffen wie die herkömmlichen Methoden.[6]

### *2.3 Ausnutzung exogener Variation*

Über die herkömmlichen Methoden hinaus sind in den vergangenen Jahrzehnten moderne Methoden zur Evaluierung kausaler Effekte entwickelt und von der angewandten Wirtschaftsforschung zunehmend auch eingesetzt worden. Ziel dieser Verfahren ist es, eine überzeugendere Identifikation von kausalen Effekten trotz unbeobachteter Störfaktoren zu ermöglichen. Hierbei wird „exogene“ Variation genutzt – also eine Variation in der Maßnahmenteilnahme, die nicht mit Faktoren zusammenhängt, die mit der Ergebnisgröße verbunden sind.

Beispielsweise wird in medizinischen Studien zur Erprobung der Wirksamkeit neuer Medikamente nur ein Teil der Patienten tatsächlich behandelt, der andere Teil erhält ein Placebo. Die Zuordnung in die Behandlungs- und die Placebo-Gruppe erfolgt dabei zufällig, um sicherzustellen, dass die Aufteilung nicht durch andere (Stör-)Faktoren beeinflusst wird. Die nicht-behandelten Patienten werden dann als sogenannte Vergleichs- oder Kontrollgruppe verwendet, mit der die behandelten Patienten verglichen werden. Das Ziel der ökonometrischen Methoden ist es, diese Art des experimentellen Designs nachzuahmen, wobei oftmals Daten verwendet werden, die gar nicht aus einem expliziten Experiment stammen. Man bildet eine Behandlungsgruppe (die an der Maßnahme teilnimmt oder von einer Intervention betroffen ist) und eine Kontrollgruppe (die nicht teilnimmt oder nicht betroffen ist), die exakt gleich sind. Die Beobachtungseinheiten sollten also nicht aufgrund von Begebenheiten oder Eigenschaften, die mit der interessierenden Zielgröße zusammenhängen, der jeweiligen Grup-

[6] Gleichwohl können Matching-Verfahren die Bildung geeigneter Behandlungs- und Kontrollgruppen verbessern, wenn sie mit einem der Ansätze kombiniert werden, die im Folgenden erörtert werden. Letztere stellen dann die Exogenität der Maßnahmenteilnahme sicher und ermöglichen somit eine kausale Interpretation. So stellen Matching-Verfahren, die Werte der Zielgröße berücksichtigen, die vor Einführung der Maßnahme beobachtet wurden, einen Spezialfall der hier diskutierten Panelmethoden dar.

pe zugewiesen worden sein. Die zentrale Herausforderung besteht darin zu schätzen, was in der kontrafaktischen Situation geschehen wäre, also welches Ergebnis ein Teilnehmer erreicht hätte, wenn er *nicht* Teilnehmer an der zu evaluierenden Maßnahme gewesen wäre.[7]

Die zentrale Idee, die den modernen ökonometrischen Methoden zur Evaluierung kausaler Effekte zugrunde liegt, ist folgende: Wenn die Aufteilung der Bevölkerung in Behandlungs- und Kontrollgruppe rein zufällig erfolgt und eine ausreichend große Anzahl von Einheiten beobachtet wird, dann stellt die Zufälligkeit sicher, dass sich die zwei Gruppen nicht systematisch in anderen Dimensionen als der Gruppenzuordnung unterscheiden. In der Tat sorgt das mathematische Gesetz der großen Zahlen dafür, dass die Merkmale der Beobachtungseinheiten in der Behandlungsgruppe dieselben sind wie in der Kontrollgruppe. Folglich kann der kausale Effekt der Maßnahmenteilnahme auf das Ergebnis direkt beobachtet werden, indem die durchschnittlichen Ergebnisse der Behandlungs- und der Kontrollgruppe verglichen werden, weil sich die zwei Gruppen lediglich in Bezug auf die Maßnahmenteilnahme unterscheiden. Das Ziel der Evaluationsmethoden, die im vorliegenden Beitrag erläutert werden, ist es, solche eindeutigen Behandlungs- und Kontrollgruppen zu bilden und somit auszuschließen, dass Schätzungen des Effektes der Maßnahmenteilnahme durch unbeobachtete Unterschiede verzerrt werden.

## 3. Explizite Randomisierung: Den Würfel entscheiden lassen

Wir beginnen mit zwei Methoden, die explizite Randomisierung – also eine rein zufällige Aufteilung (wie beim Würfeln) auf Teilnahme bzw. Nicht-Teilnahme an einer Intervention – nutzen, um Behandlungs- und Kontrollgruppen zu bilden, die sich weder in beobachteten noch in nicht-beobachteten Eigenschaften voneinander unterscheiden.

### *3.1 Kontrolliert randomisierte Experimente: Die „ideale" Welt*

Der erste Ansatz besteht darin, ein kontrolliert zufälliges Experiment (häufig auch als „RCT" für Randomized Controlled Trial bezeichnet) durchzuführen.[8] Das entscheidende Merkmal von kontrollierten Experimenten ist die zufällige Zuordnung der Teilnehmer in Behandlungs-

[7] Die Idee der kontrafaktischen oder potentiellen Ergebnisse ist das fundamentale Konzept moderner Evaluationsstudien. Der Grundstein für das heutige Verständnis kausaler Wirkungsmechanismen, das der Identifikation kausaler Effekte zugrunde liegt, wurde in Beiträgen von Rubin (1974, 1977) und Holland (1986) gelegt.

[8] Grundlagen und weiterführende Diskussionen zum Einsatz von kontrolliert randomisierten Experimenten in der angewandten Wirtschaftsforschung bieten Harrison und List (2004), Banerjee und Duflo (2009), List (2011) und Ludwig, Kling und Mullainathan (2011). Arni (2012) erörtert die Nutzung von Randomisierung zur kausalen Evaluierung von Pilotprojekten in der Praxis. Eine politikorientierte Einführung bieten Haynes u.a. (2012). Für eine anwendungsorientierte Diskussion mit Fokus auf die Bildungspolitik vgl. Bouguen und Gurgand (2012).

und Kontrollgruppe. Aus konzeptioneller Sicht stellen solche kontrollierten Experimente eine wichtige Benchmark mit bestimmten optimalen Bedingungen dar, an der sich alle anderen hier erörterten Methoden messen lassen müssen.

### *3.1.1 Idee und Intuition*

Um schlussfolgern zu können, dass eine bestimmte Politikmaßnahme einen kausalen Effekt auf wirtschaftliche und soziale Ergebnisgrößen der Maßnahmenteilnehmer bzw. der von der Intervention betroffenen Beobachtungseinheiten hat, würden wir idealerweise gerne dieselbe Einheit in der „kontrafaktischen Welt" beobachten und die zwei Ergebnisse vergleichen. Wir würden gerne beobachten, was geschieht, wenn die Einheit an der Maßnahme teilnimmt, und was geschehen wäre, wenn *dieselbe* Einheit zur gleichen Zeit *nicht* an der Maßnahme teilgenommen hätte. Interessant wären dabei die beobachteten Unterschiede in der Zielgröße. Man könnte dann argumentieren, dass der Unterschied bei Ergebnisgrößen, beispielsweise dem Arbeitsmarkterfolg der beobachteten Personen, wirklich auf die Auswirkungen der Intervention zurückzuführen ist. Da diese kontrafaktische Beobachtung aber nun einmal nicht möglich ist, benötigt man Forschungsdesigns, die diesem kontrafaktischen Vergleich so nahe wie möglich kommen.

In diesem Zusammenhang haben kontrollierte Experimente faszinierende Eigenschaften, so dass sie manchmal als das exakteste aller Forschungsdesigns oder sogar als der „Goldstandard" bezeichnet werden (z.B. Angrist 2004). Im Basisansatz versuchen Forscher zwei Gruppen zu bilden, die einander „äquivalent" sind. Die eine Gruppe – die Behandlungsgruppe – wird einer spezifischen Maßnahme ausgesetzt, die andere Gruppe – die Kontrollgruppe – nicht. Abgesehen von der „Behandlung" durch die politische Maßnahme sollen die zwei Gruppen in allen anderen wichtigen Eigenschaften gleich sein. Wie in der kontrafaktischen Situation könnten dann Ergebnisunterschiede zwischen den Gruppen der Maßnahme zugeschrieben werden. Da die Gruppen jedoch aus verschiedenen Beobachtungseinheiten zusammengesetzt werden, ist es nicht möglich, sie so zu bilden, dass sie sich exakt in allen wichtigen Charakteristika gleichen, beispielsweise im Hinblick auf Unternehmenskultur und betriebliche Normen oder auf sozialen Hintergrund und Präferenzen von Personen. Umso wichtiger ist es zu zeigen, dass die zwei Gruppen zumindest mit genügend hoher Wahrscheinlichkeit gleich sind.

Dies lässt sich dadurch erreichen, dass man ausreichend viele Einheiten zufällig aus der Bevölkerung zieht und dann ebenfalls *zufällig* der Behandlungs- und der Kontrollgruppe

zuordnet. Tut man dies konsequent, ist sichergestellt, dass sich die zwei Gruppen nur durch Zufall unterscheiden.

### *3.1.2 Eine Beispielstudie: Weiterbildungsgutscheine in der Schweiz*

Die Vorzüge eines expliziten Experiments lassen sich an einer Beispielstudie illustrieren, in der Weiterbildungsgutscheine evaluiert worden sind, wie sie beispielsweise im Kanton Genf flächendeckend als Subventionierungsmaßnahme zur Förderung der Erwachsenenbildung eingesetzt werden. Schwerdt u.a. (2012) untersuchen ein Feldexperiment, in dem mehr als 1.400 per Zufall ausgewählte Personen aus der ganzen Schweiz Gutscheine für Erwachsenenbildungsmaßnahmen erhalten haben. Die Weiterbildungsbeteiligung der Behandlungsgruppe des Experiments wird der Weiterbildungsbeteiligung einer Vergleichsgruppe von mehr als 9.000 Personen gegenübergestellt, die nicht in den Genuss der Gutscheine kamen. Das wesentliche Merkmal dieses Gutscheinprogramms ist, dass es als kontrolliertes Experiment durchgeführt wurde. Ohne die zufällige Zuteilung ließe sich der kausale Effekt von Weiterbildungsmaßnahmen kaum abschätzen, da die Teilnahme an solchen Maßnahmen im Normalfall auf vielerlei anderen Wegen mit dem Arbeitsmarkterfolg zusammenhängen dürfte.

Da sowohl die Mitglieder der Behandlungs- als auch der Vergleichsgruppe Teilnehmer der Schweizer Arbeitskräfteerhebung waren, bedurfte es keiner gesonderten Datenerhebung, um Informationen über das Weiterbildungsverhalten und den Arbeitsmarkterfolg zu gewinnen. So ließ sich auch vermeiden, dass den Mitgliedern der Behandlungs- und Kontrollgruppe bewusst wurde, Teil eines Experiments zu sein. Den Gutscheinempfängern wurde lediglich in einem Brief vom Statistischen Bundesamt der Schweiz mitgeteilt, dass sie den Gutschein als Teil einer Maßnahme zur Förderung des lebenslangen Lernens in der Schweiz erhalten.

Die zufällige Zuteilung der Weiterbildungsgutscheine garantiert, dass sich die Mitglieder der Behandlungsgruppe im Durchschnitt nicht von den Mitgliedern der Vergleichsgruppe bezüglich ihrer beobachtbaren und nicht-beobachtbaren Charakteristika unterscheiden. Somit liefert ein einfacher Vergleich der Weiterbildungsbeteiligung von Behandlungs- und Kontrollgruppe Aufschluss über den kausalen Effekt von Weiterbildungsgutscheinen auf die Weiterbildungsbeteiligung. Das Experiment zeigt, dass der Gutschein eine Steigerung der Weiterbildungsbeteiligung um 13 Prozentpunkte bewirkt hat. Allerdings führen die Weiterbildungsgutscheine auch zu einem Mitnahmeeffekt im Sinne einer teilweisen Verdrängung von arbeitgeberfinanzierter Weiterbildung. Neben der unmittelbaren Weiterbildungsbeteiligung wird auch die Wirkung der Weiterbildungsgutscheine auf andere Ergebnisvariablen evaluiert. Allerdings zeigen sich hier im Durchschnitt keine signifikanten Auswirkungen auf das Arbeitseinkom-

men, die Beschäftigungswahrscheinlichkeit oder die zukünftige Weiterbildungsbeteiligung gut ein Jahr nach der Maßnahme. Schließlich deuten die Befunde darauf hin, dass Geringqualifizierte am ehesten von der Maßnahme profitieren könnten, aber das Gutscheinangebot nur in geringem Maße nutzen.

### *3.1.3 Weitere Beispiele und wichtige Aspekte*

Zahlreiche weitere Beispiele für Feldexperimente kommen aus dem Bereich der Evaluierung arbeitsmarktpolitischer Maßnahmen.[9] Insbesondere die aktive Arbeitsmarktpolitik, die das Ziel verfolgt, Arbeitslose wieder in Beschäftigung zu bringen, wurde bereits häufig experimentell evaluiert. So gibt es einige Studien, in denen die Wirksamkeit von öffentlichen Arbeitsvermittlungsmaßnahmen mittels kontrollierter Feldexperimente evaluiert wird. In diesen Feldexperimenten erhalten zum Beispiel zufällig ausgewählte Arbeitslose in den Niederlanden eine intensivierte Beratung bei gleichzeitiger stärkerer Kontrolle (Gorter und Kalb 1996; van den Berg und van der Klaauw 2006). Ähnliche experimentelle Evidenz zur Wirksamkeit von staatlichen Aktivierungsmaßnahmen liegt auch für die USA (Klepinger, Johnson und Joesch 2002), Dänemark (Graversen und van Ours 2008, 2011) und Schweden (Hägglund 2009) vor. Hierbei werden teilweise mehrere Behandlungsgruppen gebildet, so dass in einem Experiment unterschiedliche Maßnahmen evaluiert werden können. So gibt es in der Studie von Graversen und van Ours (2008) neben einer Kontrollgruppe insgesamt vier verschiedene Behandlungsgruppen. Beispielsweise bekommt eine der Behandlungsgruppen ein zweiwöchiges Bewerbertraining, während für eine andere Behandlungsgruppe die reguläre Verpflichtung, mindestens zwei potenzielle Arbeitgeber pro Woche zu kontaktieren, auf vier Kontakte verdoppelt wird.[10]

Auch für den bildungs-, wohnungs- und gesundheitspolitischen Bereich gibt es zahlreiche Beispiele experimenteller Evaluationsstudien. So wurden für das Project STAR im US-Bundesstaat Tennessee Schüler und Lehrer zufällig auf Klassen unterschiedlicher Größe aufgeteilt (Finn und Achilles 1990; Krueger 1999; Chetty u.a. 2011). Angrist und Lavy (2009) untersuchen ein israelisches Experiment, in dem Schüler monetäre Anreize für das Erreichen eines Schulabschlusses erhielten. Fryer (2011), Bettinger (2012) und Levitt u.a. (2012) untersuchen Experimente mit finanziellen und nicht-monetären Anreizen für Schüler in einigen

[9] Siehe List und Rasul (2011) mit einem ausführlichen Überblick über die Literatur zu Feldexperimenten in der Arbeitsmarktökonomik.

[10] Siehe Card, Kluve und Weber (2010) für eine Metaanalyse von 97 Studien, die aktive Arbeitsmarktpolitik evaluieren.

US-amerikanischen Städten. Experimente mit finanziellen Anreizen für Lehrer untersuchen unter anderem Muralidharan und Sundararaman (2011) in Indien und Fryer u.a. (2012) in den USA. Rockoff u.a. (2012) analysieren mit einem experimentellen Ansatz, wie Schulleiter auf die Bereitstellung von objektiven Informationen über Leistungen der Lehrer reagieren. Leuven, Oosterbeek und van der Klaauw (2010) realisieren ein Experiment mit finanziellen Anreizen für Studenten an der Universität Amsterdam.

Für verschiedene bildungs- und entwicklungspolitische Maßnahmen sind in den vergangenen Jahren Experimente in Entwicklungsländern durchgeführt worden (Banerjee und Duflo 2009). In den USA ist „Moving to Opportunity" ein breit untersuchtes experimentelles Programm, das Familien mit geringen Einkommen die Möglichkeit bot, in weniger benachteiligte Nachbarschaften umzuziehen (Kling, Liebman und Katz 2007; Ludwig u.a. 2013). Im Gesundheitsbereich untersuchen Augurzky u.a. (2012) ein randomisiertes Feldexperiment mit rund 700 übergewichtigen Personen, um zu evaluieren, ob finanzielle Anreize beim Abnehmen helfen können.

