Köln Z Soziol (2012) 64:123–153
DOI 10.1007/s11577-012-0158-5

ABHANDLUNGEN

Kölner Zeitschrift für Soziologie und Sozialpsychologie

Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule

Joscha Legewie

Zusammenfassung: Im Vordergrund eines Großteils quantitativer Sozialforschung steht die Schätzung von kausalen Effekten. Um ein besseres Verständnis des Problems der kausalen Inferenz zu entwickeln, wird in diesem Beitrag das Kausalitätsproblem anhand einer klassischen Frage der Bildungssoziologie veranschaulicht: Dem Effekt der sozialen Zusammensetzung der Mitschüler auf die Leistungen von Schülern. Dabei werden nach einer Einführung in die Frage von Peer-Effekten in der Schule der *kontrafaktische Ansatz zur Kausalität* sowie das fundamentale Problem der Kausalanalyse anhand dieses Beispiels verdeutlicht und anschließend sowohl Experimente als auch eine Reihe von statistischen Verfahren zur Lösung des Selektionsproblems diskutiert. Im Einzelnen behandelt der Beitrag neben der Kontrolle nach Kovariaten durch die heute gängigen Regressionsmodelle, Matchingverfahren (etwa Propensity Score Matching), Fixed-Effekt und Difference-in-Difference-Modelle sowie instrumentelle Variablen und das Regression Discontinuity Design. Das Augenmerk der Einführung liegt nicht auf dem mathematischen Hintergrund oder den Schätzverfahren sondern vielmehr auf der generellen Logik der Ansätze sowie den impliziten Annahmen. Abschließend wird der adäquate Umgang mit möglichen Selektionsprozessen anhand einer beispielhaften Analyse zu Kontexteffekten in der Schule veranschaulicht.

Schlüsselwörter: Kausalanalyse · Kontexteffekte · Experimente · Regressionsanalyse · Matching · Fixed-Effekt Modelle · Difference-in-Difference Ansatz · Instrumentelle Variablen · Regression Discontinuity Design · Bildungssoziologie

The estimation of causal effects: an introduction to methods of causal inference based on peer effects in education

Abstract: This article discusses the problem of causal inference based on contextual peer effects in education. For this purpose, I first introduce the counterfactual approach to causality and the fundamental problem of causal inference. Subsequently, experiments and a number of statistical methods are discussed as possible approaches to estimate causal effects. In particular, conditioning on observables using common linear regression models as well as matching procedures

J. Legewie (✉)
Department of Sociology, Columbia University,
MC9649, 606 W 122nd Street, New York, NY 10027, USA
E-Mail: jpl2136@columbia.edu

(such as propensity score matching), fixed-effects models, the difference-in-difference approach, instrumental variables, and the regression discontinuity design are discussed. The introduction of these methods is neither focused on the mathematical background nor on the estimation procedure but rather on the general logic and the assumptions connected with a causal interpretation of the estimated effects. The paper closes with an analysis of compositional peer effects in German elementary schools to illustrate the adequate treatment of selection processes.

Keywords: Causal inference · Experiment · Regression models · Fixed-effect · Difference-in-difference · Regression discontinuity design · Sociology of education · Peer effects

1 Einleitung

Im Vordergrund eines Großteils quantitativer Sozialforschung steht, neben der deskriptiven Beschreibung von Entwicklungen und Mustern, etwa sozialer Ungleichheit, die Schätzung von kausalen Effekten. Hat etwa der Anteil von weiblichen Lehrkräften einen Effekt auf die Leistungsentwicklung von Jungen (Helbig 2010)? Oder wirkt sich die Komposition des Schul- und Klassenkontextes jenseits von individuellen Merkmalen auf die Leistung der Schüler aus (Legewie und DiPrete 2012; Schulze et al. 2009; Sacerdote 2011)? Diese und ähnliche Fragen sind in der Bildungssoziologie und anderen Bereichen entscheidend sowohl für die empirische Auseinandersetzung mit vorhandenen Theorien als auch für konkrete Handlungsempfehlungen an die Politik.

In diesem Beitrag werden das Problem der Kausalität und verschiedene Verfahren zur Schätzung von kausalen Effekten anhand des Effektes der sozialen Zusammensetzung der Schülerschaft auf die Leistungen von Schülern veranschaulicht. Seit der Veröffentlichung von Colemans Bericht 1966 (Coleman 1966) spielen sogenannte Peer-Effekte eine zentrale Rolle in der Bildungssoziologie. Gerade in Deutschland kommt der Frage aufgrund des traditionell dreigliedrigen Schulsystems eine wichtige Bedeutung zu und wurde im Hinblick auf die politische Diskussion, etwa zur Abschaffung der Hauptschule oder zur Verlängerung der Grundschulzeit, wiederholt thematisiert. Zum einen leistet die Abhandlung somit einen eigenen Beitrag zu der Diskussion über den Effekt der sozialen Zusammensetzung der Schülerschaft im deutschen Kontext. Zum anderen veranschaulicht das Anwendungsbeispiel das Problem der Kausalität sowie verschiedene Verfahren zur Schätzung von kausalen Effekten. Damit macht der Beitrag auf die einer kausalen Interpretation von Regressionskoeffizienten zugrunde liegenden Annahmen aufmerksam und treibt einen sensibleren Umgang mit den statistischen Methoden zur Schätzung von kausalen Effekten voran.[1]

1 Englischsprachige Einführungen in die hier diskutierten Methoden finden sich in Artikellänge bei Gangl (2010) und Sobel (1996, 2000) und in Buchlänge bei Morgan und Winship (2007) mit soziologischen Beispielen und bei Angrist und Pischke (2008) mit Beispielen aus den Wirtschaftswissenschaften. Eine weniger komplexe, aber sehr gute Einführung in das Thema findet sich auch bei Gelman und Hill (2007, Kap. 9 und 10). Weitere Literaturhinweise befinden sich in den Abschnitten zu den einzelnen Analyseverfahren.

Nach einer Einführung in das zentrale Anwendungsbeispiel des Beitrags werden zunächst die theoretischen Grundlagen der Kausalität sowie der *kontrafaktische Ansatz zur Kausalität* (*Counterfactual Approach to Causality*) und das fundamentale Problem der Kausalanalyse thematisiert. Anschließend werden mögliche Lösungen des Selektionsproblems diskutiert. Dabei handelt es sich im Einzelnen um Zufallsexperimente, die Kontrolle nach beobachteten Variablen durch den standard Regressionsansatz sowie Matchingverfahren, Fixed-Effekt und Difference-in-Difference Modelle und schließlich instrumentelle Variablen und das Regression-Discontinuity-Design. Der Beitrag schließt mit einer eigenen Analyse zum zentralen Anwendungsbeispiels des Beitrags: dem Einfluss des Klassenkontextes auf die schulischen Leistungen von Schülern.

2 Kompositionseffekte im Schulkontext: Ein Anwendungsbeispiel

Als Anwendungsbeispiel greift dieser Beitrag eine alte bildungssoziologische Diskussion auf, die auch in Deutschland wiederholt thematisiert wurde: Die Auswirkungen der sozialen Zusammensetzung der Schülerschaft auf die Leistungen von Schülern. Das Ziel dieses Anwendungsbeispiels ist es, sowohl einen eigenen Beitrag zu dieser wichtigen Diskussion zu leisten als auch das Problem der Kausalität sowie die verschiedenen Verfahren anhand eines konkreten Beispiels zu veranschaulichen.

Seit der Veröffentlichung von Colemans Bericht 1966 (Coleman 1966) spielt die Frage, welchen Einfluss die soziale Zusammensetzung der Schülerschaft auf Schul- oder Klassenebene auf die Leistungen von Schülern hat, eine wichtige Rolle. Coleman argumentiert in diesem Bericht, dass neben dem Familienhintergrund als wichtigster Einflussfaktor die Leistungen von Schülern auch vom sozioökonomischen Status der Mitschüler beeinflusst werden. Auch eine Reihe von deutschen Studien haben sich mit diesem Thema auseinandergesetzt. Diese Studien kommen in der Regel zu der Schlussfolgerung, dass sich die soziale Zusammensetzung der Schule auch in Deutschland auf schulische Leistungen und Bildungschancen auswirkt (siehe etwa Schulze et al. 2009; Baumert et al. 2006).

Der Einfluss der Mitschüler auf die Leistungen einzelner Schüler wird theoretisch in der Regel über drei Prozesse erklärt (siehe etwa Rumberger und Palardy 2005). Erstens beeinflussen die Mitschüler die Lernkultur in der Schule und tragen somit zur Lernmotivation aller Schüler bei. Eine höhere Leistungsmotivation ist üblicher unter Schülern mit höherem Familienstatus, sodass die Zusammensetzung der Klasse die Lernkultur in der Schule beeinflusst. Zweitens bewerten Schüler ihre eigenen Leistungen anhand von Mitschülern und verwenden diese als Referenzgruppe. Somit werden die eigenen Leistungen und Bemühungen anhand der Klassenkameraden gemessen und möglicherweise dem Klassenniveau angepasst. Drittens reagieren Lehrer nicht nur auf einzelne Schüler, sondern auch auf das Verhalten der gesamten Klasse. Somit passen sich die Lehrmethoden und der Aufbau des Unterrichts der Zusammensetzung der Klasse an.

Colemans Ergebnisse sowie darauf folgende Studien wurden auf methodischer Ebene wiederholt kritisiert. Im Vordergrund der Kritik stand der Selektionsprozess, über den sich Schüler auf Basis ihres Wohnorts, der Bemühungen ihrer Eltern sowie weiterer Merkmale in Schulen selektieren. Dieser Prozess führt dazu, dass die Merkmale von Schülern einer Schule (wie etwa die sozioökonomische Zusammensetzung der Schülerschaft) mit

unbeobachteten Merkmalen sowohl auf der individuellen- als auch auf der Schulebene korreliert sind (Sørensen und Morgan 2006, S. 153 ff.). Der Effekt der sozialen Zusammensetzung auf Schulleistungen lässt sich somit nur schwer durch einfache Kontrolle nach beobachtbaren Merkmalen auf der Schüler- und Schulebene schätzen. Eine ähnliche Kritik lässt sich in Bezug auf deutsche Studien zu dieser Fragestellung vorbringen. Diese verwenden in der Regel die heute populären Mehrebenenmodelle, welche zwar Vorteile im Vergleich zu früheren Analysemethoden bieten und zahlreiche interessante Analysemöglichkeiten eröffnen, aber nicht das grundsätzliche Problem beim Schätzen von kausalen Effekten angehen. Wie einfache Regressionsmodelle erlauben Mehrebenenmodelle lediglich, nach einer Reihe von beobachteten Variablen zu kontrollieren. Somit stellt sich auch in Bezug auf diese Studien die Frage, inwiefern mögliche Selektionsprozesse die Ergebnisse beeinflussen.

In den folgenden Abschnitten wird dieses Problem genauer thematisiert und zur Veranschaulichung sowohl der Grundlagen als auch verschiedener Verfahren der Kausalanalyse herangezogen. Schließlich wird am Ende des Beitrags eine Analyse durchgeführt, die eine explizite Auseinandersetzung mit dem Kausalitätsproblem in den Vordergrund stellt. Das Anwendungsbeispiel illustriert somit nicht nur viele, der im Verlauf des Beitrags angesprochenen Probleme und einen adäquaten Umgang mit Selektionsprozessen, sondern leistet auch einen eigenen Beitrag zu der Diskussion über den Effekt der sozialen Zusammensetzung der Schülerschaft im deutschen Kontext. Gerade aufgrund des traditionell dreigliedrigen Schulsystems und der Tatsache, dass ethnische und soziale Segregation in den USA weitaus deutlicher ausgeprägt sind, scheint die Frage wichtig zu sein, ob die für den US-amerikanischen Raum gefundenen Effekte auch in Deutschland vorhanden sind.

3 Kausalanalyse

3.1 Die Grundlagen

Sowohl in der Bildungs- als auch in anderen Bereichen der Soziologie bezieht sich ein Großteil der quantitativen Forschung auf Fragestellungen, die sich auf den Effekt bestimmter Attribute auf eine abhängige Variable beziehen. Welche Auswirkung hat etwa die soziale Zusammensetzung der Schule/Klasse auf die Leistung von Schülern? Statistisch steht damit die Frage im Vordergrund, ob X (ein bestimmtes Attribut, wie etwa die soziale Zusammensetzung der Schule/Klasse) einen kausalen Effekt auf Y (eine abhängige Variable wie etwa die Leistung von Schülern) hat und wie groß dieser Effekt ist. Im Allgemeinen werden dabei zunächst Hypothesen über Ursache-Wirkungs-Zusammenhänge auf Basis von theoretischen Argumenten oder bisherigen Forschungsergebnissen formuliert. So argumentierte Coleman und zahlreiche ihm folgenden Studien etwa, dass ein höherer sozialer Status der Mitschüler zu einem verbesserten Lernumfeld beiträgt, das sich wiederum positiv auf Schulleistungen auswirkt. Um das theoretische Argument empirisch zu belegen wird in der Regel X und Y gemessen und anschließend der Mittelwert von Y in Abhängigkeit von X – formal $E(Y|X)$ – mit Hilfe von Regressionsanalysen untersucht. Das Interesse in der Soziologie besteht zumeist darin, die Regressionsko-

effizienten nicht vorhersagend (predictive interpretation), sondern vielmehr kausal zu interpretieren (causal interpretation). Diese kausale Interpretation von Regressionskoeffizienten ist mit weitreichenden Annahmen verbunden, die häufig nicht hinreichend beachtet werden.

In der empirischen Sozialforschung hat sich eine Auseinandersetzung mit Kausalitätsfragen im Sinne des *kontrafaktische Ansatzes zur Kausalität* durchgesetzt. Diese Perspektive ist für die empirische Forschung besonders fruchtbar und wird im Folgenden dargestellt, um anschließend auf das fundamentale Problem der Kausalanalyse sowie mögliche Lösungsansätze einzugehen und diese anhand des Anwendungsbeispiels zu veranschaulichen.

3.2 Der kontrafaktische Ansatz zur Kausalität und das fundamentale Problem der kausalen Inferenz

Der kontrafaktische Ansatz zur Kausalität (counterfactual approach to causality oder auch potential outcome framework) hat sich über die letzten Jahre als allgemein anerkanntes Verständnis von Kausalität in der Forschungspraxis durchgesetzt. Im Vordergrund dieses Ansatzes steht die Annahme, dass jede Analyseeinheit mehrere potenzielle Ergebnisse hat; eins unter jedem möglichen Treatmentstatus (Morgan und Winship 2007, S. 5). Bei einem binären Treatment bedeutet das also, dass es für jede Einheit zwei potenzielle Ergebnisse gibt: y_i^0 wenn Beobachtungseinheit i nicht das Treatment erhält und somit in die Kontrollgruppe fällt und y_i^1 wenn Einheit i das Treatment erhält und somit in die Treatmentgruppe fällt. Damit lässt sich der individuelle Treatment-Effekt für Einheit i als Differenz zwischen den beiden potenziellen Ergebnissen $y_i^1 - y_i^0$ definieren. In Bezug auf unser Anwendungsbeispiel bedeutet dieses Verständnis von Kausalität etwa, dass jeder Schüler verschiedene potenzielle Schulleistungen in Abhängigkeit von der sozialen Zusammensetzung der Mitschüler aufweist. Eine zentrale Annahme dieses Ansatzes wird als *stable unit treatment value assumption* (SUTVA) bezeichnet. Diese Annahme besagt, dass die potenziellen Ergebnisse nur von dem Treatmentstatus der jeweiligen Beobachtungseinheit beeinflusst werden, nicht aber von dem Treatmentstatus der anderen Einheiten (für weitere Details siehe Morgan und Winship 2007, S. 37 ff.).

Das *fundamentale Problem der Kausalanalyse* bezieht sich auf die Tatsache, dass es sich um theoretische „was-wäre-wenn" Ergebnisse handelt, die per Definition nicht beide beobachtbar sind. Für jede Einheit lässt sich nur eines der theoretischen Ergebnisse beobachten, während es sich bei dem anderen um ein unbeobachtetes, *kontrafaktisches* Ergebnis im Sinne einer „was-wäre-wenn" Frage handelt (daher die Bezeichnung des Ansatzes). Jeder Schüler befindet sich etwa nur in einer Schule, sodass sich die Schulleistungen nur für ein bestimmtes der verschiedenen potenziellen Szenarien beobachten lassen. Damit lässt sich der individuelle Treatment-Effekt niemals direkt messen. Es lässt sich nicht feststellen, welche Schulleistungen ein bestimmter Schüler gehabt hätte, wenn er auf eine Schule mit einer anderen sozioökonomischen Zusammensetzung gegangen wäre.