Ein wichtiges Merkmal von experimentellen Studien ist der Zeitraum, in dem die Zielgrößen beobachtet werden können. Der Beobachtungszeitraum von Schwerdt u.a. (2012) lässt beispielsweise nur eine Evaluierung kurzfristiger Effekte von Weiterbildungsgutscheinen bis zu einem Jahr nach Erhalt des Gutscheins zu. Die untersuchten Maßnahmen in Graversen und van Ours (2008) können nur im Hinblick auf einen möglichen Wiederbeschäftigungseffekt innerhalb von 43 Wochen evaluiert werden. In anderen Feldexperimenten ist dagegen ein Forschungsdesign gewählt, das auf einen sehr langen Beobachtungszeitraum ausgerichtet ist.

So wurde beispielsweise im Jahr 1962 in Ypsilanti, Michigan das Perry-Vorschulprogramm für frühkindliche Bildung eingeführt, um die soziale und kognitive Entwicklung von Kindern aus extrem benachteiligten Bevölkerungsschichten in armen Stadtteilen zu verbessern. Im Rahmen eines kontrollierten Experiments wurden 123 Kinder zufällig einer Behandlungs- und einer Kontrollgruppe zugeordnet (Belfield u.a. 2006). Die 58 Kinder der Behandlungsgruppe erhielten für ein oder zwei Schuljahre ein hochwertiges Vorschulprogramm, das verschiedene Fördermaßnahmen umfasste. Die Kinder aus der Kontrollgruppe erhielten diese Fördermaßnahmen nicht. Die Beobachtungseinheiten aus beiden Gruppen wurden bis ins Erwachsenenalter wiederholt befragt, zuletzt im Alter von 40 Jahren. Im Vergleich verschiedener sozialer und wirtschaftlicher Ergebnisgrößen der Erwachsenen aus der Behandlungs- und der Kontrollgruppe zeigen sich in mehreren Dimensionen deutliche positive Effekte des frühkindlichen Bildungsprogramms. Im Erwachsenenalter hatten die geförderten Schüler unter

anderem durchschnittlich höhere Löhne, eine höhere Wahrscheinlichkeit, die High-School abzuschließen, sowie eine niedrigere Kriminalitätsrate als die nicht-geförderten Schüler.

Allerdings ist nicht klar, inwieweit die Ergebnisse verallgemeinert werden können. Während die Zuordnung in die Behandlungs- und Kontrollgruppe rein zufällig erfolgte, wurde die Auswahl der zugrunde liegenden Stichprobe explizit auf Kinder aus extrem benachteiligten Schichten begrenzt. Dies schränkt die „externe Validität“ (Gültigkeit) der Studienergebnisse ein: Die kausale Schlussfolgerung ist begrenzt auf die spezielle Gruppe der benachteiligten afroamerikanischen Schüler. Wenn die Auswahl der Maßnahmenteilnehmer völlig zufällig gewesen wäre, hätten die Effekte möglicherweise anders ausgesehen. Allgemeingültige Ergebnisse, die für die gesamte Bevölkerung zutreffend sind, waren allerdings auch nicht das Ziel dieses spezifischen Experiments. Es ist oftmals interessant und aus einer politischen Sichtweise relevant, den Fokus auf bestimmte Untergruppen zu legen.

Ein Nachteil von expliziten Experimenten besteht darin, dass diese meist in einer etwas künstlichen Umgebung durchgeführt werden, die sich von der „realen Welt“ unterscheidet. Sowohl die Beobachtungseinheiten der Behandlungs- als auch der Kontrollgruppe könnten von ihrem „normalen“ Verhalten abweichen, weil sie wissen, dass sie Teil eines Experiments sind. Idealerweise sollte es den Teilnehmern nicht bewusst sein, dass sie an einem Experiment teilnehmen. Diese Anforderung erfüllen beispielsweise die Forschungsdesigns der Studien von Schwerdt u.a. (2012) und Gorter und Kalb (1996). Die Anforderung trifft auch auf Situationen zu, in denen die zufällige Aufteilung aus einem anderen Grund als der experimentellen Studie geschah. Dies ist beispielsweise bei der zufälligen Zuordnung von Mitbewohnern im College-Wohnheim der Fall, die zur Untersuchung von Peer-Gruppen-Effekten unter Studenten genutzt wurde (Sacerdote 2001; Zimmerman 2003). Aber abgesehen von diesen speziellen Fällen ist die Tatsache, dass Teilnehmer eines Experiments ihr Verhalten genau deswegen ändern können, weil sie wissen, dass sie beobachtet werden (der sogenannte Hawthorne-Effekt), ein Beispiel dafür, dass aus kontrollierten Experimenten stammende Evidenz nicht immer verallgemeinert werden kann. Dies kann insbesondere dann zutreffen, wenn den Teilnehmern ein Ergebnis des Experiments günstiger erscheint als ein anderes Ergebnis.

Die externe Validität experimenteller Befunde ist mitunter auch dadurch beeinträchtigt, dass eine flächendeckende Politikmaßnahme allgemeine Gleichgewichtseffekte hervorrufen kann. Damit ist gemeint, dass zum Beispiel ein Experiment, das regional oder auf eine andere Weise begrenzt ist, zwar das Bildungsniveau erhöhen und auf diesem Weg schließlich die Löhne der geförderten Teilnehmer steigern kann, dass jedoch eine flächendeckende Intervention dieser Art, die alle Regionen betrifft, möglicherweise nicht die gleichen Lohneffekte her-

vorruft. Grund dafür kann sein, dass ein beträchtlicher Anstieg des Angebots an hochgebildeten Arbeitskräften die Bildungsrenditen im allgemeinen Gleichgewicht schmälert.

Darüber hinaus kann es zahlreiche Umstände geben, die die zufällige Ziehung der Stichprobe und die zufällige Zuordnung in die Beobachtungsgruppen erschweren. Heckman u.a. (2010) zeigen zum Beispiel, dass im Perry-Vorschulprogramm eine völlig zufällige Zuordnung nicht gelungen ist, da Schüler aus der Behandlungs- und der Kontrollgruppe teilweise im Nachhinein neu zugeordnet wurden. Kontrollierte Experimente sind oft von Problemen fehlerhafter Umsetzung betroffen, die der Validität der Ergebnisse abträglich sind.

Manchmal werden auch ethische Bedenken bezüglich der Fairness der Zuteilung erhoben: „Wie kann man denjenigen aus der Kontrollgruppe die Förderung und die Vorteile durch die Maßnahme verwehren?“ Natürlich sind solche Einwände nur dann berechtigt, wenn die positiven Effekte durch die Intervention bewiesen und allgemein akzeptiert sind. Wie bei experimentellen Untersuchungen zur Wirksamkeit neuer Medikamente müssen die positiven Effekte zunächst überzeugend nachgewiesen werden, bevor die ganze Bevölkerung damit „behandelt“ wird. Sobald die förderliche Wirkung eines Medikaments bewiesen ist, werden die Tests beendet und das Medikament wird für die breite Anwendung frei gegeben. Auch wenn immer mehr Evidenz die förderliche Wirkung der frühkindlichen Bildung heute ganz offensichtlich erscheinen lässt, konnte in ganz ähnlicher Weise diese Einschätzung etwa vor dem Beginn des Perry-Vorschulprogramms nicht getroffen werden, als die relativen Vorteile der Förderung und Nicht-Förderung durch das Programm noch heftig umstritten waren.

Auch bei überzeichneten staatlichen Förderprogrammen (vgl. Abschnitt 3.2) erscheint es im Vergleich zu alternativen Zuweisungsverfahren am „fairsten“, wenn die Zuteilung zwischen gleichermaßen förderungswürdigen Objekten zufällig geschieht.[11] Darüber hinaus kann ein „Phase-in“-Design angewendet werden, bei dem zunächst nur ein zufällig ausgewählter Teil der Zielgruppe eine Förderung empfängt. Der zunächst als Kontrollgruppe fungierende andere Teil kann etwa ein Jahr später in den Genuss der Maßnahme kommen. So lassen sich gleichwohl zumindest kurzfristige Effekte der Maßnahme evaluieren. Beispielsweise verwenden Rosendahl Huber, Sloof und Van Praag (2012) eine um ein bis zwei Monate verschobene Durchführung, um ein Entrepreneurship-Bildungsprogramm in niederländischen Grundschulen zu evaluieren.

[11] Auch aus rechtlicher Sicht kann die randomisierte Zuteilung staatlicher Fördermittel beispielsweise bezüglich des Gleichbehandlungsgrundsatzes als unbedenklich gelten, solange der Staat keine Willkür walten lässt; die zufällige Zuteilung unter gleichermaßen förderwürdigen Objekten kann sogar verhindern, dass sachwidrige Kriterien in die Zuteilungsentscheidung einfließen (Wissenschaftlicher Beirat beim Bundesministerium für Wirtschaft und Energie 2013).

Einen weiteren Weg, Fairness-Bedenken entgegenzutreten, eröffnet ein Forschungsdesign, bei dem die Maßnahmenteilnahme „rotiert“. Es werden ausschließlich Behandlungsgruppen gebildet, die jeweils einer von mehreren verschiedenen Interventionen ausgesetzt werden. So können die Wirkungen verschiedener Maßnahmen oder unterschiedlicher Ausgestaltungen einer Maßnahme (die gegebenenfalls mit denselben Kosten verbunden sind) gegeneinander getestet werden, indem das Ergebnis nach der Teilnahme an der einen Maßnahme mit dem Ergebnis der anderen Maßnahme verglichen wird. Da es jedoch keine Kontrollgruppe gibt, die keiner Maßnahme ausgesetzt ist, kann allerdings der Effekt der Maßnahmenteilnahme im Vergleich zur Nicht-Teilnahme nicht identifiziert werden. Forschungsdesigns dieser Art haben den Vorteil, dass sie nicht von einschlägigen ethischen Bedenken betroffen sind. Sie könnten deshalb von politischen Entscheidungsträgern durchaus häufiger genutzt werden.

## *3.2 Zufallsvergabe überzeichneter Programme: Wenn man nicht alle bedienen kann*

Die zweite hier betrachtete Methode ist ein Spezialfall des expliziten kontrollierten Experiments, bei dem die Tatsache genutzt wird, dass die Zuordnung in überzeichnete Programme oft durch zufällige Lotterien oder Auslosungen geregelt ist.

### *3.2.1 Idee und Intuition*

Bei bestimmten Fördermaßnahmen – insbesondere wenn es sich um staatliche Pilotprojekte oder Programme gemeinnütziger Stiftungen handelt – reichen die finanziellen Mittel oft nicht aus, um die Teilnahme jedes Interessenten zu ermöglichen. Wenn es mehr Bewerber für eine Fördermaßnahme als zur Verfügung stehende Plätze gibt, das Programm also überzeichnet ist, kann die Teilnahme durch ein zufälliges Losverfahren bestimmt werden, so dass jeder Bewerber die gleiche Chance zur Teilnahme hat.

Wenn nur die Bewerber für das Programm betrachtet werden, stellt sich diese Situation als explizite Randomisierung heraus. Wenn die Teilnahme bzw. Nicht-Teilnahme an einem überzeichneten (Pilot-)projekt auf Basis einer zufälligen Auslosung erfolgt, sollten sowohl die Gewinner als auch die Verlierer der Auslosung beobachtet werden – vorzugsweise vor, aber insbesondere auch nach der Durchführung des Programms. Denn dann ist wieder die Identifikation von kausalen Effekten der Intervention möglich, wenn man die Ergebnisgrößen der zwei zufällig eingeteilten Gruppen vergleicht.

### *3.2.2 Eine Beispielstudie: Vergabe von Privatschulgutscheinen im Losverfahren*

Ein bekanntes Beispiel für die zufällige Auslosung bei überzeichneten Programmen sind Fördermaßnahmen, die den Bewerbern die Chance bieten, einen Gutschein zu gewinnen, der zumindest für einen Teil des Schulgeldes an Privatschulen eingelöst werden kann. Als Beispiele für solche Gutschein-Programme, die mit einer randomisierten Evaluierung einhergehen, bieten sich verschiedene Programme in den USA an – in New York City, Washington, DC und Dayton, Ohio. Diese (privat finanzierten) Programme waren an Familien mit niedrigem Einkommen gerichtet. Jedes dieser Programme zur Übernahme eines Teils des Schulgeldes für Privatschulen war überzeichnet, d.h. es gab mehr Bewerber als zu vergebende Gutscheine. Die Teilnahme wurde deshalb durch ein Losverfahren bestimmt. In jedem Fall erhob man den sozioökonomischen Hintergrund und die erwünschten Ergebnisgrößen der Gewinner der Gutscheine (Behandlungsgruppe) und der Verlierer der Auslosung, die keinen Gutschein erhielten (Kontrollgruppe), sowohl vor der Durchführung des Programms als auch zu mehreren Zeitpunkten danach.

Peterson u.a. (2003) nutzen die Daten dieser Gutschein-Programme, um zu untersuchen, ob sich die Leistungen der Schüler, die Zufriedenheit der Eltern und die Bildungsumgebung der Schulen nach dem Programm signifikant zwischen Behandlungs- und Kontrollgruppe unterscheiden. Weil die Zuordnung in die Gruppen zufällig erfolgte, können solche Unterschiede als der kausale Effekt des Gutschein-Gewinns interpretiert werden. Unter anderem deuten die Ergebnisse darauf hin, dass die Zufriedenheit der Eltern jener Schüler, die einen Gutschein erhielten, gestiegen ist und dass sich die Leistungen derjenigen Schüler, die den Gutschein zum Privatschulbesuch nutzten, in standardisierten Tests (nur) in der Untergruppe der afroamerikanischen Schüler verbessert haben. Das Design der Studien ermöglichte es den Autoren, sowohl den Effekt des Angebots eines Gutscheins (den „Intention-to-treat"-Effekt) als auch den Effekt der tatsächlichen Nutzung des Gutscheins zum Wechsel von einer staatlichen auf eine Privatschule (den „Treatment-on-the-treated"-Effekt) zu schätzen.[12]

### *3.2.3 Weitere Beispiele und wichtige Aspekte*

Einen ähnlichen Ansatz benutzen Angrist, Bettinger und Kremer (2006), um die Vergabe von Bildungsgutscheinen in Kolumbien im Wege einer Lotterie zu evaluieren. Cullen, Jacob und Levitt (2006) und Deming u.a. (2014) nutzen dieselbe Identifikationsstrategie, um die

[12] Weitere Diskussionen der Ergebnisse der Gutschein-Auslosung in New York City finden sich bei Krueger und Zhu (2004) und Peterson und Howell (2004).

Auswirkungen einer größeren Wahlfreiheit zwischen öffentlichen Schulen in Chicago beziehungsweise Charlotte-Mecklenburg zu evaluieren. In ähnlicher Weise analysieren Abdulkadiroğlu u.a. (2011) Lotterien zum Zugang zu überzeichneten unabhängigeren öffentlichen Schulen (Charter schools und Pilot schools) in Boston.

Ein weiteres Beispiel für eine Studie, die auf der zufälligen Auslosung bei überzeichneten Programmen basiert, stammt aus dem Bereich der staatlichen Förderung von Beratung und Weiterbildung für Unternehmensgründungen. Fairlie, Karlan und Zinman (2012) evaluieren das Projekt „Growing America through Entrepreneurship“ (GATE), das von 2003 bis 2005 in sieben US-amerikanischen Städten angeboten wurde. Es beinhaltete spezielle Trainingsmaßnahmen und Beratungsangebote für Unternehmensgründer in einem Gesamtwert von rund 1.300 $ pro Teilnehmer. Allerdings war die Teilnehmerzahl begrenzt. Nur knapp die Hälfte der 4.000 Antragsteller konnte gefördert werden. Da die Teilnehmer per Losentscheid ausgewählt wurden, konnte die Maßnahme durch einen Vergleich der geförderten mit den nichtgeförderten Antragstellern evaluiert werden.

All diese Studien haben eine Gemeinsamkeit, die typisch ist für Evaluierungen, die auf randomisierten Auslosungen bei überzeichneten Programmen beruhen: Die zugrunde liegende Population besteht lediglich aus denjenigen Beobachtungseinheiten, die sich für die Teilnahme an dem Programm beworben haben. Das entspricht üblicherweise keiner zufälligen Ziehung aus der gesamten Bevölkerung. Personen oder Unternehmen, die das Programm besonders schätzen, für sich einen besonderen Bedarf sehen oder besonderen Wert auf das Ergebnis der Intervention legen, dürften eher geneigt sein, sich zu bewerben, als die durchschnittliche Bevölkerung. Folglich stellen die Ergebnisse von solchen Studien bei überzeichneten Programmen valide Schätzungen des kausalen Effektes für die Gruppe der Bewerber dar, was zumeist von großem politischem Interesse ist. Gleichzeitig bleibt aber die externe Gültigkeit der Ergebnisse für die gesamte Bevölkerung unklar.