Der einfachste Ansatz, dieses Problem zu umgehen, ist, das Ergebnis für *verschiedene* Einheiten mit *verschiedenem* Treatmentstatus zu vergleichen. Es werden also die Leistungen von verschiedenen Schülern in Schulen mit hohem (Treatmentgruppe) und niedrigem (Kontrollgruppe) sozialem Status der Mitschüler verglichen und auf Basis der

Leistungsdifferenz zwischen den Schulen auf den durchschnittlichen kausalen Effekt geschlossen. Dieser oft als naive Schätzer bezeichnete Ansatz beruht allerdings auf der starken Annahme, dass es keine weiteren Differenzen zwischen den beiden Gruppen gibt, die für das Ergebnis von Bedeutung sind. Entsprechende Differenzen entstehen, wenn sich bestimmte Einheiten auf Basis von Merkmalen in den Treatmentstatus selektieren, die auch mit der abhängigen Variable zusammenhängen und somit ein *Selektionsbias* entsteht. Dies ist etwa der Fall, wenn sich Schüler aufgrund von Merkmalen, wie etwa ihrem Familienhintergrund oder ihrem Wohnort, in bestimmte Schulen selektieren und diese Merkmale auch mit der Schulleistung als abhängiger Variable zusammenhängen. Damit unterscheiden sich die beiden Gruppen nicht nur in Bezug auf die sozioökonomische Zusammensetzung der Schülerschaft (dem Treatmentstatus), sondern auch in Hinblick auf andere Merkmale, die für Leistungsdifferenzen zwischen Treatment- und Kontrollgruppe verantwortlich sind. Der eigentliche Effekt der sozialen Zusammensetzung wird somit unter- oder überschätzt. Bei Beobachtungsdaten[2] ist diese Annahme in der Regel nicht vertretbar, da verschiedene Faktoren Einfluss auf den Selektionsprozess haben.

Eine elegante Lösung des Selektionsproblems auf Ebene des Forschungsdesigns stellen Zufallsexperimente dar. Alternative Ansätze basieren entweder auf natürlichen Experimenten oder verlassen sich auf statistische Anpassungen, um das Selektionsproblem zu lösen. Angefangen mit einer ausführlicheren Diskussion von Zufallsexperimenten und den gängigen Regressionsmodellen werden im Folgenden verschiedene dieser Ansätze zur Lösung des Selektionsproblems vorgestellt und anhand des bereits diskutierten Beispiels sowie weiterer Anwendungen verdeutlicht.

4 Zufallsexperimente

Zufallsexperimente werden allgemein als elegante Lösung des Selektionsproblems auf Ebene des Forschungsdesigns und somit als anzustrebender Standard angesehen. Experimentelle Daten zeichnen sich dadurch aus, dass der Treatmentstatus als Teil des Forschungsdesigns zugewiesen wird und nicht wie bei Beobachtungsdaten bereits festgelegt ist und einfach beobachtet wird. Dabei wird in der Regel der Treatmentstatus zufällig den Beobachtungseinheiten zugewiesen, sodass jede Einheit die gleiche Chance hat, Teil der Kontroll- oder der Treatmentgruppe zu sein (Komplikationen dieses Designs sind vielfältig). Die zufällige Zuordnung des Treatmentstatus löst das Selektionsproblem insofern, als die Kontroll- und Treatmentgruppe sich lediglich in Hinblick auf den Treatmentstatus unterscheiden (abgesehen von zufälligen Variationen). Der Treatmentstatus ist also exogen, da er durch äußere Ursachen bestimmt wird (die Zuweisung des Forschers). Er hängt nicht mit anderen pre-treatment Merkmalen zusammen. Damit lässt sich der durchschnittliche kausale Effekt des Treatments durch einen einfachen Mittelwertvergleich zwischen den beiden Gruppen berechnen.[3] In Bezug auf unser Anwendungsbeispiel

2 Bei Beobachtungsdaten handelt es sich um Daten, wo der Treatmentstatus beobachtet wird und nicht unter Kontrolle des Sozialforschers liegt.

3 Auch bei experimentellen Daten ist die Kontrolle nach Kovariaten mit Hilfe von Regressionen in der Regel sinnvoll (Angrist und Pischke 2008, 23): Erstens ist die Zuweisung des Treat-

würde dies bedeuten, dass Schüler zufällig Schulen zugewiesen werden und somit die soziale Zusammensetzung der Schülerschaft in keinerlei systematischen Zusammenhang mit anderen pre-treatment Merkmalen steht. Das Experiment gleicht also die Kontroll- und Treatmentgruppe bis auf Unterschiede im Treatmentstatus (und zufälliger Variationen) an. Damit lässt sich davon ausgehen, dass sich die beiden Gruppen im Hinblick auf ihre Motivation, Fähigkeiten, Familienhintergrund sowie allen anderen Merkmalen, die nicht selbst vom Treatmentstatus beeinflusst werden, ähneln.

Während Experimente in anderen wissenschaftlichen Bereichen eine entscheidende Rolle spielen, sind entsprechende Designs in den Sozialwissenschaften oft aus ethischen Gründen nicht vertretbar und zudem häufig mit hohen Kosten verbunden.[4] So ist es etwa bis auf Ausnahmen schwer vorstellbar, Schüler zufällig Schulen zuzuweisen. Dennoch wurden zahlreiche sozialwissenschaftliche Experimente durchgeführt und sogenannte „randomized field trials" spielen heute in verschiedenen Bereichen eine entscheidende Rolle.

Ein frühes Beispiel setzt sich etwa mit der Frage auseinander, welche Auswirkungen die Klassengröße auf die Lernfortschritte von Schülern hat und erlaubt es, auch Peer-Effekte zu schätzen. Im Tennessee STAR Experiment der 1980er Jahre wurden Schüler zufällig entweder Klassen mit der in Tennessee gängigen Klassengröße von etwa 22–25 Schülern (Kontrollgruppe) oder Klassen mit einer reduzierten Größe von etwa 13–17 Schülern (Treatmentgruppe) zugewiesen.[5] Boozer und Cacciola (2001) verwenden die Daten vom Tennessee STAR Experiment, um Peer-Effekte zu schätzen und zeigen, dass sich die Leistungen von Mitschülern eindeutig auf die Schulleistungen von einzelnen Schülern auswirken.

Dieses Beispiel verdeutlicht, dass experimentellen Designs auch in den Sozialwissenschaften eine große Bedeutung zukommt. Trotz der zufälligen Zuweisung des Treatments sind allerdings weitere Dinge zu beachten. So ist es durchaus möglich, dass auch die zufällige Zuweisung gerade bei kleinen Fallzahlen zu systematischen Unterschieden zwischen den beiden Gruppen führt oder andere Prozesse die zufällige Zuweisung untergraben. Dementsprechend sollte eine Überprüfung des Bias zwischen der Kontroll- und Treatmentgruppe auch bei experimentellen Forschungsdesigns Teil der Datenanalyse sein.[6]

---

mentstatus oft nicht zufällig in Bezug auf die gesamte Stichprobe, sondern nur nach Kontrolle bestimmter Variablen. Zweitens können pre-treatment Variablen die Unsicherheit (also den Standardfehler) des geschätzten kausalen Effektes verringern. Man beachte allerdings, dass eine Kontrolle nach post-treatment Variablen in der Regel nicht angebracht ist (Gelman und Hill 2007, 190 f.).

4 Dies bezieht sich insbesondere auf Feldexperimente und weniger auf Laborexperimente, wie sie etwa in der Spieltheorie üblich sind (Diekmann 2008).

5 Die Ergebnisse dieses experimentellen Designs zeigen, dass kleinere Klassen einen klaren Vorteil für die Lernentwicklung von Schülern bieten, was frühere Ergebnisse auf Basis von nicht-experimentellen Designs widerlegt (Finn und Achilles 1990).

6 Vom STAR Experiment ist beispielweise bekannt, dass Eltern Druck auf die beteiligen Schulen und Lehrer ausgeübt haben, um ihre Kinder den kleineren Klassen zuzuweisen, was vereinzelt zu Klassenwechseln während der Laufzeit der Studie geführt hat.

5 Statistische Verfahren zur Schätzung von kausalen Effekten

Wenn keine experimentellen Daten zur Verfügung stehen und somit das Selektionsproblem nicht durch das Forschungsdesign alleine gelöst werden kann, ist der empirische Sozialforscher auf kompliziertere statistische Verfahren angewiesen. Dies ist in den Sozialwissenschaften der Regelfall und trifft auch meist auf das hier diskutierte Anwendungsbeispiel zu. Die statistischen Verfahren verfolgen das Ziel, durch statistische Anpassung zufällige Variationen im Treatmentstatus zu identifizieren, um auf Basis dieser Variationen den kausalen Effekt zu schätzen. In der Literatur findet sich eine große Anzahl von Studien, die verschiedene dieser Verfahren verwenden, um Peer-Effekte zu schätzen. Der üblichste Ansatz zur Lösung des Selektionsproblems, wenn keine experimentellen Daten zur Verfügung stehen, ist die Betrachtung des Treatment-Effektes nach der Kontrolle beobachtbarer Variablen. Dabei besteht die Hoffnung darin, dass die Zuweisung des Treatmentstatus nach Kontrolle der Variablen so gut wie zufällig ist.

5.1 Standard Regressionsansatz

Eine mögliche Methode, nach beobachtbaren Variablen zu kontrollieren, sind Regressionen in der allgemein bekannten Form

$$y_i = \alpha + \theta T_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

wobei es sich bei T um die Treatmentvariable von Interesse handelt und demensprechend bei θ um den kausalen Effekt von T auf y. x_1, x_2 bis x_k bezeichnen Kontrollvariablen und β die entsprechenden Koeffizienten, die in der Regel auch unter den hier diskutierten Annahmen nicht kausal interpretiert werden können.[7] Unzählige Studien haben mit dieser Methode die Auswirkungen der sozialen Zusammensetzung des Schulkontextes untersucht (zum Beispiel Rumberger und Palardy 2005; oder auch eine Reihe von deutschen Studien wie etwa Schulze et al. 2009; Baumert et al. 2006).

Die entscheidende Frage ist allerdings, unter welchen Umständen eine kausale Interpretation des Koeffizienten θ als Treatment-Effekt angebracht ist. Im Allgemeinen ist dies der Fall, wenn die Annahme der konditionalen Unabhängigkeit (conditional independence assumption, CIA) erfüllt ist. Dies bedeutet, dass nach Kontrolle der anderen unabhängigen Variablen die Verteilung der Einheiten über die Treatment- und Kontrollgruppe in Hinblick auf die abhängige Variable so gut wie zufällig ist (Gelman und Hill 2007, S. 182 f.). Die Annahme ist nur erfüllt, wenn nach allen pre-treatment Variablen kontrolliert wird, die sowohl mit dem Treatmentstatus als auch mit der abhängigen Variablen in Verbindung stehen.[8]

7 Die diskutierten Methoden und Annahmen beziehen sich nicht nur auf lineare Regressionsmodelle mit kontinuierlichen abhängigen Variablen, sondern auch allgemein auf Generalisierte Lineare Modelle (etwa logistische Regressionen), Quantilregressionen und andere Verfahren.

8 Alternativ wird diese Annahme auch als ignorability of the treatment assignment oder selection on observables bezeichnet. Ersteres bezieht sich darauf, dass der Selektionsprozess nach Kontrolle der weiteren unabhängigen Variablen ignoriert werden kann, da kein systematischer Selektionsbias mehr in der Zuweisung des Treatmentstatus vorliegt. Die zweite Bezeichnung

In Bezug auf unser Anwendungsbeispiel bedeutet diese Annahme, dass eine kausale Interpretation der Koeffizienten nur möglich ist, wenn nach allen Variablen kontrolliert wird, die sowohl mit der Selektion von Schülern in bestimmte Schulen als auch mit der Schulleistung zusammenhängen. Nur unter dieser Bedingung ist die Verteilung von Schülern über Schulen mit unterschiedlicher sozialer Zusammensetzung (dem Treatmentstatus) nach Kontrolle der entsprechenden Variablen zufällig. Diese Annahme wurde gerade in Bezug auf unser Anwendungsbeispiel aber auch in Bezug auf andere Bereiche wiederholt als unplausibel bezeichnet. Das grundsätzliche Problem besteht darin, dass zahlreiche Variablen nicht beobachtet oder ungenau gemessen werden. So bezieht etwa keine der mir bekannten Studien die Kriterien der Schulauswahl, die Motivation der Eltern oder die Fachinteressen der Schüler als möglichen Einflussfaktor auf sowohl die Schulauswahl als auch auf die Leistungen eines Schülers ein. Die einfache Kontrolle nach dem Bildungshintergrund der Eltern ist sicherlich kein vollständiger Ersatz.

Eine entscheidende Rolle für die Glaubwürdigkeit der Ergebnisse auf Basis von Regressionsmodellen spielt daher die Auswahl der Kontrollvariablen. So haben Vergleiche von experimentellen und nicht-experimentellen Methoden gezeigt (Shadish et al. 2008; für ähnliche Ergebnisse in Bezug auf Matchingverfahren siehe Smith und Todd 2005), dass diese Auswahl entscheidenden Einfluss auf die Höhe des Bias der geschätzten Effekte hat. Grundsätzlich lässt sich sagen, dass alle pre-treatment-Variablen verwendet werden sollten, die sowohl die abhängige Variable als auch den Treatmentstatus beeinflussen. Bei der Auswahl dieser Variablen sollte auf vorhandene Kenntnisse über den Selektionsprozess zurückgegriffen werden. Ein geringer Bias ist nur dann zu erwarten, wenn dem Sozialforscher reichhaltige pre-treatment-Variablen zur Verfügung stehen, die weit über standarddemographische Merkmale hinausgehen, unmittelbare Relevanz für den Selektionsprozess haben und präzise gemessen werden. Die gängige Praxis, lediglich nach einer Reihe von standarddemographischen Merkmalen zu kontrollieren, führt hingegen zu eindeutigen Verzerrungen (Shadish et al. 2008). Ein weiterer, verbreiteter Fehler besteht darin, nach Variablen zu kontrollieren, die kausal dem Treatment nicht eindeutig vorgeordnet sind und somit möglicherweise selbst vom Treatment beeinflusst werden (Angrist und Pischke 2008, S. 64 f.). Dies kann sowohl zu einer Über- als auch Unterschätzung der entscheidenden Koeffizienten führen.

In Bezug auf den Effekt der sozialen Zusammensetzung der Schülerschaft sind somit nicht nur demographische Merkmale wie etwa Geschlecht, Familienhintergrund, Alter und Migrationshintergrund, sondern auch andere Merkmale relevant. Als Beispiel lassen sich etwa die Bildungsaspiration der Eltern, das Fachinteresse des Kindes, die Kriterien für die Schulauswahl (Distanz zum Wohnort, Fächerangebote und so weiter) oder kognitive Fähigkeiten und Schulleistungen vor der Einschulung in die entsprechende Schule anführen. Gerade bei Merkmalen wie der Bildungsaspiration der Eltern und den Fachinteressen des Kindes ist allerdings zu beachten, dass diese möglicherweise selbst von der sozialen Zusammensetzung direkt oder indirekt beeinflusst werden. Sie sind daher nur als Kontrollvariablen geeignet, wenn sie zeitlich vor der abhängigen Variable gemessen

hingegen verweist auf die Implikation der Annahme, dass die Selektion in die Treatment- und Kontrollgruppe nur auf beobachtbaren Variablen beruht und nicht auf weiteren Variablen, die unbeobachtet sind.

wurden. Querschnittsdaten erlauben es in vielen Fällen nicht, diese Kriterien zu erfüllen, da sich in der Regel nur schwer sagen lässt, ob bestimmte Variablen selbst vom Treatmentstatus beeinflusst wurden.

Diese Ausführungen verdeutlichen, wie wichtig es ist, sich zum einen, der mit einer kausalen Interpretation der Koeffizienten verbundenen Annahmen bewusst zu sein, und zum anderen, sich explizit mit dem Selektionsprozess auseinanderzusetzen. Dieser Prozess verschafft Klarheit darüber, nach welchen Variablen kontrolliert werden sollte und ermöglicht es, die Plausibilität der Annahme einzuschätzen. Die Ausführungen machen allerdings auch die grundsätzliche Schwäche des Ansatzes deutlich, insofern als keinerlei Möglichkeit besteht, nach unbeobachteten Variablen zu kontrollieren. Somit ist die Schätzung von kausalen Effekten durch die Kontrolle nach beobachtbaren Variablen grundsätzlich mit Zweifeln behaftet. Nichtsdestotrotz bietet der Regressionsansatz zahlreiche Möglichkeiten, solange die zentralen Annahmen sowie der Selektionsprozess und die gezielte Auswahl von Kontrollvariablen explizit thematisiert werden.

5.2 Matching und Propensity Score Matching

Matching bezieht sich auf eine Reihe von Verfahren, die genau wie Standard-Regressionsmodelle zur Schätzung von kausalen Effekten nach einer Reihe von möglichen konfundierenden Variablen kontrollieren. Sie folgen damit dem gleichen Grundprinzip und beruhen auf der gleichen Grundannahme (CIA). In einem ersten Schritt werden möglichst ähnliche Beobachtungseinheiten in der Kontroll- und Treatmentgruppe auf Basis der relevanten pre-treatment-Variablen gematched. In einem zweiten Schritt werden dann die so gepaarten Einheiten zur Schätzung des kausalen Effektes verwendet. Die grundlegende Idee ist damit denkbar einfach: Da sich die Beobachtungseinheiten bei nicht-experimentellen Daten in der Regel nicht nur durch den Treatmentstatus unterscheiden, sondern auch in anderer Hinsicht, werden zur Schätzung des kausalen Effektes nur Einheiten mit unterschiedlichen Treatmentstatus verglichen, die sich im Hinblick auf relevante pre-treatment Merkmale möglichst ähnlich sind. Damit stehen Matchingverfahren keineswegs im Gegensatz zu Regressionsmodellen, sondern dienen vielmehr dazu, die Daten vor der eigentlichen Analyse so aufzubereiten, dass nur möglichst ähnliche Einheiten verglichen werden. Die eigentliche Analyse wird im Anschluss unter anderem mit den gleichen Regressionsmodellen durchgeführt, die auch ohne vorheriges Matching zur Anwendung gekommen wären. In Anhang A wird das Vorgehen bei der Umsetzung von Matchingverfahren genauer diskutiert.