Der deutlichste Vorteil von Studien, die auf randomisierten Auslosungen überzeichneter Programme beruhen, besteht darin, dass sie nicht die Durchführung eines separaten Experiments erfordern, sondern auf jener Randomisierung aufbauen, die ohnehin stattfindet. Zudem zählen sie eher zu den Feldexperimenten, die in der „realen Welt“ stattfinden, als zu den Experimenten in künstlichen Umgebungen. Ähnlich wie explizite Experimente können auch Evaluierungen randomisierter Verlosungen überzeichneter Programme Hawthorne-Effekten unterliegen und keine allgemeinen Gleichgewichtseffekte auffangen. Darüber hinaus ist es oft schwierig, diejenigen, die bei der Auslosung verloren haben, zur Teilnahme an weiteren Tests und Befragungen zu motivieren.

# 4. „Natürliche“ Experimente: Das Würfeln nachahmen

Nicht in allen Fällen ist es möglich oder sinnvoll, Effekte wirtschaftspolitischer Maßnahmen mit Hilfe kontrollierter Experimente oder Loszuteilung zu evaluieren. Die folgenden beiden Evaluationsverfahren zielen daher darauf, die zufällige Zuordnung zu Behandlungs- und Kontrollgruppe, wie sie in kontrollierten Experimenten möglich ist, mit nicht-experimentell erhobenen Daten nachzuahmen. Sie nutzen Situationen, in denen beispielsweise die Natur, institutionelle Regelungen oder politische Eingriffe eine zufällige bzw. für den interessierenden Zusammenhang exogene Variation hervorrufen – also eine Variation in der Maßnahmenteilnahme aufgrund von Ursachen, die selbst in keiner Weise mit der eigentlich interessierenden Ergebnisgröße zusammenhängen. Aufgrund der „so gut wie zufälligen“ Variation spricht man bei solchen Methoden auch von „natürlichen“ oder „Quasi“-Experimenten.

## *4.1 Instrumentvariablen-Ansatz: Hilfe von außen*

Die dritte Methode in unserer Abhandlung, der Instrumentvariablen-Ansatz, nutzt eine solche Variation in der Teilnahmewahrscheinlichkeit, die ihren Ursprung in einer konkreten Begebenheit hat, welche selbst aber nicht mit der interessierenden Ergebnisvariable in Zusammenhang steht.[13] Dieses Vorgehen hilft, jegliche Teile der Variation in der Maßnahmenteilnahme, die von Verzerrungen durch Endogenität betroffen sind, zu eliminieren.

### *4.1.1 Idee und Intuition*

Das Identifizieren kausaler Effekte stellt eine beachtliche Herausforderung dar, wenn eine absichtliche Randomisierung nicht stattgefunden hat. Hier hilft dann der Instrumentvariablen-Ansatz (oder kurz IV-Ansatz, auch Instrumentalvariablen-Ansatz) weiter, bei dem man es sich zunutze machen kann, dass die Natur manchmal zufällige Zuordnungen herbeiführt.

Die zentrale Idee ist recht einfach: Man stellt sich die Variable, welche die Maßnahmenteilnahme bzw. die Behandlung abbildet, so vor, dass sie zwei Teile hat: Ein Teil ist von den im zweiten Abschnitt erläuterten Endogenitätsproblemen betroffen, zum Beispiel weil die Variable mit einer ausgelassenen Variable zusammenhängt. Der andere Teil ist von diesem Endogenitätsproblem nicht betroffen und kann somit für die kausale Identifikation genutzt werden. Mit dem IV-Ansatz versucht man nun, diesen zweiten Teil der Variation der Behandlungsvariable zu isolieren, der nicht von Endogenität betroffen ist. Dies wird dadurch erreicht,

[13] Neben den oben gegebenen allgemeinen Literaturhinweisen vgl. Angrist und Krueger (2001) für eine einfache Einführung in den Instrumentvariablen-Ansatz.

dass man nur jenen Teil der Variation der Variable in die Analysen einbezieht, der durch eine beobachtete zusätzliche Variable (das „Instrument“) bestimmt wird, die ihrerseits nicht anderweitig mit der Ergebnisvariable zusammenhängt. Wenn Informationen über ein solches Instrument zur Verfügung stehen, ist es möglich, jene Variation in der Maßnahmenteilnahme zu isolieren, die im betrachteten Modell exogen ist. Findet sich nun ein Zusammenhang zwischen diesem exogenen Teil der Variation und der Ergebnisvariable, so kann dieser als kausaler Effekt der Maßnahme auf das Ergebnis interpretiert werden.

Das Entscheidende bei IV-Schätzungen ist, ein überzeugendes Instrument zu finden. Das ist eine Variable, die mit der Teilnahme an der politischen Maßnahme zusammenhängt (diese Eigenschaft wird als „Relevanz“ des Instruments bezeichnet), nicht jedoch mit der Ergebnisvariable – abgesehen von dem möglichen indirekten Zusammenhang, der durch die Maßnahmenteilnahme entsteht (diese Eigenschaft wird als „Exogenität“ des Instruments bezeichnet). Wenn sich solch ein Instrument findet, lässt sich der Behandlungseffekt durch jenen Teil der Variation in der Maßnahmenteilnahme identifizieren, der durch die Variation in der Instrumentvariable ausgelöst wird. Dadurch werden Probleme wie Verzerrungen der Ergebnisse aufgrund umgekehrter Kausalität oder ausgelassener Variablen vermieden und konsistente Schätzergebnisse erreicht.

### *4.1.2 Eine Beispielstudie: Britisches Investitionsförderprogramm*

Ein gutes Beispiel zur Veranschaulichung des Instrumentvariablen-Ansatzes ist die Studie von Criscuolo u.a. (2012), in der das britische Investitionsförderungsprogramm „Regional Selective Assistance“ (RSA) evaluiert wird. Das RSA-Programm gewährt in bestimmten Regionen staatliche Beihilfen für Investitionsvorhaben. Die Teilnahme an dem Programm ist natürlich nicht zufällig: Vor allem sollen wirtschaftlich strukturschwache Regionen gefördert werden, und Unternehmen, die von negativen Nachfrageschocks betroffen sind, dürften sich in besonderem Maße um die RSA-Förderung bemühen. Falls sich der Nachfrageschock auch direkt negativ beispielsweise auf die künftige Beschäftigungsentwicklung der Unternehmen auswirkt, dann unterschätzt der einfache Zusammenhang zwischen Förderung und Beschäftigungsentwicklung die tatsächliche Förderwirkung.

Die britische Förderpolitik unterliegt den Vorgaben der Europäischen Union, die eine derartige Subventionierung nur in Ausnahmefällen zulässt. In welchen Regionen eine Förderung durch RSA möglich ist, wird daher auch durch Kriterien bestimmt, die auf EU-Ebene festgelegt werden. Diese Kriterien, die bestimmen, welche Regionen in welcher Höhe vom RSA-Programm profitieren können, werden etwa alle sieben Jahre angepasst. Diese zeitliche

Variation in den Ausnahmeregelungen der EU erzeugt eine regionalspezifische Variation in der Teilnahmewahrscheinlichkeit am RSA-Programm, die zur Evaluierung des Förderprogramms genutzt werden kann. Die regionalspezifischen Änderungen in den Ausnahmeregelungen der EU werden also als Instrumentvariable für die Teilnahme an der britischen Fördermaßnahme verwendet. Dass dieses Instrument „relevant" ist – dass es also die Teilnahme am RSA-Programm signifikant vorhersagt –, lässt sich leicht zeigen. Die identifizierende Annahme des Ansatzes besteht darin, dass das Instrument auch „exogen" ist, dass also die Anpassungen der Ausnahmeregelungen auf der Ebene der EU nicht systematisch zusammenhängen mit regionalspezifischen Veränderungen in den betrachteten Zielgrößen wie Beschäftigung, Investitionsvolumen und Produktivität der britischen Unternehmen.

Für Firmen mit mehr als 150 Beschäftigten finden sich in der Studie keine Effekte der Fördermaßnahme. Aber für kleine Unternehmen zeigt sich, dass sich die Förderung positiv auf Investitionen und Beschäftigung auswirkt, wenn auch nicht auf die Produktivität. Interessant ist ein Vergleich mit den Ergebnissen herkömmlicher multivariater Regressionsmodelle, die deutlich kleinere Schätzergebnisse ergeben als die Schätzungen des IV-Ansatzes, was auf die angesprochene Endogenität der Teilnahme am RSA-Programm hindeutet.

### *4.1.3 Weitere Beispiele und wichtige Aspekte*

Der IV-Ansatz hat eine sehr breite Anwendungsmöglichkeit, wobei verschiedene Instrumente auf ganz unterschiedliche Weise eine exogene Variation „von außen" bringen können. Da die Quelle exogener Variation oftmals sehr speziell ist, sollen hier einige sehr unterschiedliche Beispiele mögliche Anwendungen verdeutlichen. So nutzen Frölich und Lechner (2010) in einem IV-Ansatz die Tatsache, dass die Umsetzungsintensität von Maßnahmen der aktiven Arbeitsmarktpolitik zwischen Schweizer Kantonen variiert, auch wenn Teile zweier Kantone de facto einen gemeinsamen lokalen Arbeitsmarkt bilden. So können sie die regionale Umsetzungsintensität arbeitsmarktpolitischer Maßnahmen als Instrument dafür nehmen, ob Individuen auf demselben lokalen Arbeitsmarkt in den Genuss der Maßnahmen gekommen sind oder nicht. Die Identifikation beruht also auf dem Unterschied zwischen Personen, die nur deshalb von Maßnahmen der aktiven Arbeitsmarktpolitik betroffen sind, weil sie auf derjenigen kantonalen Seite eines lokalen Arbeitsmarktes wohnen, der eine höhere Umsetzungsintensität aufweist, und Personen auf demselben lokalen Arbeitsmarkt, die nur deshalb keinen Zugang zu den Maßnahmen haben, weil sie auf der anderen Seite der kantonalen Grenze wohnen.

In einer umweltpolitischen Anwendung nutzen Aichele und Felbermayr (2012) die nationalen Ratifizierungen des Internationalen Strafgerichtshofs als Instrumentvariable, um den Effekt der Ratifizierung des Kyoto-Protokolls auf die $CO_2$-Bilanz zu evaluieren. Die Kernidee dieser Identifikationsstrategie ist, dass generelle länderspezifische Präferenzunterschiede im Hinblick auf gemeinschaftliche internationale Politikinitiativen einen Zusammenhang zwischen der Wahrscheinlichkeit der Ratifizierung beider Abkommen bedingen, dass es aber keinen direkten Effekt zwischen der nationalen Ratifizierung des Internationalen Strafgerichtshofs und der $CO_2$-Bilanz eines Landes gibt.

In zahlreichen Studien werden Veränderungen der Gesetze über Pflichtschuljahre genutzt, um den Effekt von Bildung auf Ergebnisgrößen wie Einkommen, Kriminalität, Gesundheit, Sterblichkeit, Bildungsstand der nächsten Generation und vieles andere mehr zu untersuchen.[14] Dabei werden die Änderungen der Pflichtschulgesetze als Instrument für individuelle Bildungsniveaus verwendet, um zu vermeiden, dass letztere mit unbeobachteten Faktoren wie angeborenen Begabungen zusammenhängen, welche selbst mit den verschiedenen Ergebnisgrößen zusammenhängen dürften.

Das Instrument der Pflichtschulgesetze veranschaulicht gut, dass IV-Schätzer im Allgemeinen als lokaler Effekt oder sogenannter „Local Average Treatment Effect“ (LATE) interpretiert werden sollten. So ist es denkbar, dass sich eine Steigerung des Bildungsniveaus anders auf die genannten Ergebnisvariablen auswirkt, wenn sie durch eine Verlängerung der Schulzeit von 8 auf 9 Jahre zustande kommt, als wenn ihr die Ausweitung einer Lehrausbildung zu einem Hochschulabschluss zugrunde liegt. Wie Angrist, Imbens und Rubin (1996) zeigen, kann man in diesem Fall, wenn sich die Auswirkungen der Bildung je nach Personen und Situationen unterscheiden, mit dem IV-Ansatz einen durchschnittlichen Effekt für diejenige Teilgruppe der Gesamtpopulation identifizieren, die ihren Behandlungsstatus – in diesem Fall die Anzahl der Bildungsjahre – aufgrund des Instruments ändert. Man bezeichnet diese Teilgruppe als „Complier“, weil sie die Behandlung „befolgen“. Im Fall des Pflichtschulgesetz-Instruments identifiziert man also den Effekt von Bildung für Schüler, deren Bildungsentscheidung vom Pflichtschulalter abhängt – und dieser Effekt muss nicht notwendigerweise dem „durchschnittlichen“ Effekt der Bildung in der Gesamtpopulation entsprechen. Das wirft wiederum Fragen der externen Validität auf.

[14] U.a. Harmon und Walker (1995); Oreopoulos (2006); Brunello, Fort und Weber (2009); Lochner und Moretti (2004); Kemptner, Jürges und Reinhold (2011); Lleras-Muney (2005); Black, Devereux und Salvanes (2005); Piopiunik (2014).

Parey und Waldinger (2011) nutzen die fachbereichsspezifische Einführung und Ausbreitung des Erasmus-Programms zur Förderung des europäischen Studentenaustauschs an deutschen Universitäten als Instrument, um den Effekt eines Auslandsstudiums auf die internationale Arbeitsmobilität zu erfassen. Um den Effekt der Ausbreitung von Breitband-Internet auf soziale Aktivitäten der Bevölkerung zu evaluieren, berücksichtigen Bauernschuster, Falck und Wößmann (2011), dass die vor dem Aufkommen des Internets in einigen Telekommunikations-Anschlussgebieten in Ostdeutschland verlegte OPAL-Technologie den Zugang zu Breitband verhindert. So erhalten sie den Anschluss von Haushalten an OPAL-Anschlüsse als Instrument, um die exogene Variation im Zugang zu Breitband zu identifizieren.

In einigen Studien hilft jene Variation weiter, die im wahren Wortsinne von der Natur hervorgerufen wurde. So nutzen Hægeland, Raaum und Salvanes (2012) den Zugang norwegischer Gemeinden zu Wasserfällen und damit zu Einnahmen aus Steuern auf Wasserkraftwerke, um den Effekt zusätzlicher Bildungsausgaben auf Schülerleistungen zu evaluieren. Die Aufteilung von Schülern auf Klassen unterschiedlicher Größe innerhalb von und zwischen Schulen ist alles andere als zufällig und hängt mit zumeist unbeobachteten Faktoren wie elterlichen Wahlentscheidungen sowie schulischen und systemischen Zuweisungen zusammen. Aufgrund der natürlichen Zufälligkeit von Geburten um den Einschulungsstichtag herum variiert die Größe der Einschulungsjahrgänge in Schuleinzugsgebieten aber auch zufällig von Jahr zu Jahr. Da damit auch die Klassengröße in einer Schule von Jahrgang zu Jahrgang variiert, nutzt Hoxby (2000) die aufgrund dieser natürlichen Schwankungen vorhergesagte Klassengröße als Instrument für exogene Variation in der tatsächlichen Klassengröße, um den kausalen Effekt von Klassengrößen auf Schülerleistungen im US-Bundesstaat Connecticut zu schätzen. Wößmann und West (2006) schätzen mit einem IV-Ansatz Klassengrößeneffekte in einer Reihe von Ländern.

West und Wößmann (2010) untersuchen, wie sich die Konkurrenz nicht-staatlicher Schulen auf die Leistung der Schüler auswirkt. Sie nutzen dabei die Tatsache, dass es in manchen Ländern um 1900 einen größeren Anteil von Katholiken gab als in anderen, die in Opposition gegen das aufkommende staatliche Schulsystem standen. Diese Länder haben noch heute mehr Schulen in nicht-staatlicher Trägerschaft. Die Wissenschaftler verwenden in ihrem internationalen Vergleich den historischen Katholikenanteil als Instrument für heutigen nicht-staatlichen Wettbewerb.