Crosnoe (2009) verwendet Matchingverfahren, um den Effekt der sozialen Zusammensetzung der Schülerschaft auf verschiedene abhängige Variablen zu schätzen. Dafür gruppiert er Schulen zunächst in die Kontroll- und Treatmentgruppe mit niedrigen oder hohem sozioökonomischen Status und verwendet dann propensity-score-Matching und Robustheitsanalysen zur Schätzung des kausalen Effektes. Dabei bezieht er auch eine Reihe von Variablen ein, die sich als pre-Treatment-Messung der abhängigen Variablen verstehen lassen, was sich in verschiedenen Studien als besonders wichtig erwiesen hat (Steiner et al. 2011).

Wie bereits angemerkt, beruht Matching in den verschiedenen Formen auf dem gleichen Prinzip wie der Regressionsansatz. In beiden Fällen werden Unterschiede in der

abhängigen Variablen zwischen der Kontroll- und Treatmentgruppe nach Kontrolle von konfundierenden Variablen zur Schätzung des kausalen Effektes verwendet. Dementsprechend basiert eine kausale Interpretation dieser Unterschiede auf der gleichen Annahme (CIA) und ist ebenfalls mit dem grundsätzlichen Problem behaftet, dass Selektion möglicherweise auf unbeobachtbaren Variablen beruht. Somit lässt sich in Bezug auf Crosnoes (2009) Anwendung von Matchingverfahren kritisch anmerken, dass auch hier zahlreiche potenziell relevante Variablen unbeobachtet sind und daher auch nicht in das Matchingverfahrens einbezogen werden.

Im Gegensatz zum Regressionsansatz bieten Matchingverfahren allerdings auch zahlreiche Vorteile: Erstens handelt es sich bei Matchingverfahren, zumindest theoretisch, nicht um eine parametrische Methode zur Kontrolle nach Kovariaten, sondern um ein nicht-parametrisches Verfahren. Dies bedeutet, dass im Falle von Regressionen Annahmen über die Art und Weise des Zusammenhangs zwischen den Kontrollvariablen und der abhängigen Variable gemacht werden müssen. Der Anwender ist etwa gezwungen zu spezifizieren, ob dieser Zusammenhang linear oder quadratisch ist. Exakte Matchingverfahren sind hingegen nicht-parametrisch und beruhen somit nicht auf diesen Annahmen, was zu einer Verringerung des Bias zwischen der Treatment- und Kontrollgruppe führen kann. Dies bietet zumindest theoretisch einen Vorteil im Hinblick auf die Vergleichbarkeit der Treatment- und Kontrollgruppe. In der Praxis sind diese Unterschiede allerdings in der Regel marginal (Shadish et al. 2008). Zweitens werden durch Matchingverfahren nur Beobachtungen zur Schätzung des Effektes verwendet, die auch zwischen der Treatment- und Kontrollgruppe vergleichbar sind (overlap, common support). Beobachtungen in der Kontrollgruppe, für die keine vergleichbaren Einheiten in der Treatmentgruppe vorhanden sind, werden in der Regel von der Analyse ausgeschlossen oder geringer gewichtet. Regressionen beziehen hingegen alle Beobachtungen in die Analyse ein, was dazu führen kann, dass der geschätzte Effekt stark von den Annahmen des Models abhängt. Drittens erleichtern Matchingverfahren die Anwendung von Sensitivitätsanalysen im Hinblick auf den möglichen Einfluss von unbeobachtbaren Variablen (DiPrete und Gangl 2004; Gangl und DiPrete 2004). Schließlich bieten Matchingverfahren den vielleicht entscheidenden Vorteil, dass der Anwender durch die direkte Modellierung des Selektionsprozesses dazu gezwungen wird, sich explizit mit dem Selektionsprozess auseinanderzusetzen und den Bias zwischen den beiden Gruppen zu untersuchen. Die Anwendung von Matchingverfahren an sich bedeutet somit in der Regel, dass der Frage des Selektionsbias eine wichtige Rolle zugesprochen wird, was in der Praxis bei der Verwendung von Regressionsmodellen nur selten der Fall ist. Außerdem sind Matchingverfahren in der Literatur eng mit dem kontrafaktischen Ansatz zur Kausalität verbunden, sodass im Allgemeinen eine Auseinandersetzung mit Matchingverfahren die Sensibilität für Probleme der Kausalanalyse erhöht.

Wie bereits angemerkt, sollten die Vorteile von Matchingverfahren allerdings nicht darüber hinwegtäuschen, dass auch hier keinerlei Möglichkeit besteht, das Problem von unbeobachteten Variablen zu umgehen. Entscheidend für die Glaubwürdigkeit der geschätzten Effekte ist genau wie bei Regressionsmodellen die Auswahl der Kovariaten (siehe ausführliche Diskussion im letzten Abschnitt).

5.3 Fixed-Effekt (FE)-Modelle

In vielen Situationen ist die zentrale Annahme beim Schätzen von kausalen Effekten durch die Kontrolle nach konfundierenden Variablen nicht plausibel. So lässt sich oft nicht ausschließen, dass Selektion auch auf unbeobachteten Variablen beruht. Diese Kritik wurde wiederholt auch in Bezug auf das hier behandelten Anwendungsbeispiel vorgebracht (Sørensen und Morgan 2006, S. 155 f.) und lässt sich auch durch Matchingverfahren nicht umgehen. Eine Möglichkeit mit diesem Problem umzugehen sind Fixed-Effekt (FE)-Modelle, die mehrere Beobachtungen innerhalb von Gruppen oder Individuen verwenden, um nach unbeobachteten Merkmalen auf der Gruppen oder Individualebene zu kontrollieren (Allison 1994, 2009; Halaby 2004). Dabei werden nur Einheiten innerhalb von Gruppen/Individuen verglichen, sodass alle beobachteten und unbeobachteten Merkmale auf der Gruppen- oder Individualebene konstant gehalten werden. FE-Modelle kommen oft bei Paneldaten mit mehreren Beobachtungen für jeden Befragten über die Zeit zum Einsatz, aber lassen sich auch in anderen Situationen verwenden. Ein Beispiel ist die am Ende dieses Beitrags diskutierte Analyse zum Effekt der sozialen Zusammensetzung des Schulkontextes. Diese Analyse beruht auf dem Argument, dass bei der Zuweisung von Schülern zu Klassen *innerhalb* von Schulen weitaus weniger Selektionsprozesse auftreten als bei der Selektion von Schülern in verschiedene Schulen (siehe auch Legewie und DiPrete 2012; Ammermueller und Pischke 2009). Fixed-Effekt-Modelle erlauben es nach allen (beobachteten und unbeobachteten) schulspezifischen Merkmalen zu kontrollieren und somit den Effekt nur auf Basis der Variationen in der Komposition von Klassen *innerhalb* von Schulen zu schätzen.

Ein weiteres Anwendungsbeispiel ist die Untersuchung von Budig und England (2001) zu den Auswirkungen von Mutterschaft auf den Stundenlohn mit Paneldaten und FE-Modellen, die nach allen zeitkonstanten Merkmalen auf der Ebene der Mutter kontrollieren. Daher werden nur die Änderung des Stundenlohns von einer bestimmten Frau vor und nach der Geburt, aber nicht die Unterschiede zwischen zwei unterschiedlichen Frauen mit und ohne Kind zur Schätzung des kausalen Effektes verwendet. Es wird somit nur ein bestimmter Teil der vorhandenen Variation verwendet und Befragte, die innerhalb des Befragungszeitraums kein Kind bekommen, werden von der Analyse ausgeschlossen. Aus diesem Grund werden FE-Modelle oft als konservativ bezeichnet. Dabei wird allerdings nicht beachtet, dass die Variationen zwischen Gruppen oder Individuen nicht einfach verschwendet, sondern vielmehr sinnvoll genutzt werden, um auch nach unbeobachteten Variablen auf der Gruppenebene zu kontrollieren. Im Gegensatz zur Schätzung von kausalen Effekten durch die Kontrolle nach beobachtbaren Variablen lässt sich somit in vielen Situationen argumentieren, dass die Annahme der konditionalen Unabhängigkeit plausibel ist (siehe etwa die Analyse am Ende dieses Beitrags). FE-Modelle sind damit den im letzten Abschnitt diskutierten Methoden zur Schätzung von kausalen Effekten, oft eindeutig überlegen: „It is hard to overstate the gain in identifying power provided by the beautifully simple method of FE estimation over standard cross-sectional estimators“ which is reflected in the „value of repeated observations in safeguarding causal inferences against bias arising from the presence of unmeasured cofounders“ (Gangl 2010, S. 34; siehe auch Halaby 2004). Nichtsdestotrotz gilt weiterhin die Annahme, dass, bedingt nach unbeobachteten und beobachteten gruppenspezifischen Merkmalen sowie