Wie diese Beispiele veranschaulichen, kann man mit IV-Spezifikationen exogene Variation verschiedenster Quellen nutzen, darunter politische Veränderungen, echte natürliche Variationen und historische Besonderheiten. In der Praxis liegt die zentrale Herausforderung

des IV-Ansatzes darin, dass man ein überzeugendes Instrument finden muss. Bei jeder Anwendung müssen die Hauptbedingungen des IV-Ansatzes – Relevanz und Exogenität des Instruments – sorgfältig bedacht werden. Dabei muss der Zusammenhang zwischen dem Instrument und der Teilnahme an der politischen Maßnahme möglichst eng sein, damit das Problem „schwacher Instrumente“ nicht aufkommt. Wenn ein überzeugendes Instrument gefunden ist, können kausale Effekte identifiziert werden, selbst wenn die beobachteten Daten lediglich im Querschnitt vorliegen.

Der Vorteil dieser quasi-experimentellen Analysen besteht darin, dass sie einige der maßgeblichen Probleme von expliziten Experimenten vermeiden. Dazu gehört, dass Feldexperimente in der Regel teuer und zeitaufwendig sind, weshalb Politiker und Verwaltungsakteure nur schwer von ihrer Notwendigkeit überzeugt werden können. Darüber hinaus sind quasi-experimentelle Studien nicht vom Hawthorne-Effekt betroffen, weil den Individuen nicht bewusst ist, dass sie Teil eines Experiments sind. Im Idealfall können natürliche Experimente auch allgemeine Gleichgewichtseffekte einfangen, was mit Feldexperimenten üblicherweise nicht möglich ist.

### *4.2 Regressions-Diskontinuitäten-Ansatz: Wenn Maßnahmen einen Sprung machen*

Die vierte Methode in unserer Abhandlung ist auf Situationen gemünzt, in denen die Teilnahme an einer Maßnahme diskontinuierlich – also „sprunghaft“ – durch einen Schwellenwert bestimmt wird.[15]

#### *4.2.1 Idee und Intuition*

Der Regressions-Diskontinuitäten-Ansatz (RD) ist ein weiterer Ansatz aus der Klasse der natürlichen Experimente. Das RD-Design kann in jener speziellen Situation angewendet werden, in der die Teilnahme an einer Maßnahme dadurch bestimmt wird, ob ein potenzieller Teilnehmer im Hinblick auf eine Variable unter oder über einem festgelegten Schwellenwert liegt, von der die Zuordnung in Teilnahme und Nicht-Teilnahme abhängig gemacht wird. Die Wahrscheinlichkeit, an der Maßnahme teilzunehmen, macht also beim Schwellenwert der Zuordnungsvariable einen Sprung. Standardbeispiele sind Programme, mit denen nur Unternehmen, deren Mitarbeiterzahl unter oder über einer bestimmten Schwelle liegt, oder nur Personen jenseits einer bestimmten Altersgrenze gefördert werden.

[15] Neben den oben gegebenen allgemeinen Literaturhinweisen siehe Imbens und Lemieux (2008) und Lee und Lemieux (2010) für umfassende Darstellungen des Regressions-Diskontinuitäten-Ansatzes. Eine sehr technische Darstellung bieten Hahn, Todd und Van der Klaauw (2001).

Die Idee beim RD-Ansatz ist es, Beobachtungseinheiten zu vergleichen, die in einer ausreichend kleinen Bandbreite gerade über und unter dem Schwellenwert liegen. Die Einheiten oberhalb bilden dann beispielsweise die Behandlungsgruppe, die aufgrund des Überschreitens des Schwellenwertes der Maßnahme ausgesetzt wird. Die Einheiten unterhalb des Schwellenwertes bilden die Kontrollgruppe. Die Intuition dahinter ist, dass sich die betrachteten Einheiten sehr ähnlich sind und sich nur darin unterscheiden, ob sie an der Maßnahme teilnehmen. Im Hinblick auf die Zuordnungsvariable weisen sie schließlich sehr ähnliche Werte auf. Wenn zum Beispiel die Teilnahmemöglichkeit an einer Forschungsförderung daran geknüpft ist, dass ein Unternehmen weniger als 250 Mitarbeiter hat, dann werden die Unternehmen gerade unter und gerade über diesem Schwellenwert miteinander verglichen. Wenn die Ergebnisgröße – zum Beispiel die Ausgaben für Forschung und Entwicklung (F&E) ein Jahr nach der Maßnahme – genau am Schwellenwert einen Sprung (eine Diskontinuität) aufweist, so kann dies als kausaler Effekt der Maßnahme interpretiert werden (vgl. Abbildung 1).

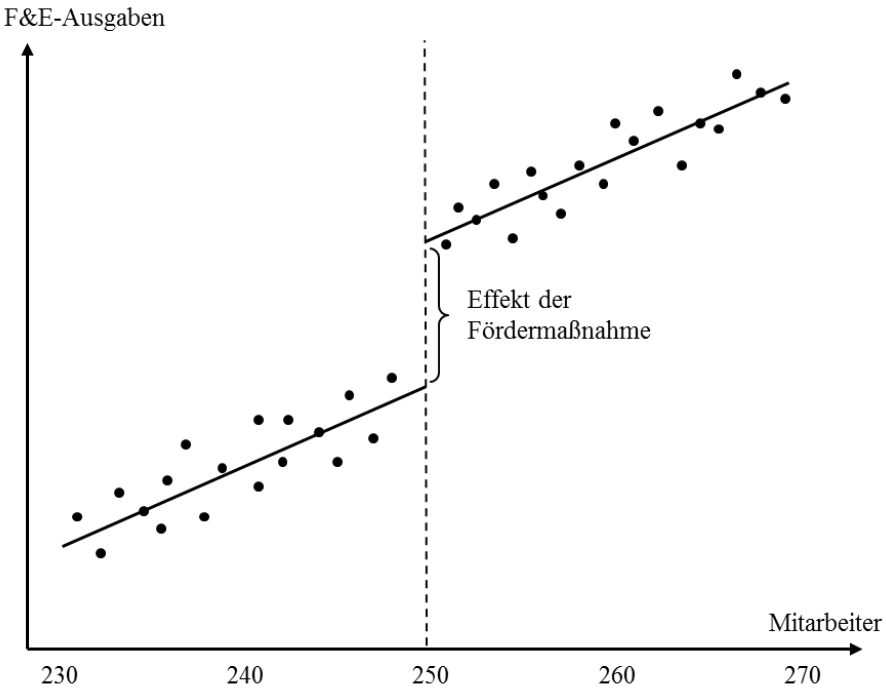


**Abbildung 1:** Der Regressions-Diskontinuitäten-Ansatz
Quelle: Eigene Darstellung.

Die zugrunde liegende Annahme besteht darin, dass sich etwa Unternehmen mit 240-249 Mitarbeitern kaum von Unternehmen mit 250-259 Mitarbeitern unterscheiden – außer eben

im Zugang zur Fördermaßnahme. Zusätzlich ermöglicht es die Tatsache, dass die Zuordnung in Behandlungs- und Kontrollgruppe nach einem nicht-kontinuierlichen Muster erfolgt, jegliche kontinuierlichen Effekte der Variable herauszurechnen, die die Teilnahmeberechtigung festlegt. Genau genommen ist die notwendige Annahme, um mit dem RD-Ansatz einen kausalen Effekt zu identifizieren, also nur, dass der Zusammenhang zwischen Mitarbeiterzahl und Ergebnisgröße in dem kleinen Intervall stetig ist und keine weiteren Diskontinuitäten um den Schwellenwert herum bestehen. Das lässt sich beispielsweise damit belegen, dass vor Einführung der Maßnahme keine beobachtbaren Unterschiede zwischen den beiden Gruppen bestanden haben.

### *4.2.2 Eine Beispielstudie: Regionalförderung der EU-Strukturfonds*

Becker, Egger und von Ehrlich (2010, 2013) untersuchen mit dem RD-Ansatz die Auswirkungen der EU-Strukturfonds auf die regionale Entwicklung. In der „Ziel-1"-Kategorie sollen die Strukturmittel Regionen mit Entwicklungsrückstand fördern. Deshalb sind die auf Konvergenz zielenden Fördermaßnahmen solchen Regionen vorbehalten, deren Bruttoinlandsprodukt (BIP) pro Kopf unter 75% des europäischen Durchschnitts liegt. Die Regel führt an dieser Schwelle zu einem abrupten Sprung in der Förderwahrscheinlichkeit.[16] Dementsprechend wird in der Studie das Wirtschafts- und Beschäftigungswachstum der Regionen in verschiedenen Bandbreiten gerade unter- und oberhalb dieser Schwelle verglichen. Die Betrachtung ist zum Beispiel nur auf die Regionen konzentriert, die ein BIP pro Kopf zwischen 70% und 80% des europäischen Durchschnitts haben.[17]

Es zeigt sich, dass das Wirtschaftswachstum an dieser Schwelle einen Sprung macht, nicht aber das Beschäftigungswachstum, was auf Wachstums-, nicht aber Beschäftigungseffekte hindeutet. Allerdings sind die positiven Effekte auf rund ein Viertel der geförderten Regionen beschränkt, die genügend Humankapital und hinreichend gute Institutionen aufweisen. Um zu untermauern, dass die Diskontinuität tatsächlich zufällige Unterschiede in der Förderung und keine sonstigen Unterschiede zwischen Behandlungs- und Kontrollgruppe abbildet, zeigen die Autoren darüber hinaus, dass eine Reihe weiterer Variablen an der Diskontinuität keine Sprünge machen, beispielsweise die Beschäftigungsquote, die Anteile der Sektoren Landwirtschaft, Industrie und Dienstleistungsgewerbe sowie Wachstum und Dichte der Bevölkerung.

[16] Dieselbe Diskontunität wird in der Studie von Pellegrini u.a. (2013) verwendet.

[17] Die Hauptspezifikation erfolgt in der Grundgesamtheit aller Regionen mithilfe einer Regression, die für hohe Polynome der Zuordnungsvariable (also des BIP pro Kopf) kontrolliert.

Wie in den meisten angewandten RD-Studien wird hier ein sogenannter unscharfer („fuzzy“) RD-Ansatz verwendet. Im Gegensatz zum scharfen („sharp“) RD-Ansatz springt die Teilnahmewahrscheinlichkeit an der zu untersuchenden Fördermaßnahme beim Überschreiten des Schwellenwertes nicht von null auf eins (oder umgekehrt). Es gibt lediglich einen signifikanten Sprung in der Wahrscheinlichkeit, an der Maßnahme teilzunehmen. Das geschieht immer dann, wenn von der Zuordnungsregel abgewichen wird. So gibt es im vorliegenden Beispiel einige wenige Regionen, die keine Förderung erhalten, obwohl ihr BIP pro Kopf unter 75% des europäischen Durchschnitts liegt, während andere Regionen, die eigentlich über dem Schwellenwert liegen, trotzdem gefördert werden. Diese Abweichler (Non-complier) sollten nicht zur Identifikation des kausalen Effektes beitragen, da zu vermuten ist, dass hier andere Gründe die Maßnahmenteilnahme beeinflussen, die ebenfalls einen direkten Einfluss auf die Zielgrößen haben. Um jene exogene Variation in der Maßnahmenteilnahme zu isolieren, die allein durch die Regel generiert wird, wird im Fall von unscharfen RD-Ansätzen in der Regel eine Instrumentvariablen-Schätzung durchgeführt. Das verwendete Instrument ist dann eine einfache Indikatorvariable, die angibt, ob eine Region auf Basis der Regel die Förderung erhalten sollte oder nicht. Dies stellt sicher, dass der Effekt ausschließlich aufgrund von exogener Variation geschätzt wird.

### *4.2.3 Weitere Beispiele und wichtige Aspekte*

Zahlreiche RD-Studien basieren auf Altersschwellenwerten für den Zugang zu politischen Maßnahmen. So nutzt Lalive (2008) einen scharfen RD-Ansatz, um die Auswirkungen einer Maßnahme zu untersuchen, die 1988 in bestimmten österreichischen Regionen die maximale Bezugsdauer des Arbeitslosengeldes für Personen jenseits der Altersgrenze von 50 Jahren von 30 auf 209 Wochen verlängert hat. Abrupte Sprünge in der Dauer der Arbeitslosigkeit am Altersschwellenwert und an den regionalen Grenzen der förderfähigen Regionen deuten darauf hin, dass die Maßnahme die Arbeitslosigkeit der berechtigten Personen deutlich verlängert hat. In ähnlicher Weise wenden auch Schmieder, von Wachter und Bender (2012) und Caliendo, Tatsiramos und Uhlendorff (2013) den RD-Ansatz an, um die Wirkung der Bezugsdauer von Arbeitslosengeld in Deutschland zu untersuchen.

Um den Effekt der Wehrdienstpflicht auf Arbeitsmarktergebnisse zu evaluieren, nutzen Bauer u.a. (2012) die Diskontinuität, dass bei der Einführung der Wehrpflicht in Deutschland vor dem 1. Juli 1937 geborene Männer von der Wehrpflicht ausgenommen waren. Yörük und Yörük (2011) nutzen die Diskontinuität, dass Bürger in den USA erst ab dem 21. Geburtstag legal alkoholische Getränke kaufen dürfen, um den Effekt des gesetzlichen Mindestalters für

den Alkoholkonsum auf den tatsächlichen Konsum von Alkohol, Tabakprodukten und Marihuana zu evaluieren. Bedard und Dhuey (2006), Mühlenweg und Puhani (2010) und McCrary und Royer (2011) identifizieren Effekte des relativen Einschulungsalters auf die weitere Schullaufbahn anhand von Einschulungsstichtagen, die dazu führen, dass einige Kinder bei der Einschulung aus exogenen Gründen fast ein Jahr älter sind als andere.

Buettner (2006) nutzt die Tatsache, dass der im Rahmen des Länderfinanzausgleichs in einer Gemeinde verbleibende Gewerbesteueranteil bei bestimmten Schwellenwerten der relativen Finanzstärke einer Gemeinde sprunghaft steigt, um Effekte des Finanzausgleichs auf die gemeindliche Steuerpolitik zu untersuchen. Um zu evaluieren, wie sich Kündigungsschutz auf das Verhalten der Beschäftigten auswirkt, prüfen Ichino und Riphahn (2005), ob Abwesenheitstage in einer großen italienischen Bank, in der Angestellte nach 12 Wochen Probezeit dem Kündigungsschutz unterliegen, an diesem Schwellenwert sprunghaft steigen.

Weitere Beispiele für die Anwendbarkeit des RD-Ansatzes finden sich in Studien, in denen anhand von Klassenteilern der kausale Effekt von Klassengrößen auf den Bildungserfolg evaluiert wird, beispielsweise Angrist und Lavy (1999) für Israel; Wößmann (2005) für zahlreiche europäische Länder; Fredriksson, Öckert und Oosterbeek (2013) für Schweden. Klassenteiler, das heißt Schwellenwerte für maximale Klassengrößen, führen zu Diskontinuitäten im Zusammenhang zwischen Jahrgangsstufengröße und durchschnittlicher Klassengröße. Bei einer maximalen Klassengröße von 30 Schülern wird mit dem 31. Schüler im Jahrgang eine zweite Klasse aufgemacht, so dass bei einem Anstieg der Jahrgangsstufengröße von 30 auf 31 Schüler die durchschnittliche Klassengröße abrupt von 30 auf 15,5 sinkt. Auch wenn diese Regel nicht strikt befolgt wird, bietet sie exogene Variation in der Klassengröße im Rahmen eines unscharfen RD-Ansatzes.

Um ein niederländisches Förderprogramm für Schulen mit einem hohen Anteil an Schülern aus benachteiligten Schichten zu evaluieren, nutzen Leuven u.a. (2007) den Schwellenwert von 70% Schülern aus benachteiligten Schichten an einer Schule im Rahmen eines RD-Ansatzes. Lavy (2010) wiederum implementiert einen geographischen RD-Ansatz, in dem er Schüler in Tel Aviv, wo freie Schulwahl zwischen öffentlichen Schulen eingeführt wurde, mit Schülern in benachbarten Wohngebieten vergleicht, wo dies nicht gilt. Um den Effekt von Klassenwiederholungen auf weitere Schülerleistungen zu evaluieren, nutzen Schwerdt und West (2013) die Tatsache, dass Drittklässler in Florida eine bestimmte Punktzahl in einem Lesetest erzielen müssen, um in die vierte Klasse vorrücken zu können. Garibaldi u.a. (2012) untersuchen die Auswirkung von Studiengebühren auf die Studiendauer, indem sie berück-

sichtigen, dass die Studiengebühren an der Bocconi-Universität in Mailand sprunghaften Veränderungen in Abhängigkeit vom Familieneinkommen unterliegen.

Um überzeugende Evaluationsergebnisse mit dem RD-Verfahren zu erhalten, muss Manipulierbarkeit ausgeschlossen werden. Die potenziellen Maßnahmenteilnehmer dürfen nicht in der Lage sein, die Variable, die den Schwellenwert bestimmt, präzise zu manipulieren und sich somit selbst bewusst in die Behandlungs- oder Kontrollgruppe zu sortieren.[18] Anderenfalls wäre die Zuordnung nicht mehr zufällig, und das könnte den identifizierten Effekt der Maßnahmenteilnahme verzerren. Eine solche Manipulationsmöglichkeit ist beispielsweise dann nicht gegeben, wenn das Datum, auf das sich der Schwellenwert bezieht, in der Vergangenheit liegt und der Schwellenwert nicht zuvor bekannt war. Nicht-präzise Manipulierbarkeit ist im Übrigen generell kein Problem, da dann die Variation in der Nähe des Schwellenwertes wiederum als zufällig angesehen werden kann.

Die Beispiele verdeutlichen, dass es viele Fälle gibt, in denen Politikmaßnahmen auf eine Art und Weise eingeführt werden, die eine Diskontinuität enthält und somit eine Evaluierung mittels des RD-Ansatzes ermöglicht. Ein Problem besteht indes bei der Anwendung des RD-Designs darin, dass es nicht immer möglich ist, genügend Beobachtungen gerade unter- und oberhalb des jeweiligen Schwellenwertes zu finden. Eine Lösung besteht darin, die Bandbreite für die um den Schwellenwert herum betrachteten Beobachtungseinheiten zu vergrößern. Dies reduziert jedoch die Wahrscheinlichkeit, dass sich die Beobachtungseinheiten diesseits und jenseits des Schwellenwerts lediglich in der Maßnahmenteilnahme unterscheiden.[19] Zudem weisen die Ergebnisse von Studien im RD-Design aufgrund der lokalen Identifikation um den Schwellenwert herum nicht notwendigerweise externe Validität auf.

## 5. Methoden mit Paneldaten: Das Unbeobachtete „fixieren“

Die verbleibenden beiden Evaluationsverfahren erlauben es, mit Endogenitätsproblemen umzugehen, indem man eine Variation nutzt, die zutage tritt, wenn die gleichen Beobachtungseinheiten mehrere Male beobachtet werden – üblicherweise (aber nicht notwendigerweise) zu mehreren Zeitpunkten. Solche zweidimensionalen Datensätze werden Paneldaten genannt. Die den paneldatenbasierten Verfahren gemeinsame Hoffnung besteht darin, dass man störende Einflussfaktoren ausschließen kann, selbst wenn sie in den Daten unbeobachtet bleiben.

[18] Ein statistischer Test, ob die Zuordnungsvariable manipuliert wurde, der auf der Kontinuität der Verteilungsfunktion der Zuordnungsvariable in der Umgebung des Schwellenwertes beruht, findet sich bei McCrary (2008).

[19] Ein statistisches Verfahren, das die optimale Bandbreite um den Schwellenwert bestimmt, bieten Imbens und Kalyanaraman (2012).

Dies ist möglich, solange die unbeobachteten Faktoren über die zweite Dimension des Datensatzes „fixiert“ sind – also zum Beispiel über die Zeit konstant bleiben.

## *5.1 Differenzen-in-Differenzen-Ansatz: Unterscheidet sich die Differenz?*

Die fünfte Methode beruht auf Datensätzen, in denen jede Beobachtungseinheit mindestens zweimal beobachtet worden ist und ein Teil der Beobachtungseinheiten seinen Teilnahmestatus an der politischen Maßnahme ändert.[20] Die Beobachtungspunkte entsprechen üblicherweise zwei Zeitpunkten, es können aber auch andere Dimensionen gewählt werden, beispielsweise verschiedene Mitarbeiter desselben Unternehmens oder verschiedene Schulfächer desselben Schülers.

### *5.1.1 Idee und Intuition*

Mit dem Differenzen-in-Differenzen-Ansatz (oft auch „Doppelter-Differenzen-Ansatz“, englisch differences-in-differences oder kurz „Diffs-in-Diffs“, DiD) lassen sich im einfachsten Fall zwei Gruppen zu zwei Zeitpunkten betrachten. In der ersten Periode hat noch keine Gruppe an der politischen Maßnahme teilgenommen. In der zweiten Periode hat dann eine der Gruppen an der Maßnahme teilgenommen, die andere nicht. So werden zum Beispiel zwei Gruppen von Unternehmen zu zwei Zeitpunkten beobachtet. Zwischen diesen Zeitpunkten erhält eine der beiden Gruppen eine Förderung. Mit dem DiD-Ansatz kann man nun den Effekt der Förderung beispielsweise auf die Beschäftigungslage in den Unternehmen identifizieren, indem er die durchschnittliche *Veränderung* der Beschäftigung in den beiden Gruppen von Unternehmen miteinander vergleicht (vgl. Abbildung 2).

Die Identifikationsstrategie des DiD-Ansatzes besteht somit darin, in zwei Dimensionen Differenzen zu bilden: Die erste Differenz ist die durchschnittliche Veränderung der Beschäftigung zwischen den beiden Beobachtungszeitpunkten, die separat für die Behandlungs- und die Kontrollgruppe berechnet wird. Die zweite Differenz ist der Unterschied zwischen den beiden Differenzen, die auf der ersten Stufe berechnet wurden. Diese „Differenz in der Differenz“ misst also, wie sich die Veränderung der Ergebnisgröße über die Zeit zwischen den beiden Gruppen unterscheidet. Sie wird als kausaler Effekt der Maßnahme interpretiert.

[20] Neben den oben gegebenen allgemeinen Literaturhinweisen siehe etwa Besley und Case (2000) und Bertrand, Duflo und Mullainathan (2004) zum Differenzen-in-Differenzen-Ansatz.

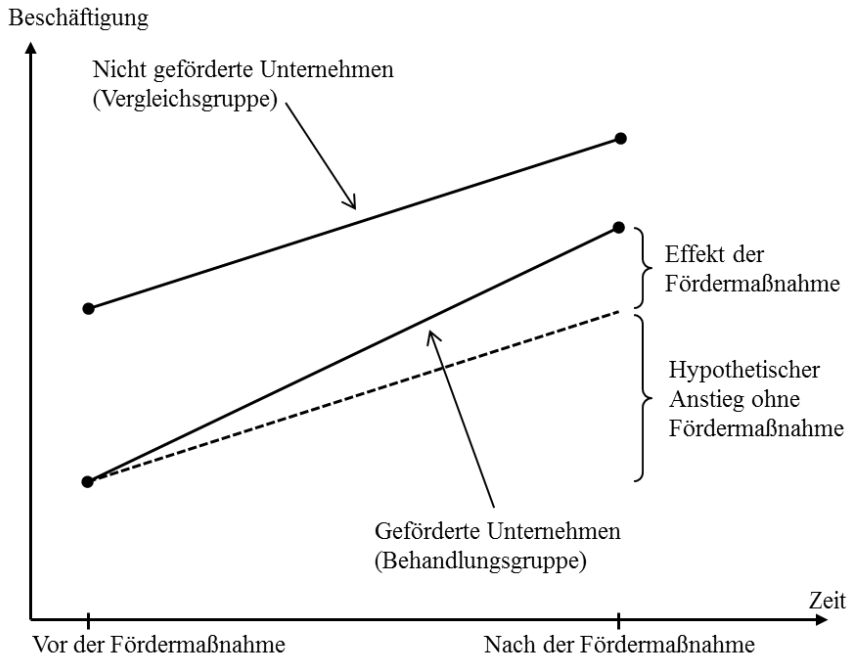


**Abbildung 2: Der Differenzen-in-Differenzen-Ansatz**

Quelle: Eigene Darstellung.

Die Idee ist einfach: Die beiden Gruppen könnten sich in der Ausgangsperiode durchaus unterscheiden. Das heißt, die gruppenspezifischen Mittelwerte der Ergebnisgröße könnten sich auch dann unterscheiden, wenn keine der Gruppen an der Maßnahme teilnimmt. Solange dieser Unterschied zwischen den Gruppen jedoch ohne die Maßnahme über die Zeit konstant bliebe, kann er ausdifferenziert werden, indem die gruppenspezifischen Mittelwerte der Ergebnisvariable voneinander abgezogen werden. Der in der zweiten Stufe verbleibende Unterschied zwischen den gruppenspezifischen Differenzen der ersten Stufe bildet dann den interessierenden kausalen Effekt ab.

Die zentrale Annahme des DiD-Ansatzes besteht also darin, dass die gruppenspezifischen *Trends* der interessierenden Ergebnisgröße ohne die Maßnahme identisch gewesen wären (nachdem gegebenenfalls Veränderungen in weiteren beobachteten Einflussfaktoren herausgerechnet wurden). In unserem Beispiel besteht die identifizierende Annahme also darin, dass sich die durchschnittliche Beschäftigung in den beiden Unternehmensgruppen gleich stark verändert hätte, wenn es keine Förderung gegeben hätte. Die gestrichelte Linie in Abbildung 2 veranschaulicht diese Annahme, dass die geförderten Unternehmen eine kontrafaktische Entwicklung der Beschäftigung erfahren hätten, die parallel zur beobachteten Beschäftigungsentwicklung der nicht-geförderten Unternehmen verläuft. Die Plausibilität dieser Identifikationsannahme hängt vom spezifischen Studiendesign ab. In jedem Fall ist die Identifikations-

annahme des DiD-Ansatzes weniger restriktiv als die Annahme, die implizit in klassischen Standardmethoden gemacht wird, nämlich dass die zwei Gruppen im Hinblick auf das *Niveau* aller relevanten unbeobachteten Faktoren identisch sind.

### *5.1.2 Eine Beispielstudie: Die bayerische „High-Tech-Offensive“*

Falck, Heblich und Kipar (2010) verwenden den DiD-Ansatz, um die Effekte einer bayerischen Politikmaßnahme auf den Innovationsgrad von Unternehmen zu evaluieren. Die bayerische „High-Tech-Offensive“, mit der die Bildung von sogenannten Clustern – regionalen Kooperationen von Unternehmen mit einer Vielzahl von Akteuren aus der Wirtschaft, der Wissenschaft und dem Finanzgewerbe – in fünf Hochtechnologiefeldern gefördert wird, besteht seit 1999. Die Autoren nutzen Daten des ifo Innovationstests, der jährliche Daten über die Innovationstätigkeit von Unternehmen des verarbeitenden Gewerbes in Deutschland liefert, für die Jahre von 1991 bis 2001 – also vor und nach der Einführung der „High-Tech-Offensive“ in Bayern. Die verwendete Stichprobe beinhaltet 709 Unternehmen aus den fünf Technologiefeldern, von denen sich 185 in Bayern und 524 in anderen Bundesländern befinden.

Die Unternehmen aus den anderen deutschen Bundesländern, die nicht von der bayerischen Politikmaßnahme betroffen waren, bilden die Vergleichsgruppe im einfachen DiD-Ansatz. Für die bayerischen Unternehmen und die Vergleichsgruppe wird jeweils berechnet, wie sich die durchschnittlichen Innovationsraten nach 1999 im Vergleich zur Zeit davor verändert haben. Diese beiden Veränderungsraten werden dann miteinander verglichen. Es zeigt sich, dass die betroffenen bayerischen Unternehmen tatsächlich ihre Innovationstätigkeit stärker ausgeweitet haben als die Unternehmen der Vergleichsgruppe. Die zentrale Identifikationsannahme ist hier, dass die Entwicklung der durchschnittlichen Innovationsraten in den bayerischen und nicht-bayerischen Unternehmen ohne die Clusterpolitik identisch gewesen wäre. Da die Daten Informationen über Innovationstätigkeit für mehrere Jahre vor 1999 beinhalten, können die Autoren zeigen, dass die Trends in den beiden Gruppen vor 1999 tatsächlich fast identisch waren, was die Identifikationsannahme plausibel erscheinen lässt.

Es wäre jedoch auch denkbar, dass sich 1999 in Bayern auch die allgemeinen wirtschaftlichen Rahmenbedingungen anderweitig, aber unbeobachtet verändert hätten, mit der Folge eines allgemeinen Anstiegs der Innovationsraten in Bayern. Um diese Möglichkeit auszuschließen, bilden Falck, Heblich und Kipar (2010) ebenfalls einen DiD-Schätzer für (bayerische und nicht-bayerische) Unternehmen, die nicht in den fünf Hochtechnologiefeldern tätig sind. Bayerische Unternehmen außerhalb der fünf Technologiefelder sind nicht von der

Clusterpolitik erfasst. Sollte sich also bei diesem DiD-Schätzer ein Unterschied zeigen, so würde dieser die Effekte unbeobachteter, bayernspezifischer Einflüsse reflektieren, die zu einem allgemeinen Anstieg in der Innovationstätigkeit in Bayern nach 1999 geführt hätten. Es zeigt sich aber, dass sich die Innovationstätigkeit bayerischer Unternehmen außerhalb der geförderten Technologiefelder nicht anders entwickelt hat als jene nicht-bayerischer Unternehmen. Durch einen Vergleich der beiden DiD-Schätzer (innerhalb und außerhalb der geförderten Technologiefelder) miteinander lässt sich der positive Effekt der bayerischen Clusterpolitik auf die Innovationstätigkeit im Rahmen eines sogenannten Differenzen-in-Differenzen-in-Differenzen-Ansatzes (oder Dreifachen-Differenzen-Ansatzes) auch formal bestätigen.

### *5.1.3 Weitere Beispiele und wichtige Aspekte*

Da er relativ leicht einsetzbar ist, wird der DiD-Ansatz in der Politikevaluierung häufig verwendet. So evaluieren Card und Krueger (1994) in einer frühen Studie mit dem DiD-Forschungsdesign die Auswirkungen einer Mindestlohnerhöhung auf die Beschäftigung. Im US-Bundesstaat New Jersey wurde im April 1992 der Mindeststundenlohn von 4,25 $ auf 5,05 $ angehoben. Im angrenzenden Bundesstaat Pennsylvania, der eine ähnliche wirtschaftliche Struktur und Entwicklung hat, blieb der Mindestlohn unverändert. Da Beschäftigte in Fast-Food-Restaurants üblicherweise auf Mindestlohnbasis arbeiten, nutzen die Autoren Beschäftigungsdaten von 410 Fast-Food-Restaurants aus beiden Staaten, erhoben im Februar 1992 (vor der Reform) und November 1992 (nach der Reform). Ein Vergleich der Beschäftigungsentwicklung von Februar bis November 1992 in den Fast-Food-Restaurants in New Jersey mit der in Pennsylvania legt nahe, dass die Erhöhung des Mindestlohns nicht zu einem Beschäftigungsrückgang geführt hat.

Häufig bilden Reformen, die nur einen Teil der Bevölkerung oder der Unternehmen betreffen, in Verbindung mit Längsschnittdaten die Grundlage für ein DiD-Forschungsdesign. So verwenden beispielsweise auch einige deutsche Studien den DiD-Ansatz, um die Auswirkungen von Mindestlöhnen in einzelnen deutschen Sektoren zu untersuchen (z.B. Boockmann u.a. 2013; Frings 2013; König und Möller 2009).

Bauernschuster (2013) nutzt die Änderung des Kündigungsschutzgesetzes zum 1. Januar 2004 in Deutschland, um die Effekte von Kündigungsschutzregelungen auf die Schaffung von Stellen in kleinen Unternehmen zu evaluieren. Hier wird berücksichtigt, dass nur Unternehmen mit 5 bis 10 Mitarbeitern von der Veränderung des Kündigungsschutzes durch die Reform betroffen waren. In weiteren Studien wird das DiD-Design verwendet, um die Effekte der Dauer oder Höhe der Arbeitslosenunterstützung auf das Suchverhalten und den

Arbeitsmarkterfolg von Arbeitslosen zu evaluieren, z.B. Winter-Ebmer (2003) und Lalive und Zweimüller (2004) für Österreich. Der DiD-Ansatz wurde auch verwendet, um Effekte einer Ausweitung des Mutterschaftsurlaubs (Dustmann und Schönberg 2012), der Einführung des Betreuungsgeldes in Thüringen (Gathmann und Sass 2012) oder der Einführung des Rechtsanspruches auf einen Kindergartenplatz (Bauernschuster und Schlotter 2013) in Deutschland zu evaluieren. Bergemann, Fitzenberger und Speckesser (2009) verwenden einen DiD-Schätzer, um die Effekte der Weiterbildung im Rahmen der aktiven Arbeitsmarktpolitik in Deutschland zu prüfen. Schreyögg und Grabka (2010) evaluieren den Effekt der Einführung der Praxisgebühr für gesetzlich Versicherte in Deutschland im Jahr 2004 auf die Häufigkeit von Arztbesuchen, wobei privat Versicherte die Kontrollgruppe bilden. Anger, Kvasnicka und Siedler (2011) nutzen aus, dass das Rauchverbot in Restaurants und Gaststätten in den deutschen Bundesländern zu unterschiedlichen Zeitpunkten eingeführt wurde, um Effekte von Nichtraucherschutzbestimmungen auf das Rauchverhalten zu untersuchen.