den weiteren Kontrollvariablen, die Verteilung der Einheiten über die Treatment- und Kontrollgruppe zufällig ist und somit das Treatment unabhängig von den anderen Beobachtungen der Gruppe/des Individuums ist. In Bezug auf das hier diskutierte Beispiel ist diese Annahme etwa verletzt, wenn sich Schüler auf Basis ihrer Leistungen, ihrer ethnischen Herkunft oder durch Bemühungen ihrer Eltern in bestimmte Klassen selektieren. Ein weiteres Problem tritt auf, wenn sich bestimmte Personen nach Kontrolle der verwendeten Kovariaten auf Basis ihrer eigenen Erwartungen über den Effekt des Treatments in die Treatmentgruppe selektieren (Gangl 2010, S. 37). Bei Budig und Englands Analyse könnte dieses Problem etwa auftreten, wenn Elternschaft in Abhängigkeit von den erwarteten Einkommensverlusten der Mutter geplant wird. Außerdem sollte beachtet werden, dass sich der geschätzte Effekt lediglich auf Gruppen oder Individuen bezieht, die Varianz im Treatmentstatus zwischen den Beobachtungen innerhalb der jeweiligen Gruppe oder des Individuums aufweisen. Bei Budig und England (2001) bezieht sich der Effekt etwa nur auf Frauen, die zwischen den Beobachtungszeitpunkten ein Kind bekommen haben, was die externe Validität der geschätzten Effekte in Frage stellt.[9] Ein weiterer Nachteil von FE-Modellen besteht darin, dass sich nicht der Effekt von gruppenspezifischen Merkmalen schätzen lässt. Dies ist allerdings nur ein Problem, wenn sich das Forschungsinteresse auf einen gruppenspezifischen Effekt bezieht, der Konstant über die verschiedenen Beobachtungen ist.

5.4 Der Difference-in-Difference-Ansatz (DD)

Beim Difference-in-Difference-Ansatz handelt es sich um eine spezielle Art von FE-Modellen, bei der in der Regel der kausale Effekt einer Intervention, wie etwa einer Gesetzesänderung oder eines Ereignisses, durch den Vergleich des Trends von Gruppen mit und ohne Intervention geschätzt wird (Meyer 1995). Dabei wird der Trend innerhalb der Gruppe mit Intervention mit dem Trend innerhalb der Gruppe ohne Intervention verglichen, sodass die Gruppe ohne Intervention als kontrafaktischer Trend für die Gruppe mit der Intervention verwendet wird. Dadurch lässt sich sowohl nach unbeobachteten Gruppenmerkmalen als auch nach Merkmalen des Zeitpunkts der Intervention kontrollieren, die den beiden Gruppen gemein sind. Als Bespiel lässt sich hier etwa die allgemeine wirtschaftliche Entwicklung in Deutschland anführen, wenn es sich bei den Gruppen um verschiedene Bundesländer handelt.

Dies lässt sich am einfachsten grafisch und anhand eines Beispiels verdeutlichen.[10] Abbildung 1 veranschaulicht das Prinzip des DD-Ansatzes, wobei die Intervention zu den Zeitpunkten 6 und 7 auftritt. Wie an der grauen Linie verdeutlicht, wird der Trend der Kontrollgruppe als kontrafaktischer Trend für die Treatmentgruppe verwendet. Der Treatment Effekt ergibt sich aus dem Unterschied zwischen dem eigentlichen und dem kontrafaktischen Trend der Treatmentgruppe. Die Berechnung des Treatmenteffektes basiert auf

9 Externe Validität bezieht sich auf die Frage, zu welchem Grad sich die Ergebnisse auf die Grandgesamtheit sowie zeitlich und geografisch andere Umstände übertragen lassen.

10 Da sich Difference-in-Difference Modelle nur schwer auf das hier diskutierte Anwendungsbeispiel beziehen lassen, werden in diesem Abschnitt andere Anwendungsbeispiele zur Verdeutlichung herangezogen.

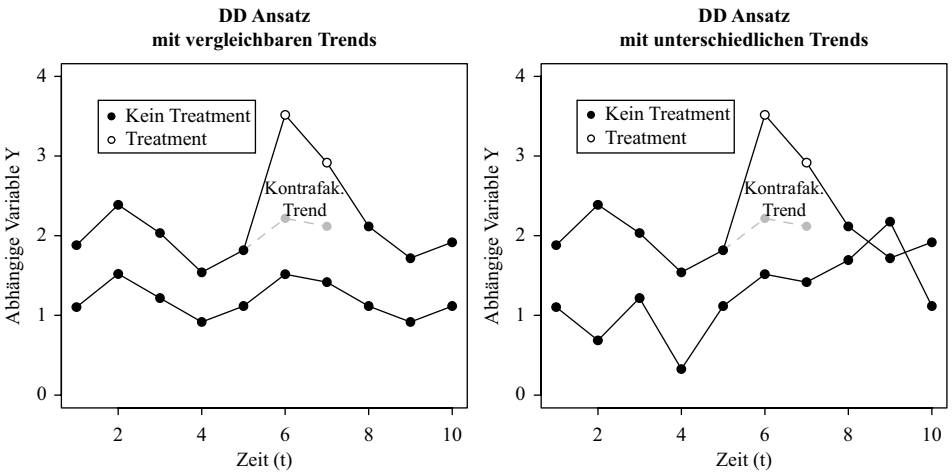


Abb. 1: Veranschaulichung des DD Ansatzes

der impliziten Annahme, dass der Trend ohne die Intervention oder das Ereignis in beiden Gruppen der Gleiche gewesen wäre. Die Plausibilität dieser Annahme lässt sich durch weitere Beobachtungen vor und nach der Intervention überprüfen. Abbildung 1 links zeigt einen Fall, in dem diese Annahme plausibel erscheint, da sich die Trends der beiden Gruppen in der Zeit vor und nach der Intervention sehr ähnlich sind. Abbildung 1 rechts hingegen zeigt einen Fall, wo sich die Trends deutlich unterscheiden und die Annahme somit unplausibel erscheint.

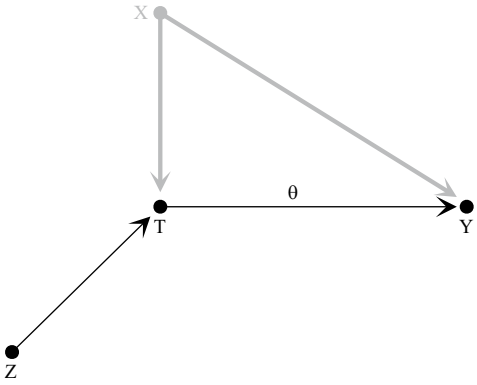
Pischke (2007) verwendet den DD-Ansatz, um die Auswirkungen der Länge des Schulsemesters auf die Leistungen von Schülern zu untersuchen. Dabei greift er auf eine Gesetzesänderung Ende der 1960er in Deutschland zurück, die während der Umstellungsphase in einigen Regionen zu einer Verkürzung des Semesters von 37 auf 24 Wochen geführt hat. Er vergleicht den Trend des Anteils der Schüler, die eine Klasse wiederholen müssen in den Regionen, die von dem verkürzten Semester betroffen sind, mit dem Trend in Regionen, die von der Gesetzesänderung nicht betroffen sind. Seine Ergebnisse zeigen, dass das verkürzte Schulsemester zu einem deutlich erhöhten Anteil von Sitzenbleibern geführt hat.[11]

5.5 Instrumentelle Variablen (IV)

Ein weiteres in der Ökonometrie verbreitetes Verfahren zur Schätzung von kausalen Effekten sind instrumentelle Variablen, die es unter bestimmten Umständen ermöglichen, das Problem der Selektion auf Basis von unbeobachteten Variablen zu umgehen. Ein Instrument Z_i ist eine Variable, die mit dem Treatmentstatus T_i korreliert ist, aber nicht auf anderem Wege (direkt oder indirekt) mit der abhängigen Variable Y_i zusammenhängt.

11 Ein weiteres Beispiel ist die berühmte Studie von Card und Krueger (1994; siehe auch 2000 für eine spätere Replikationen mit anderen Ergebnissen) zu den Auswirkungen von Mindestlöhnen auf die Arbeitsmarktbeteiligung.