Meghir und Palme (2005) evaluieren eine schwedische Bildungsreform aus den vierziger Jahren, mit der die Pflichtschulzeit verlängert, ein einheitlicher Lehrplan eingeführt und die frühe Mehrgliedrigkeit im Schulsystem abgeschafft worden ist. Der DiD-Ansatz ist hier anwendbar, weil die Reform in verschiedenen Bezirken zu verschiedenen Zeitpunkten griff. In ähnlicher Weise nutzen Pekkala Kerr, Pekkarinen und Uusitalo (2013) regional unterschiedliche Einführungszeitpunkte, um eine finnische Schulreform zu evaluieren, mit der in den siebziger Jahren die Zweigliedrigkeit des Schulsystems abgeschafft worden ist. Piopiunik (2013) untersucht die Vorverlegung der Aufteilung zwischen Haupt- und Realschule in Bayern im Jahr 2000 im Vergleich zu den anderen Bundesländern (und zu den Gymnasien). Hübner (2012) und Dwenger, Storck und Wrohlich (2012) untersuchen im Rahmen eines DiD-Ansatzes Effekte von Studiengebühren, indem sie die Einführung von Studiengebühren in einigen Bundesländern zugrunde legen. Um ein italienisches Gesetz zu evaluieren, das die öffentliche Förderung von privaten Investitionen in Krisenregionen regelt, vergleichen Bronzini und de Blasio (2006) das Investitionsverhalten von geförderten und nicht-geförderten Unternehmen vor und nach der Einführung des Förderprogramms.

Häufig werden Politikmaßnahmen nur in einer geringen Anzahl von aggregierten Einheiten eingeführt, beispielsweise in einem einzelnen Bundesland oder in einer einzelnen Region. Um in diesem Fall eine möglichst gut vergleichbare Kontrollgruppe für komparative Fallstudien mit aggregierten Zielgrößen zu erhalten, schlagen Abadie, Diamond und Hainmueller (2010) vor, eine „synthetische“ Kontrollregion zu betrachten. Diese künstliche Region ergibt sich als gewichteter Durchschnitt aller anderen Regionen. Dabei werden die anderen Regionen so

gewichtet, dass sich die Zielgröße vor Einführung der Maßnahme in der künstlichen Region möglichst genauso entwickelt hat wie in der Region, die die Maßnahme einführt. Beispielsweise untersuchen Abadie, Diamond und Hainmueller (2010) mit dieser Methode, wie sich der Zigarettenverkauf in Kalifornien nach der Einführung eines Antirauchergesetzes im Vergleich zu einem synthetischen Vergleichsbundesstaat entwickelt hat.

Der DiD-Ansatz muss jedoch nicht immer eine Zeitdimension enthalten. So untersuchen beispielsweise Boeri und Jimeno (2005), ob strengere Kündigungsschutzregelungen in Italien zu weniger Entlassungen führen. Dabei nutzen sie, dass nur für unbefristet Beschäftigte in Firmen mit mehr als 15 Mitarbeitern ein strengerer Kündigungsschutz gilt. Ein Vergleich der Kündigungswahrscheinlichkeit in Firmen mit mehr oder weniger als 15 Mitarbeitern liefert hier die erste Differenz des DiD-Ansatzes. Der Unterschied zwischen den beiden ersten Differenzen für Beschäftigte mit befristeten und unbefristeten Verträgen ergibt dann den DiD-Schätzer.

Hanushek und Wößmann (2006) vergleichen Leistungsunterschiede zwischen Grundschulen (in denen es noch in keinem Land eine Aufteilung auf verschiedene Schulformen gibt) und Sekundarschulen (in denen manche Länder die Schüler bereits aufgeteilt haben) in Ländern mit und ohne frühe Mehrgliedrigkeit im Schulsystem, um den Effekt der frühen Mehrgliedrigkeit auf Schulleistungen und Ungleichheit im Schulsystem zu untersuchen. Jürges, Schneider und Büchel (2005) wenden den DiD-Ansatz an, um den kausalen Effekt zentraler Abschlussprüfungen auf die Leistung der Schüler zu schätzen. Dabei nutzen sie, dass in einigen deutschen Bundesländern Mathematik im Rahmen einer verpflichtenden zentralen Abschlussprüfung geprüft wird, während dies für Naturwissenschaften nirgendwo gilt. Die beiden betrachteten Differenzen sind folglich zum einen die Schulfächer, zum zweiten die Bundesländer. Schwerdt und Wößmann (2014) untersuchen den Informationswert von Noten in zentralen Abschlussprüfungen auf dem Arbeitsmarkt, indem sie Einkommensunterschiede zwischen Personen mit guten und schlechten Abiturnoten (erste Differenz) zwischen Bundesländern mit und ohne zentrale Abschlussprüfungen (zweite Differenz) vergleichen. Hanushek, Wößmann und Zhang (2011) untersuchen, wie sich die Arbeitsmarkteffekte von allgemeiner und berufsspezifischer Bildung im Lebensverlauf verändern, indem sie die Beschäftigungsquoten von jüngeren und älteren Personen (erste Differenz) mit allgemeiner und berufsspezifischer Ausbildung (zweite Differenz) vergleichen.

## *5.2 Panelmethoden mit fixen Effekten: Informationen im Überfluss*

Die sechste Gruppe an Evaluationsverfahren schließlich verallgemeinert den vorangegangenen Ansatz und erfordert in der Regel große Datensätze, die die Berücksichtigung unbeobachteter individueller fixer Effekte ermöglichen.[21]

### *5.2.1 Idee und Intuition*

Umfangreiche Paneldatensätze ermöglichen eine – über das Ausmaß traditioneller DiD-Ansätze hinaus – eingehendere Behandlung von Fällen, in denen unbeobachtete Unterschiede über die Zeit (oder eine andere Dimension) konstant bleiben. Dabei können über eine Vorher-Nachher-Betrachtung hinaus viele Perioden und über die ausschließliche Ja-Nein-Ausprägung der Behandlung hinaus stetige Unterschiede im Grad, zu dem Individuen einer Maßnahme ausgesetzt sind, berücksichtigt werden.

Zu den prominentesten Faktoren, die Forscher beispielsweise bei der Evaluierung von Maßnahmen, die Individuen fördern sollen, üblicherweise nicht beobachten können, gehören Fähigkeiten und Persönlichkeitsmerkmale. Bei vielen Fragen der empirischen Sozialforschung stellt ein möglicher „Ability bias“ – eine Verzerrung der Ergebnisse aufgrund der unbeobachteten Fähigkeiten – ein typisches Problem dar. Wenn man zum Beispiel evaluieren möchte, ob eine Weiterbildungsmaßnahme den Arbeitsmarkterfolg der Teilnehmer verbessert, muss man ausschließen können, dass die Ergebnisse durch Unterschiede in den Fähigkeiten von Teilnehmern und Nicht-Teilnehmern zustande kommen. Mangels eines klaren und umfassenden Maßes für die Fähigkeiten lässt sich die Verzerrung durch unbeobachtete individuelle Heterogenität ohne kontrolliertes oder natürliches Experiment in einer Querschnittsbetrachtung zu einem einzigen Zeitpunkt nur schwer aufheben.

Wenn aber der Einflussfaktor, der – wie die individuellen Fähigkeiten – zur unbeobachteten Heterogenität führt, über die Zeit unverändert (konstant) ist, so kann er „fixiert“ werden, indem man in der Analyse jede Variation *zwischen* den beobachteten Einheiten ignoriert und nur Veränderungen über die Zeit, also „innerhalb“ der Einheiten betrachtet. Zu diesem Zweck werden Paneldaten benötigt, die es ermöglichen, die gleiche Einheit zu mehreren Zeitpunkten zu beobachten. Fixe Effekte jeder Beobachtungseinheit lassen sich herausrechnen, indem man mit einer Indikatorvariable in den Schätzungen die mittleren Unterschiede *zwischen* den Einheiten berücksichtigt, so dass nur noch *Veränderungen* der Einfluss- und Ergebnisgrößen über die Zeit verwendet werden, um den interessierenden Effekt zu identifizieren. Auf diese

[21] Eine umfassende Einführung in Panelmethoden bietet das Lehrbuch von Baltagi (2013).

Weise kann man mit unbeobachteter, jedoch zeitlich unveränderlicher Heterogenität zwischen den Beobachtungseinheiten umgehen. Bei komplexeren Schätzmodellen können solche fixen Effekte beispielsweise auf Ebene der Beschäftigten, der Unternehmen, der Branche und der Region eingeführt werden. Wie beim DiD-Ansatz muss bei den generelleren Panelmethoden mit fixen Effekten die fixierte Dimension nicht unbedingt die Zeit sein. Auch familienspezifische fixe Effekte bei mehreren Geschwistern oder schülerspezifische fixe Effekte bei mehreren Schulfächern kommen in Frage.

### *5.2.2 Eine Beispielstudie: Hermes-Bürgschaften für Exportkredite*

Ein Beispiel für die Anwendung von Panelmethoden mit fixen Effekten ist die Studie von Felbermayr und Yalcin (2013), in der Hermes-Bürgschaften in Deutschland evaluiert werden. Bei Hermesdeckungen handelt es sich um eine Exportkreditgarantie, die der Staat Exportunternehmen zur Verfügung stellt, um Handelsgeschäfte in risikoreichen Ländern zu ermöglichen. Durch diese Garantie sollen Zahlungsausfallrisiken gedeckt werden. In der Folge sollen Exportgeschäfte stattfinden, für die es andernfalls aufgrund des hohen Risikos keine Finanzierung gäbe. Die Wissenschaftler untersuchen, ob Hermesdeckungen tatsächlich Exportgeschäfte anregen und somit zu mehr Beschäftigung in den exportierenden Unternehmen führen.

Um den kausalen Effekt der Hermesdeckungen zu identifizieren, müssen unbeobachtete Einflussfaktoren berücksichtigt werden, die sowohl die Inanspruchnahme einer Hermesdeckung als auch die Exporttätigkeit von Unternehmen beeinflussen. So kann es sein, dass Unternehmen einer bestimmten exportorientierten Branche eher risikoavers sind und daher tendenziell öfter Hermes-Bürgschaften in Anspruch nehmen. Typischerweise kann die Risikoaversion von Unternehmen allerdings nicht oder nur schlecht beobachtet werden, und wenn sie mit der interessierenden Ergebnisvariable zusammenhängt, würde sich in einer einfachen Querschnittsanalyse eine Verzerrung durch ausgelassene Variablen ergeben. Alternativ könnten Hermes-Bürgschaften gerade dann in Anspruch genommen werden, wenn ein bestimmtes Zielland seit Jahrzehnten übermäßig viel importiert, so dass genau aus diesem Grund Zahlungsausfälle drohen. In diesem Falle würde es aufgrund umgekehrter Kausalität zu einer Überschätzung des Effektes der Hermes-Bürgschaften auf die Exporte kommen.

Felbermayr und Yalcin (2013) nutzen einen Panelansatz, um Verzerrungen aufgrund solcher Endogenitätsprobleme zu eliminieren. Der verwendete Paneldatensatz enthält drei Dimensionen: die Branche, das Zielland der Exporte und das Jahr. Der Einfachheit halber sei ein Exportmarkt durch eine bestimmte Kombination aus Branche und Zielland definiert. In einem multivariaten Regressionsmodell werden nun fixe Effekte für jeden Exportmarkt

aufgenommen. Dadurch werden alle über die Zeit konstanten (beobachteten wie unbeobachteten) Einflüsse eines Exportmarktes auf die Exporte herausgerechnet. Der geschätzte Effekt der Hermesdeckungen wird in diesem Fall nur durch die zeitliche Variation (von 2000 bis 2009) in der Exporthöhe innerhalb eines Exportmarktes identifiziert.

Diese Effekte für jeden Exportmarkt sind fixe Effekte auf der Ebene der Paneldimension. Zusätzlich ist es oftmals noch sinnvoll, weitere fixe Effekte auf anderer Aggregationsebene aufzunehmen. Beispielsweise können innerhalb einer Branche andere unbeobachtete Einflüsse einen vom Zielland unabhängigen Anstieg der Exporte über die Zeit und der Hermes-Bürgschaften verursachen. Daher nehmen Felbermayr und Yalcin (2013) ebenfalls fixe Effekte für jede Kombination von Branche und Jahr in ihr Schätzmodell auf. Letztlich werden für jede binäre Interaktion der drei Dimensionen – Branche, Zielland und Jahr – fixe Effekte in das multivariate Regressionsmodell aufgenommen. Allein die Aufnahme fixer Effekte für die tertiäre Interaktion der drei Dimensionen ist natürlich nicht möglich, da genau die Variation auf dieser Ebene für die Identifikation des Effektes genutzt wird.

Die Ergebnisse der Studie zeigen, dass eine Erhöhung der Hermesdeckungen um ein Prozent die Exporte im Durchschnitt in der Tat um 0,012 Prozent steigert, wobei allerdings nur weniger als 3 Prozent des deutschen Exportmarktes von den Garantien abgedeckt werden. Ohne Berücksichtigung der fixen Effekte würde fälschlicherweise ein deutlich höherer Effekt geschätzt.

### *5.2.3 Weitere Beispiele und wichtige Aspekte*

Fixe Effekte für jede Beobachtungseinheit, die es ermöglichen, zeitinvariante Unterschiede zwischen den Beobachtungeinheiten herauszurechnen, sind der klassische Fall des Panelansatzes mit fixen Effekten. Dabei kann die Beobachtungsebene sehr unterschiedlich sein. Buettner und Wamser (2013) nutzen beispielsweise ein Panelmodel mit konzernspezifischen fixen Effekten, um zu zeigen, dass Steuerunterschiede zwischen Ländern lediglich einen kleinen Einfluss auf die interne Gewinnverschiebung in multinationalen Konzernen haben. Die Identifikation basiert also auf einer konzernspezifischen Variation der Steuerunterschiede zwischen den Konzerntöchtern in unterschiedlichen Ländern über die Zeit. Die Beobachtungsebene bei Bhattacharya, Gathmann und Miller (2013) sind russische Provinzen: Die Studie zeigt, dass die erhöhte Sterberate in Russland zwischen 1990 und 1994 auf die Beendigung von Gorbatschows Anti-Alkohol-Kampagne 1988 zurückzuführen ist. Um zeitinvariante Unterschiede in den Sterberaten auszublenden, werden provinzspezifische fixe Effekte berücksichtigt. Aichele und Felbermayr (2014) untersuchen Handelseffekte des Kyoto-

Protokolls, indem sie bilaterale Handelsströme auf Industrieebene beobachten und fixe Effekte für Länder und Jahre berücksichtigen.

Riphahn (2012) berücksichtigt fixe Effekte für Bundesländer und Geburtskohorten, um zu untersuchen, wie sich die Abschaffung von Schulgebühren für Gymnasien, die in verschiedenen Bundesländern zu verschiedenen Zeitpunkten zwischen 1947 und 1962 stattfand, auf Gymnasialabschlüsse ausgewirkt hat. Reinhold und Jürges (2010) verwenden dieselbe Variation in einem Instrumentvariablen-Ansatz, um den Effekt der durch die Gebührenabschaffung bedingten höheren Bildung auf Rauchverhalten und Übergewicht zu schätzen. Hanushek, Link und Wößmann (2013) nutzen alle Wellen des PISA-Tests von 2000 bis 2009, um anhand eines Modells mit fixen Länder- und Zeiteffekten die Auswirkung von Reformen in der Schulautonomie auf die Schülerleistungen zu schätzen.

Wenn Paneldaten Zeitpunkte enthalten, zu denen alle Beobachtungseinheiten noch nicht von der Maßnahme betroffen sind, kann oft ein einfacher Differenzen-in-Differenzen-Ansatz geschätzt werden. Alternativ kann eine Kombination beider Ansätze eingesetzt werden. Dabei werden fixe Effekte für alle Beobachtungseinheiten anstelle eines einfachen Indikators für die Behandlungsgruppe aufgenommen. Ein Indikator für die Behandlungsgruppe wird dann lediglich in Verbindung mit einem Indikator für Beobachtungszeitpunkte nach Einführung der Politikmaßnahme in das Modell aufgenommen. Neuere Studien verbinden ebenfalls Panelanalysen mit Matching-Verfahren, die auf Beobachtungen der Ergebnisvariable vor dem Behandlungszeitpunkt basieren. Hierbei wird eine Vergleichsgruppe gebildet, die einen vergleichbaren Verlauf der Ergebnisvariablen vor Einführung der Politikmaßnahme aufweist. Dieser Ansatz wird häufig bei der Evaluierung von öffentlichen Weiterbildungsprogrammen verwendet (z.B. Lechner, Miquel und Wunsch 2011; Stenberg, de Luna und Westerlund 2012; Biewen u.a. 2014).

Die geschätzten fixen Effekte können oftmals auch selbst von Interesse sein. Neue empirische Studien belegen beispielsweise, dass die Qualität von Lehrern einen erheblichen Einfluss auf Schülerleistungen hat. Dabei hängt die Qualität der Lehrer aber kaum mit leicht beobachtbaren Lehrereigenschaften wie Ausbildung oder Berufserfahrung zusammen. Um einen Indikator für Lehrerqualität zu gewinnen, werden in Studien aus den Vereinigten Staaten sehr große administrative Datensätze verwendet, die jährliche Informationen über Testleistungen der Schüler enthalten und eine Lehrer-Schüler-Zuordnung erlauben. Auf Basis dieser Daten werden Regressionsmodelle geschätzt, die Schülerleistungen auf fixe Schüler- und Lehrer-

effekte regressieren.[22] Die fixen Schülereffekte schließen damit zeitinvariante Unterschiede in der Leistungsfähigkeit von Schülern aus. Die geschätzten Lehrereffekte dienen dann als Indikator für die allgemeine Effektivität eines Lehrers, der nicht durch unbeobachtete und zeitinvariante Unterschiede zwischen Schülern verzerrt ist.

Wie im Fall des DiD-Ansatzes können Panelmethoden mit fixen Effekten aber auch verwendet werden, wenn Beobachtungeinheiten zwar nicht zu verschieden Zeitpunkten, aber entlang anderer Dimensionen wiederholt beobachtet werden. Beispielsweise nutzen einige Wissenschaftler in ihren Studien aus, dass man die Leistungen einzelner Schüler in mehreren Fächern kennt, um die Bedeutsamkeit verschiedener Lehrereigenschaften wie Geschlecht, Lehrmethoden oder Fachwissen für die Leistungen des Schülers zu erfassen (z.B. Dee 2005; Schwerdt und Wuppermann 2011; Metzler und Wößmann 2012). Durch fixe Effekte für Schüler werden in diesem Fall unbeobachtete Unterschiede in der fächerübergreifenden Leistungsfähigkeit von Schülern herausgerechnet.

Einige Evaluationsstudien auf dem Arbeitsmarkt verwenden Informationen über Geschwister oder sogar eineiige Zwillinge, um durch fixe Familieneffekte implizit unbeobachtete Einflüsse des familiären Hintergrunds oder genetische Unterschiede auszuschließen. So lassen sich beispielsweise Effekte von frühkindlichen Entwicklungen und verschiedenen Bildungsmaßnahmen auf weiterführende Bildung und auf den Arbeitsmarkterfolg analysieren (z.B. Card 1999; Garces, Thomas und Currie 2002; Black, Devereux und Salvanes 2007; Schlotter 2011; Figlio u.a. 2013). Während Geschwister- und Zwillingsstudien geeignet sind, viele Probleme aufgrund ausgelassener Variablen zu lösen, schränken weitere Punkte doch die Identifikation ein. Grundsätzlich sollte beispielsweise bei Zwillingsstudien in Frage gestellt werden, ob es überzeugend ist, die Quelle der Variation als exogen – also unabhängig vom betrachteten Zusammenhang – anzusehen. Wenn zwei vermeintlich identische Zwillinge mit demselben familiären Umfeld ein unterschiedliches Ausmaß an Bildung erhalten, scheint es wahrscheinlich, dass sie sich letztendlich doch in gewissen Merkmalen unterscheiden, die der Wissenschaftler nicht beobachten kann. Darüber hinaus kann die Aufnahme fixer Effekte das Problem von Messfehlern in erklärenden Variablen, das zu einer Unterschätzung der tatsächlichen Effekte führt, verstärken (Ashenfelter und Krueger 1994). Schließlich ist nicht klar, inwieweit Ergebnisse, die auf Zwillingspaaren basieren, die gemeinsam in einer spezifischen

[22] Z.B. Rockoff (2004); Rivkin, Hanushek und Kain (2005); Chetty, Friedman und Rockoff (2011); ein Überblick findet sich bei Hanushek und Rivkin (2012).

Situation aufgewachsen sind, auf die gesamte Bevölkerung übertragen werden können. Hier kann es also wieder an externer Validität mangeln.

Trotz der beachtlichen Fortschritte, die in den vergangenen Jahren gemacht wurden, bleibt festzuhalten, dass Panelmethoden nur insofern unbeobachtete Heterogenität berücksichtigen können, als die relevanten unbeobachteten Charakteristika nicht systematisch über die Paneldimensionen variieren. Unbeobachtete Einflüsse, die zeitlich (oder über Fächer oder Zwillinge hinweg) veränderlich sind, können nicht allein durch übliche Panelmethoden beseitigt werden. Sie erfordern experimentelle Methoden.

## 6. Schlussbemerkungen: Der Bedarf an zusätzlicher Politikevaluierung

Nicht zuletzt in Zeiten knapper Kassen besteht ein Bedarf zu lernen, wo öffentliche Mittel zielführend eingesetzt werden. Die einfache politische Frage lautet: Was funktioniert und was nicht? Wenn Politikmaßnahmen aber dafür bestimmt sind, Ergebnisse zu verbessern, sollten sie nicht auf Wunschdenken beruhen, sondern auf nachgewiesener Effektivität. Überzeugende Evidenz über die Effekte bestimmter politischer Maßnahmen zu erhalten, ist aber keine leichte Aufgabe. Als erste Voraussetzung müssen relevante Daten über mögliche Ergebnisse erhoben werden. Des Weiteren muss klar sein, dass empirische Evidenz, die eine einfache Korrelation zwischen einer Politikmaßnahme und einer Zielgröße offenlegt, nicht als hinreichender Beweis für einen Effekt der Maßnahme angesehen werden kann. Um zu verstehen, welche Auswirkungen politische Maßnahmen hätten, spielen reine Korrelationen keine Rolle, sondern ausschließlich kausale Zusammenhänge. Politiker interessiert, was geschieht, wenn eine bestimmte Politikmaßnahme ergriffen wird. Stellen sich die erhofften Effekte tatsächlich ein? Als Grundlage für eine evidenzbasierte Politik werden Antworten auf diese kausalen Fragestellungen benötigt. Die im vorliegenden Beitrag vorgestellten ökonometrischen Methoden können dabei helfen, solchen Antworten näherzukommen.

Selbstverständlich ist nicht jede Anwendung dieser Methoden automatisch überzeugend. Immer muss gefragt werden, ob die jeweilige Einteilung in Behandlungs- und Vergleichsgruppe im jeweiligen Anwendungsfall „so gut wie zufällig“ ist – ob also die jeweilige Annahme, die den interessierenden Effekt identifiziert, glaubwürdig ist. Umgekehrt sind die hier vorgestellten Methoden auch nicht die einzigen, die eine kausale Evaluierung erlauben. Wann immer experimentelle oder quasi-experimentelle Forschungsdesigns nicht einsetzbar sind, mögen auch klassische multivariate Regressionsanalysen wichtige Information über deskriptive Zusammenhänge liefern. Aber gerade die hier vorgestellten Methoden helfen zu verstehen,

inwiefern die Ergebnisse klassischer Verfahren von den tatsächlichen kausalen Effekten abweichen können und insofern vorsichtig interpretiert werden müssen.

Insgesamt werden die modernen Evaluierungsmethoden in der deutschen Wirtschaftspolitik noch viel zu selten eingesetzt. Wenn Politik effektiver und effizient gestaltet werden soll, dann müssen alle wichtigen politischen Maßnahmen überzeugender Evaluierung unterzogen werden. So kann sowohl die Politik als auch die Wissenschaft im Zeitablauf lernen, wie sich Politik und auch Evaluierung verbessern lassen. Für eine überzeugende Evaluierung ist es dabei unumgänglich, dass die Forschung unabhängig und ergebnisoffen, transparent und nachvollziehbar erfolgt (vgl. dazu Wissenschaftlicher Beirat beim Bundesministerium für Wirtschaft und Energie 2013; Kirchgässner 2013; Schmidt 2009; sowie den Ethikkodex des Vereins für Socialpolitik). Sind die Daten, Programme und Ergebnisberichte der Evaluierungen für andere Wissenschaftler zugänglich, so lassen sich Evaluierungsergebnisse replizieren und die Methoden gegebenenfalls im wissenschaftlichen Wettbewerb weiter verbessern. Auch im politischen Umfeld muss eine Evaluierungskultur entstehen, die erlaubt, negative Evaluationsbefunde nicht als etwas grundsätzlich Negatives zu betrachten, sondern als Erkenntnisgewinn. Solche politischen wie wissenschaftlichen Lernprozesse können am Ende dazu beitragen, dass Wirtschaftspolitik tatsächlich der Verbesserung der wirtschaftlichen und sozialen Lage der Betroffenen dient.

# Literatur


Abadie, Alberto, Alexis Diamond und Jens Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105 (490): 493-505.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane und Parag A. Pathak (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics* 126 (2): 699-748.

Aichele, Rahel und Gabriel Felbermayr (2012). Kyoto and the carbon footprint of nations. *Journal of Environmental Economics and Management* 63 (3): 336-354.

Aichele, Rahel und Gabriel Felbermayr (2014). Kyoto and carbon leakage: An empirical analysis of the carbon content of bilateral trade. *Review of Economics and Statistics*: forthcoming.

Allmendinger, Jutta und Annette Kohlmann (2005). Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung. *Allgemeines Statistisches Archiv* 88 (2): 159-182.

Anger, Silke, Michael Kvasnicka und Thomas Siedler (2011). One last puff? Public smoking bans and smoking behavior. *Journal of Health Economics* 30 (3): 591-601.

Angrist, Joshua D. (2004). American education research changes tack. *Oxford Review of Economic Policy* 20 (2): 198-212.

Angrist, Joshua D., Eric Bettinger und Michael Kremer (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review* 96 (3): 847-862.

Angrist, Joshua D., Guido W. Imbens und Donald B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91 (434): 444-455.

Angrist, Joshua D. und Alan B. Krueger (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics*, hrsg. von Orley Ashenfelter und David Card. Amsterdam: North Holland: 1277-1366.

Angrist, Joshua D. und Alan B. Krueger (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15 (4): 69-85.

Angrist, Joshua D. und Victor Lavy (1999). Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114 (2): 533-575.

Angrist, Joshua D. und Jörn-Steffen Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D. und Jörn-Steffen Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24 (2): 3-30.

Angrist, Joshua und Victor Lavy (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review* 99 (4): 1384-1414.

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

Ashenfelter, Orley und Alan B. Krueger (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* 84 (5): 1157-1173.

Augurzky, Boris, Thomas K. Bauer, Arndt R. Reichert, Christoph M. Schmidt und Harald Tauchmann (2012). Does money burn fat? Evidence from a randomized experiment. IZA Discussion Paper 6888. Bonn: Institute for the Study of Labor.

Baltagi, Badi H. (2013). *Econometric analysis of panel data*. Fifth ed. Chichester: John Wiley & Sons.

Banerjee, Abhijit V. und Esther Duflo (2009). The experimental approach to development economics. *Annual Review of Economics* 1 (1): 151-178.

Bauer, Thomas K., Stefan Bender, Alfredo R. Paloyo und Christoph M. Schmidt (2012). Evaluating the labor-market effects of compulsory military service. *European Economic Review* 56 (4): 814-829.

Bauer, Thomas K., Michael Fertig und Christoph M. Schmidt (2009). *Empirische Wirtschaftsforschung: Eine Einführung*. Berlin: Springer.

Bauernschuster, Stefan (2013). Dismissal protection and small firms' hirings: Evidence from a policy reform. *Small Business Economics* 40 (2): 293-307.

Bauernschuster, Stefan, Oliver Falck und Ludger Wößmann (2011). Surfing alone? The Internet and social capital: Evidence from an unforeseeable technological mistake. CESifo Working Paper 3469. Munich: CESifo.

Bauernschuster, Stefan und Martin Schlotter (2013). Public child care and mothers' labor supply: Evidence from two quasi-experiments. CESifo Working Paper 4191. Munich: CESifo.

Becker, Sascha O., Peter H. Egger und Maximilian von Ehrlich (2010). Going NUTS: The effect of EU Structural Funds on regional performance. *Journal of Public Economics* 94 (9-10): 578-590.

Becker, Sascha O., Peter H. Egger und Maximilian von Ehrlich (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy* 5 (4): 29-77.

Bedard, Kelly und Elizabeth Dhuey (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121 (4): 1437-1472.
Belfield, Clive R., Milagros Nores, Steve W. Barnett und Lawrence J. Schweinhart (2006). The High/Scope Perry Preschool Program. *Journal of Human Resources* 41 (1): 162-190.
Bergemann, Annette, Bernd Fitzenberger und Stefan Speckesser (2009). Evaluating the dynamic employment effects of training programs in East Germany using conditional difference-in-differences. *Journal of Applied Econometrics* 24 (5): 797-823.
Bertrand, Marianne, Esther Duflo und Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 114 (1): 249-275.
Besley, Timothy und Anne Case (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *Economic Journal* 110 (467): F672-F694.
Bettinger, Eric P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics* 94 (3): 686-698.
Bhattacharya, Jay, Christina Gathmann und Grant Miller (2013). The Gorbachev anti-alcohol campaign and Russia's mortality crisis. *American Economic Journal: Applied Economics* 5 (2): 232-260.
Biewen, Martin, Bernd Fitzenberger, Aderonke Osikominu und Marie Paul (2014). The effectiveness of public sponsored training revisited: The importance of data and methodological choices. *Journal of Labor Economics* forthcoming.
Black, Sandra E., Paul J. Devereux und Kjell G. Salvanes (2005). Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *American Economic Review* 95 (1): 437-449.
Black, Sandra E., Paul J. Devereux und Kjell G. Salvanes (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics* 122 (1): 409-439.
Boeri, Tito und Juan F. Jimeno (2005). The effects of employment protection: Learning from variable enforcement. *European Economic Review* 49 (8): 2057-2077.
Boockmann, Bernhard, Raimund Krumm, Michael Neumann und Pia Rattenhuber (2013). Turning the switch: An evaluation of the minimum wage in the German electrical trade using repeated natural experiments. *German Economic Review* 14 (3): 316-348.
Bouguen, Adrien und Marc Gurgand (2012). Randomized controlled experiments in education. EENEE Analytical Report 11. http://www.eenee.org/doc/eenee_ar11.pdf (30.6.2013).
Brinkmann, Christian, Reinhard Hujer und Susanne Koch (2006). Evaluation aktiver Arbeitsmarktpolitik in Deutschland - eine Einführung. *Zeitschrift für ArbeitsmarktForschung* 39 (3/4): 319-327.
Bronzini, Raffaello und Guido de Blasio (2006). Evaluating the impact of investment incentives: The case of Italy's Law 488/1992. *Journal of Urban Economics* 60 (2): 327-349.
Brunello, Giorgio, Margherita Fort und Guglielmo Weber (2009). Changes in compulsory schooling, education and the distribution of wages in Europe. *Economic Journal* 119 (536): 516-539.
Buettner, Thiess (2006). The incentive effect of fiscal equalization transfers on tax policy. *Journal of Public Economics* 90 (3): 477-497.
Buettner, Thiess und Georg Wamser (2013). Internal debt and multinational profit shifting: Empirical evidence from firm-level panel data. *National Tax Journal* 66 (1): 63-95.
Bundesministerium für Arbeit und Soziales und Institut für Arbeitsmarkt- und Berufsforschung (2011). *Sachstandsbericht der Evaluation der Instrumente*. Berlin: Bundesministerium für Arbeit und Soziales. http://www.bmas.de/SharedDocs/Downloads/DE/arbeitsmarktpol_instr_iab_studie.pdf?__blob=publication File (30.6.2013).
Caliendo, Marco, Konstantinos Tatsiramos und Arne Uhlendorff (2013). Benefit duration, unemployment duration and job match quality: A regression-discontinuity approach. *Journal of Applied Econometrics* 28 (4): 604-627.
Card, David (1999). The causal effect of education on earnings. In *Handbook of Labor Economics*, hrsg. von Orley Ashenfelter und David Card. Amsterdam: North-Holland: 1801-1863.
Card, David, Jochen Kluve und Andrea Weber (2010). Active labour market policy evaluations: A meta-analysis. *Economic Journal* 120 (548): F452-F477.
Card, David und Alan B. Krueger (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84 (4): 772-793.
Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach und Danny Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126 (4): 1593-1660.
Chetty, Raj, John N. Friedman und Jonah E. Rockoff (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper 17699. Cambridge, MA: National Bureau of Economic Research (December).
Criscuolo, Chiara, Ralf Martin, Henry Overman und John Van Reenen (2012). The causal effects of an industrial policy. NBER Working Paper 17842. Cambridge, MA: National Bureau of Economic Research.