Abb. 2: Veranschaulichung von IV-Ansatz



Dieses Instrument kann dazu verwendet werden, den kausalen Effekt des Treatments zu schätzen, indem nur der Teil der Variation in T_i verwendet wird, der sich durch das Instrument Z_i erklären lässt. Wenn die Annahmen in Bezug auf das Instrument erfüllt sind (siehe weiter unten), ist dieser Teil der Variation im Treatmentstatus nicht mit Selektionsproblemen behaftet.

Dieser Ansatz wird in Abb. 2 verdeutlicht. Wir interessieren uns für den Effekt eines Treatments T, wie etwa der sozialen Zusammensetzung einer Schule, auf eine abhängige Variable Y, wie etwa die Lernfortschritte von Schülern. Eine einfache Regression von Y auf T führt allerdings zu einem Fehler in der Schätzung von θ, da ein Selektionsbias durch X vorhanden ist (X steht an dieser Stelle für jeden erdenklichen Selektionsbias). Eltern mit hoher Bildungsaspiration für ihre Kinder investieren etwa mehr Zeit in die Auswahl einer geeigneten Schule und unterstützen gleichzeitig ihre Kinder intensiv bei Hausaufgaben. Damit beeinflusst X, die Bildungsaspiration der Eltern, sowohl die soziale Zusammensetzung der Schule (Treatmentstatus T) als auch die schulischen Leistungen von Schülern (abhängige Variable Y) und führt somit zu einem Bias in der Schätzung des Treatmenteffektes. Durch eine einfache Kontrolle nach X ließe sich das Problem des Selektionsbias lösen. Diese Möglichkeit besteht allerdings nicht, wenn X unbeobachtet ist und möglicherweise einen unbekannten Bias darstellt. In diesem Fall bietet das Instrument Z eine Alternative, wenn folgende Annahmen erfüllt sind: a) Es besteht ein Zusammenhang zwischen Z und T. Schwache Instrumente, also Instrumente, die nur in geringen Zusammenhang mit dem Treatmentstatus stehen, sind oft mit Problemen behaftet. Diese Annahme lässt sich leicht empirisch überprüfen. b) Das Instrument Z hängt nicht auf anderem Wege (direkt oder indirekt) mit der abhängigen Variable Y zusammen. Diese als exclusivity assumption bezeichnete Annahme ist in der Regel empirisch nicht überprüfbar und daher der kritische Punkt im Hinblick auf die Plausibilität von Instrumenten. Hinzu kommt, dass schon eine leichte Verletzung dieser Annahme bei schwachen Instrumenten zu einem großen Bias in der Schätzung des kausalen Effektes führen kann. In diesem Fall ist in der Regel eine einfache Regression zu bevorzugen.

Einige Studien verwenden instrumentelle Variablen um den Effekt der sozialen Zusammensetzung auf die Schulleistung von Schülern zu schätzen. Imberman et al. (2009) nutzen etwa die ursprüngliche Zuweisung von Schülern zu Schulen nach der erzwungenen Evakuierung bestimmter Gegenden durch Hurrikan Katrina als Instrument für die ver-

änderte Schulkomposition einige Monate später. Dieses Instrument beruht auf dem Argument, dass direkt nach der Evakuierung die Zuweisung zu Schulen weitgehend chaotisch ablief und von daher so gut wie zufällig war, wohingegen die Schüler sich in den folgenden Monaten wieder in bestimmte Schulen selektiert haben. Legewie und DiPrete (2011) verwenden hingegen Variationen in der Komposition zwischen aufeinanderfolgenden Kohorten von Schulanfängern innerhalb von Schulen als Instrument für die spätere Komposition von Kohorten. Diese Strategie beruht auf der Annahme, dass natürliche Unterschiede in der Komposition von aufeinanderfolgenden Kohorten innerhalb einer Schule auftreten, die bis auf einen generellen Trend weder Eltern noch den Schulen selbst bekannt sind.

Diese beiden Ansätze bieten unter Umständen eine Alternative zu den bisher diskutierten Verfahren und erlauben es auch, wenn eindeutige Selektionsprozesse vorliegen, den kausalen Effekt zu schätzen. Der Nutzen des Ansatzes hängt allerdings unmittelbar von der Validität des Instruments ab und kann unter Umständen vorhandene Probleme noch verschlimmern. Außerdem ist zu beachten, dass wie bei Fixed-Effekt Modellen nur ein bestimmter Teil der Varianz des Treatmentstatus zur Schätzung des kausalen Effektes verwendet wird und somit die externe Validität der Ergebnisse möglicherweise fraglich ist. Im Speziellen wird der Effekt nur auf Basis derjenigen Beobachtungen geschätzt, deren Treatmentstatus sich auch durch das Instrument beeinflussen lässt (den sogenannten *compliers*). Dieser Effekt wird in der Literatur als *Local Average Treatment Effect* (LATE) bezeichnet (Angrist und Krueger 2001, S. 77 f.). Eine wichtige Frage für die externe Validität der Ergebnisse ist, ob sich der Effekt für diese spezielle Gruppe von Beobachtungen von dem Effekt für andere Beobachtungen unterscheidet.

5.6 Regression Discontinuity Design (RD)

Das Regression Discontinuity Design (RD) bezieht sich auf Situationen, in denen der Treatmentstatus (zumindest teilweise) durch den Grenzwert einer pre-treatment-Variable bestimmt wird.[12] Als Beispiel lässt sich etwa auf Interventionen verweisen, die in Abhängigkeit von einem bestimmten Merkmal, wie etwa dem Abschneiden in einem Test oder dem Ergebnis einer medizinischen Untersuchung, zur Anwendung kommen. Ein Grenzwert bestimmt dabei den Treatmentstatus der Beobachtungen. Der Treatment-Effekt lässt sich in solchen Situationen durch einen Vergleich zwischen Beobachtungseinheiten berechnen, die gerade über oder unter dem Grenzwert liegen und sich somit im Hinblick auf den Treatmentstatus unterscheiden. Dabei werden Sprünge oder Diskontinuitäten in der Regressionsfunktion (daher der Name) beim Grenzwert verwendet, um den Treatment-Effekt zu schätzen.[13] Diese Methode beruht auf der Annahme, dass kleine Abwei-

12 Die zugänglichsten Einführungen in das Thema finden sich in den bereits genannten Überblickswerken (Gelman und Hill 2007, S. 212 ff.; Angrist und Pischke 2008, 251 ff.). Andere Einführungen sind selten und meist auf einem technisch hohem Niveau wie etwa Imbens und Lemieux (2008).

13 In der Regel unterscheidet man zwischen „sharp“ und „fuzzy“ RD-Designs. Im ersten Fall wird der Treatmentstatus komplett durch den Grenzwert determiniert und der Treatment-Effekt lässt sich durch lokale lineare Regressionsmodelle schätzen, wobei die Beobachtungen nahe dem

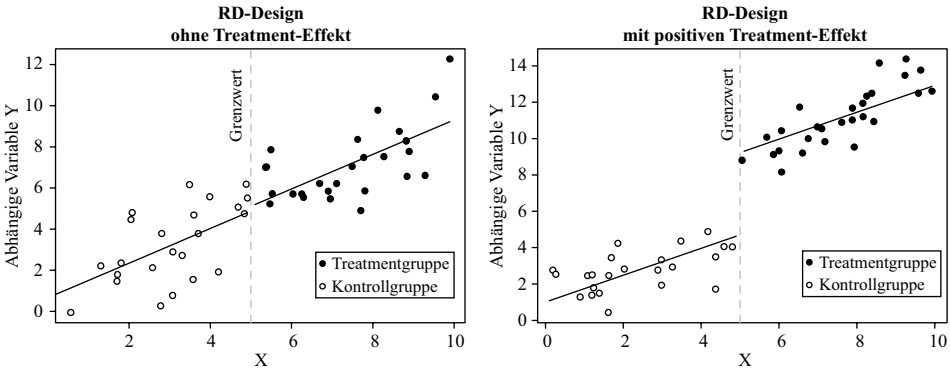


Abb. 3: Veranschaulichung vom RD-Design

chungen in der entscheidenden pre-treatment-Variable zufällig zustande kommen oder der Grenzwert zu einem gewissen Grade beliebig ist, sodass sich Einheiten nahe dem Grenzwert auch in anderer Hinsicht ähnlich sind. Eine weitere Annahme besteht darin, dass sich die Beobachtungseinheiten nicht über den Grenzwert bewusst sind und aktiv darauf Einfluss nehmen können, ob sie über oder unter den Grenzwert fallen und sich somit selbst in den Treatmentstatus selektieren. Außerdem ist zu beachten, dass genau wie bei FE-Modellen und IV der Treatment-Effekt nur auf Basis eines bestimmten Teils der Variation geschätzt wird, insofern als Beobachtungen um den Grenzwert verglichen werden. Daher stellt sich auch hier die Frage der externen Validität.