Cullen, Julie Berry, Brian A Jacob und Steven Levitt (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica* 74 (5): 1191-1230.
Deaton, Angus (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature* 48 (2): 424-455.
Dee, Thomas S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review* 95 (2): 158-165.
Deming, David J., Justine S. Hastings, Thomas J. Kane und Douglas O. Staiger (2014). School choice, school quality and postsecondary attainment. *American Economic Review* 104 (3): 991-1013.
DiNardo, John und David S. Lee (2011). Program evaluation and research designs. In *Handbook of Labor Economics, Volume 4, Part A*, hrsg. von Ashenfelter Orley und Card David. Amsterdam: North Holland: 463-536.
Dustmann, Christian und Uta Schönberg (2012). Expansions in maternity leave coverage and children's long-term outcomes. *American Economic Journal: Applied Economics* 4 (3): 190-224.
Dwenger, Nadja, Johanna Storck und Katharina Wrohlich (2012). Do tuition fees affect the mobility of university applicants? Evidence from a natural experiment. *Economics of Education Review* 31 (1): 155-167.
Fairlie, Robert W., Dean Karlan und Jonathan Zinman (2012). Behind the GATE Experiment: Evidence on Effects of and Rationales for Subsidized Entrepreneurship Training. NBER Working Paper 17804. Cambridge, MA: National Bureau of Economic Research.
Falck, Oliver, Stephan Heblich und Stefan Kipar (2010). Industrial innovation: Direct evidence from a cluster-oriented policy. *Regional Science and Urban Economics* 40 (6): 574-582.
Felbermayr, Gabriel J. und Erdal Yalcin (2013). Export credit guarantees and export performance: An empirical analysis for Germany. *The World Economy* 36 (8): 967-999.
Figlio, David N., Jonathan Guryan, Krzysztof Karbownik und Jeffrey Roth (2013). The effects of poor neonatal health on children's cognitive development. NBER Working Paper 18846. Cambridge, MA: National Bureau of Economic Research.
Finn, Jeremy D. und Charles M. Achilles (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal* 27 (3): 557-577.
Fredriksson, Peter, Björn Öckert und Hessel Oosterbeek (2013). Long-term effects of class size. *Quarterly Journal of Economics* 128 (1): 249-285.
Frings, Hanna (2013). The employment effect of industry-specific, collectively bargained minimum wages. *German Economic Review* 14 (3): 258-281.
Frölich, Markus und Michael Lechner (2010). Exploiting regional treatment intensity for the evaluation of labor market policies. *Journal of the American Statistical Association* 105 (491): 1014-1029.
Fryer, Roland G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics* 126 (4): 1755-1798.
Fryer, Roland G., Jr., Steven D. Levitt, John List und Sally Sadoff (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper 18237. Cambridge, MA: National Bureau of Economic Research.
Garces, Eliana, Duncan Thomas und Janet Currie (2002). Longer-term effects of Head Start. *American Economic Review* 92 (4): 999-1012.
Garibaldi, Pietro, Francesco Giavazzi, Andrea Ichino und Enrico Rettore (2012). College cost and time to complete a degree: Evidence from tuition discontinuities. *Review of Economics and Statistics* 94: 699-711.
Gathmann, Christina und Björn Sass (2012). Taxing childcare: Effects on family labor supply and children. CESifo Working Paper 3776. Munich: CESifo.
Gorter, Cees und Guyonne R. J. Kalb (1996). Estimating the effect of counseling and monitoring the unemployed using a job search model. *Journal of Human Resources* 31 (3): 590-610.
Graversen, Brian K. und Jan van Ours (2008). How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation program. *Journal of Public Economics* 92 (10-11): 2020-2035.
Graversen, Brian K. und Jan van Ours (2011). An activation program as a stick to job finding. *LABOUR* 25 (2): 167-181.
Hægeland, Torbjørn, Oddbjørn Raaum und Kjell G. Salvanes (2012). Pennies from heaven? Using exogenous tax variation to identify effects of school resources on pupil achievement. *Economics of Education Review* 31 (5): 601-614.
Hägglund, Pathric (2009). Experimental evidence from intensified placement efforts among unemployed in Sweden. IFAU Working Paper 2009:16. Uppsala: IFAU - Institute for Labour Market Policy Evaluation.
Hahn, Jinyong, Petra Todd und Wilbert Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69 (1): 201-209.
Hanushek, Eric A., Susanne Link und Ludger Wößmann (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics* 104: 212-232.
Hanushek, Eric A. und Steven G. Rivkin (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics* 4: 7.1-7.27.

Hanushek, Eric A. und Ludger Wößmann (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal* 116 (510): C63-C76.
Hanushek, Eric A., Ludger Wößmann und Lei Zhang (2011). General education, vocational education, and labor-market outcomes over the life-cycle. NBER Working Paper 17504. Cambridge, MA: National Bureau of Economic Research (October).
Harmon, Colm und Ian Walker (1995). Estimates of the economic return to schooling for the United Kingdom. *American Economic Review* 85 (5): 1278-1286.
Harrison, Glenn W. und John A. List (2004). Field experiments. *Journal of Economic Literature* 42 (4): 1009-1055.
Haynes, Laura, Owain Service, Ben Goldacre und David Torgerson (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. London: Cabinet Office Behavioural Insights Team.
Heckman, James J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature* 48 (2): 356-398.
Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev und Adam Yavitz (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Journal of Quantitative Economics* 1 (1): 1-46.
Holland, Paul W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81 (396): 945-960.
Hoxby, Caroline Minter (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115 (3): 1239-1285.
Hübner, Malte (2012). Do tuition fees affect enrollment behavior? Evidence from a 'natural experiment' in Germany. *Economics of Education Review* 31 (6): 949-960.
Ichino, Andrea und Regina T. Riphahn (2005). The effect of employment protection on worker effort: Absenteeism during and after probation. *Journal of the European Economic Association* 3 (1): 120-143.
Imbens, Guido W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48 (2): 399-423.
Imbens, Guido W. und Karthik Kalyanaraman (2012). Optimal bandwidth choice for the regression eiscontinuity estimator. *Review of Economic Studies* 79 (3): 933-959.
Imbens, Guido W. und Jeffrey M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47 (1): 5-86.
Jürges, Hendrik, Kerstin Schneider und Felix Büchel (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association* 3 (5): 1134-1155.
Kemptner, Daniel, Hendrik Jürges und Steffen Reinhold (2011). Changes in compulsory schooling and the causal effect of education on health: Evidence from Germany. *Journal of Health Economics* 30 (2): 340-354.
Kirchgässner, Gebhard (2013). Zur Rolle der Ökonometrie in der wissenschaftlichen Politikberatung. *Perspektiven der Wirtschaftspolitik* 14 (1-2): 3-30.
Klepinger, Daniel H., Terry R. Johnson und Jutta M. Joesch (2002). Impacts of unemployment insurance work-search requirements: The Maryland experience. *Industrial and Labor Relations Review* 56 (1): 3-22.
Kling, Jeffrey R., Jeffrey B. Liebman und Lawrence F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica* 75 (1): 83-119.
König, Marion und Joachim Möller (2009). Impacts of minimum wages: A micro data analysis for the German construction sector. *International Journal of Manpower* 30 (7): 716-741.
Krueger, Alan B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics* 114 (2): 497-532.
Krueger, Alan B. und Pei Zhu (2004). Another look at the New York City school voucher experiment. *American Behavioral Scientist* 47 (5): 658-698.
Lalive, Rafael (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics* 142 (2): 785-806.
Lalive, Rafael und Josef Zweimüller (2004). Benefit entitlement and unemployment duration: The role of policy endogeneity. *Journal of Public Economics* 88 (12): 2587-2616.
Lavy, Victor (2010). Effects of free choice among public schools. *Review of Economic Studies* 77 (3): 1164-1191.
Lechner, Michael, Ruth Miquel und Conny Wunsch (2011). Long-run effects of public sector sponsored training in West Germany. *Journal of the European Economic Association* 9 (4): 742-784.
Lee, David S. und Thomas Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* 48 (2): 281-355.
Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek und Dinand Webbink (2007). The effect of extra funding for disadvantaged pupils on achievement. *Review of Economics and Statistics* 89 (4): 721-736.
Leuven, Edwin, Hessel Oosterbeek und Bas van der Klaauw (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association* 8 (6): 1243-1265.

Levitt, Steven D., John A. List, Susanne Neckermann und Sally Sadoff (2012). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. NBER Working Paper 18165. Cambridge, MA: National Bureau of Economic Research.

List, John A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives* 25 (3): 3-16.

List, John A. und Imran Rasul (2011). Field experiments in labor economics. In *Handbook of Labor Economics*, hrsg. von Ashenfelter Orley und Card David: Elsevier: 103-228.

Lleras-Muney, Adriana (2005). The relationship between education and adult mortality in the United States. *Review of Economic Studies* 72 (1): 189-221.

Lochner, Lance und Enrico Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review* 94 (1): 155-189.

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling und Lisa Sanbonmatsu (2013). Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity. *American Economic Review* 103 (3): 226-231.

Ludwig, Jens, Jeffrey R. Kling und Sendhil Mullainathan (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives* 25 (3): 17-38.

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

McCrary, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142 (2): 698-714.

McCrary, Justin und Heather Royer (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review* 101 (1): 158-195.

Meghir, Costas und Mårten Palme (2005). Educational reform, ability, and family background. *American Economic Review* 95 (1): 414-423.

Metzler, Johannes und Ludger Wößmann (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics* 99 (2): 486-496.

Mühlenweg, Andrea M. und Patrick A. Puhani (2010). The evolution of the school-entry age effect in a school tracking system. *Journal of Human Resources* 45 (2): 407-438.

Muralidharan, Karthik und Venkatesh Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119 (1): 39-77.

Oreopoulos, Philip (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review* 96 (1): 152-175.

Parey, Matthias und Fabian Waldinger (2011). Studying abroad and the effect on international labour market mobility: Evidence from the introduction of ERASMUS. *Economic Journal* 121 (551): 194-222.

Pekkala Kerr, Sari, Tuomas Pekkarinen und Roope Uusitalo (2013). School tracking and development of cognitive skills. *Journal of Labor Economics* 31 (3): 577-602.

Pellegrini, Guido, Flavia Terribile, Ornella Tarola, Teo Muccigrosso und Federica Busillo (2013). Measuring the effects of European Regional Policy on economic growth: A regression discontinuity approach. *Papers in Regional Science* 92 (1): 217-233.

Peterson, Paul E. und William G. Howell (2004). Efficiency, bias, and classification schemes: A response to Alan B. Krueger and Pei Zhu. *American Behavioral Scientist* 47 (5): 699-717.

Peterson, Paul E., William G. Howell, Patrick J. Wolf und David E. Campbell (2003). School vouchers: Results from randomized experiments. In *The Economics of School Choice*, hrsg. von Caroline M. Hoxby. Chicago, IL: University of Chicago Press: 107-144.

Piopiunik, Marc (2013). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. Ifo Working Paper 153. Munich: Ifo Institute.

Piopiunik, Marc (2014). Intergenerational transmission of education and mediating channels: Evidence from a compulsory schooling reform in Germany. *Scandinavian Journal of Economics*: forthcoming.

Präsident des Bundesrechnungshofes, Hrsg. (2013). *Anforderungen an Wirtschaftlichkeitsuntersuchungen finanzwirksamer Maßnahmen nach § 7 Bundeshaushaltsordnung*. Schriftenreihe des Bundesbeauftragten für Wirtschaftlichkeit in der Verwaltung, Band 18. Stuttgart: W. Kohlhammer.

Reinhold, Steffen und Hendrik Jürges (2010). Secondary school fees and the causal effect of schooling on health behavior. *Health Economics* 19 (8): 994-1001.

Riphahn, Regina T. (2012). Effect of secondary school fees on educational attainment. *Scandinavian Journal of Economics* 114 (1): 148-176.

Rivkin, Steven G., Eric A. Hanushek und John F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica* 73 (2): 417-458.

Rockoff, Jonah E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94 (2): 247-252.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane und Eric S. Taylor (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review* 102 (7): 3184-3213.

Rosendahl Huber, Laura, Randolph Sloof und Mirjam Van Praag (2012). The effect of early entrepreneurship education: Evidence from a randomized field experiment. IZA Discussion Paper 6512. Bonn: Institute for the Study of Labor.

Rubin, Donald B. (1974). Estimating the causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66 (5): 688-701.

Rubin, Donald B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 2 (1): 1-26.

Sacerdote, Bruce (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116 (2): 681-704.

Schlotter, Martin (2011). The effect of preschool attendance on secondary school track choice in Germany: Evidence from siblings. ifo Working Paper 106. Munich: ifo Institute.

Schlotter, Martin, Guido Schwerdt und Ludger Wößmann (2011). Econometric methods for causal evaluation of education policies and practices: a non-technical guide. *Education Economics* 19 (2): 109-137.

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

Schmieder, Johannes F., Till von Wachter und Stefan Bender (2012). The effects of extended unemployment insurance over the business cycle: Evidence from regression discontinuity estimates over 20 years. *Quarterly Journal of Economics* 127 (2): 701-752.

Schreyögg, Jonas und Markus Grabka (2010). Copayments for ambulatory care in Germany: a natural experiment using a difference-in-difference approach. *The European Journal of Health Economics* 11 (3): 331-341.

Schwerdt, Guido, Dolores Messer, Ludger Wößmann und Stefan C. Wolter (2012). The impact of an adult education voucher program: Evidence from a randomized field experiment. *Journal of Public Economics* 96 (7-8): 569-583.

Schwerdt, Guido und Martin R. West (2013). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. PEPG Working Papers Series 12-09. Cambridge, MA: Harvard University, Program on Education Policy and Governance.

Schwerdt, Guido und Ludger Wößmann (2014). The information value of central school exams. CESifo Working Paper, forthcoming. Munich: CESifo.

Schwerdt, Guido und Amelie C. Wuppermann (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review* 30 (2): 365-379.

Stenberg, Anders, Xavier de Luna und Olle Westerlund (2012). Can adult education delay retirement from the labour market? *Journal of Population Economics* 25 (2): 677-696.

Stock, James H. und Mark W. Watson (2011). *Introduction to econometrics*. 3 ed. United States: Addison Wesley.

van den Berg, Gerard J. und Bas van der Klaauw (2006). Counseling and monitoring of unemployed workers: Theory and evidence from a controlled social experiment. *International Economic Review* 47 (3): 895-936.

West, Martin R. und Ludger Wößmann (2010). 'Every Catholic child in a Catholic school': Historical resistance to state schooling, contemporary private competition and student achievement across countries. *Economic Journal* 120 (546): F229-F255.

Winter-Ebmer, Rudolf (2003). Benefit duration and unemployment entry: A quasi-experiment in Austria. *European Economic Review* 47 (2): 259-273.

Wissenschaftlicher Beirat beim Bundesministerium für Wirtschaft und Energie (2013). *Evaluierung wirtschaftspolitischer Fördermaßnahmen als Element einer evidenzbasierten Wirtschaftspolitik*. Berlin: Bundesministerium für Wirtschaft und Energie. http://www.bmwi.de/BMWi/Redaktion/PDF/Publikationen/Studien/wissenschaftlicher-beirat-evaluierung-wirtschaftspolitischer-foerderma_C3_9Fnahmen,property=pdf,bereich=bmwi2012,sprache=de,rwb=true.pdf (13.4.2014).

Wößmann, Ludger (2005). Educational production in Europe. *Economic Policy* 20 (43): 446-504.

Wößmann, Ludger und Martin R. West (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review* 50 (3): 695-736.

Yörük, Baris K. und Ceren Ertan Yörük (2011). The impact of minimum legal drinking age laws on alcohol consumption, smoking, and marijuana use: Evidence from a regression discontinuity design using exact date of birth. *Journal of Health Economics* 30 (4): 740-752.

Zimmerman, David J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics* 85 (1): 9-23.