Abbildung 3 veranschaulicht die Grundidee des RD-Designs. Zunächst wird in beiden Grafiken deutlich, dass der Treatmentstatus komplett durch den Grenzwert der Variable X bestimmt wird. Variablen über dem Grenzwert fallen in die Treatmentgruppe und Variablen unter dem Grenzwert in die Kontrollgruppe. Die linke Abbildung zeigt einen Fall, in dem das Treatment keinen Effekt auf die abhängige Variable hat und es daher zu keinem Sprung in der Regressionsfunktion kommt. Die rechte Abbildung zeigt hingegen einen hypothetischen Fall, in dem wir einen klaren Treatment-Effekt beobachten, der in der Diskontinuität der Regressionsfunktion um den Grenzwert (dem Sprung) zum Ausdruck kommt. Die beiden Beispiele zeigen einfache lineare Zusammenhänge. Das RD-Design lässt sich allerdings auf andere funktionale Zusammenhänge übertragen und erlaubt es auch, unterschiedliche Zusammenhänge für die Werte über und unter dem Grenzwert zu modellieren. Außerdem lassen sich lokale Regressionsmodelle verwenden, bei denen Beobachtungen nahe dem Grenzwert höher gewichtet werden.

Pop-Eleches und Urquiola (2008) verwenden ein RD-Design um den Effekt von ‚besseren Schulen' und deren sozialen Zusammensetzung auf die Lernentwicklung von

Grenzwert höher gewichtet werden. Im zweiten Fall besteht eine Wahrscheinlichkeitsbeziehung zwischen der pre-Treatment Variable und dem Treatmentstatus, sodass die Beobachtungen über dem Grenzwert nicht zwangsläufig in die Treatmentgruppe fallen, aber eine deutlich höhere Wahrscheinlichkeit haben. Die Schätzung des Treatment-Effekt beim „fuzzy" RD-Designs greift auf ähnliche Methoden wie instrumentelle Variablen zurück, wobei das Instrument über den Grenzwert definiert wird.

Schülern zu schätzen. Dafür nutzen sie die Regelungen des Übergangs in die Oberschule im rumänischen Schulsystem. Die Zuweisung von Schülern zu Schulen innerhalb von Regionen hängt hier von einem nationalen Test in der 8. Klassenstufe, dem Notendurchschnitt sowie einer vorbestimmten Anzahl von Plätzen für jede Schule ab. Schüler, die besser abschneiden, haben die freie Auswahl, wobei diese meist auf die beste Schule der Region fällt, bis die Plätze in dieser Schule belegt sind. Dadurch ergibt sich ein eindeutiger Grenzwert, der es Schülern erlaubt, die bessere Schule zu besuchen und somit die Treatment- von der Kontrollgruppe unterscheidet. Zur Berechnung des kausalen Effektes werden Schüler miteinander verglichen, die gerade über und gerade unter dem Grenzwert liegen. Auf Basis dieses eleganten RD-Designs kommen die Autoren zu dem Ergebnis, dass bessere Schulen einen eindeutig positiven Effekt auf die Lernentwicklung der Schüler haben. Dabei ist allerdings zu beachten, dass hier spezielle Schüler verglichen werden, was möglicherweise die externe Validität der Ergebnisse in Frage stellt. Es handelt sich um die Schüler, die entweder die besten in der schlechteren Schule oder die schlechtesten in der besseren Schule sind, was die Ergebnisse für mögliche Referenzgruppenprozesse anfällig macht.

Insgesamt handelt es sich beim RD-Design um ein Design, dass nur in sehr speziellen Situationen zur Anwendung kommen kann und auf genaue Kenntnisse des Selektionsprozesses angewiesen ist. In den entsprechenden Situationen bringt ein RD-Design aber in der Regel sehr verlässliche Schätzer hervor (s. Cook und Wong 2008 für einen Vergleich von Ergebnissen auf Basis von experimentellen Daten und mit Hilfe eines RD-Designs).

6 Kompositionseffekte im Schulkontext: Eine Analyse zum Anwendungsbeispiel

Im Verlauf dieses Beitrags wurden verschiedene Studien zum Effekt der sozial Zusammensetzung auf die Schulleistungen von Schülern vorgestellt, die auf unterschiedliche Methoden zur Schätzung des kausalen Effektes zurückgreifen. Diese Methoden verfolgen im Allgemeinen das Ziel, die Probleme des Standard-Regressionsansatzes oder von Matchingmethoden zu umgehen. Dabei wird in der Regel versucht, einen bestimmten Teil der Variation in der Komposition von Schulen zu identifizieren, für den auf glaubwürdige Art und Weise gezeigt werden kann, dass er so gut wie zufällig ist. Jede dieser Analysestrategien ist genau wie die hier vorgestellte Alternative mit eigenen Annahmen verbunden, die sich nicht immer direkt überprüfen lassen. Wichtig ist vielmehr eine adäquate Auseinandersetzung mit diesen Annahmen und dem entscheidenden Selektionsprozess. Dementsprechend werden im Folgenden die Annahmen explizit thematisiert und mögliche Selektionsprozesse im Detail diskutiert. Die Analyse in Anlehnung an Legewie und DiPrete (2012) veranschaulicht somit zum einen viele der angesprochen Probleme sowie einen adäquaten Umgang mit Selektionsprozessen und leistet zum anderen einen eigenen Beitrag zu der Diskussion über den Effekt der sozialen Zusammensetzung der Schülerschaft im deutschen Kontext.

6.1 Datengrundlage und Variablen

Als Datengrundlage dient der ELEMENT-Datensatz. Die ELEMENT-Studie (Lehmann und Lenkeit 2008) ist eine Längsschnittuntersuchung, die Schüler von Berliner Grundschulen von der 4. bis zur 6. Jahrgangsstufe in den Jahren 2004, 2005 und 2006 begleitet. Die Stichprobe besteht aus 3169 Grundschülern in 71 Grundschulen und 1724 Gymnasiasten in 31 grundständigen Gymnasien. Die folgende Analyse greift nur auf die Grundschulstichprobe zurück. In jeder der Schulen wurden die Schüler in mindestens zwei Klassen befragt, sodass Schüler sowohl in Schulen als auch in Klassen eingebettet sind. Als abhängige Variablen wird die Lese- und Mathematikkompetenz der 5. Klasse verwendet. Als entscheidende unabhängige Variable wurde der durchschnittliche sozioökonomische Status auf der Schul- und Klassenebene verwendet. Die Kontrollvariablen auf individueller sowie Schul- und Klassenebene werden in Tab. 1 im Anhang B beschrieben.

6.2 Schätzen des Kausalen Effektes: Selektionsprozess und Analyseverfahren

Die folgende Analyse greift auf Variationen in der sozioökonomischen Zusammensetzung von Klassen innerhalb der gleichen Schule zurück, um das Problem der Selektion von Schülern in Schulen zu umgehen. Dieses Vorgehen basiert auf dem Argument, dass Schüler in Berliner Grundschulen so gut wie zufällig Klassen innerhalb von Schulen zugewiesen werden oder zumindest nicht auf Basis ihres sozioökonomischen Status. Sollte dies der Fall sein, ergeben sich durch diesen Zufallsprozess Variationen in der sozialen Zusammensetzung von Klassen, die nicht mit anderen Merkmalen korrelieren und somit zur Schätzung des kausalen Effektes geeignet sind. Entscheidend für dieses Argument ist der Selektionsprozess, über den sich Schüler in Klassen innerhalb einer Schule selektieren. Um sich genauer mit diesem Selektionsprozess auseinanderzusetzen, habe ich 1) die Grundschulverordnung in Berlin untersucht, 2) Simulationen verwendet und 3) neun qualitative Interviews mit Grundschulleitern in Berlin durchgeführt. Diese ausführliche Auseinandersetzung mit dem Selektionsprozess erlaubt es mir, nicht nur die Annahme der zufälligen Zuweisung zu beurteilen und abzuschätzen inwiefern ein Selektionsbias vorliegt, sondern auch gezielte Sensitivitätsanalysen durchzuführen.

6.2.1 Untersuchung des Selektionsprozesses

Die Berliner Grundschulverordnung (§ 8) betont, dass Klassen heterogen im Hinblick auf Geschlecht, Muttersprache und Leistung sein sollten. Diese Regulierungen schließen eine Zuweisung auf Basis von Leistungen aus und begrenzen den Einfluss der anderen Selektionsprozesse. Es besteht allerdings weiterhin die Möglichkeit, dass Selektionsprozesse am Werk sind, die eine quasi-zufällige Zuweisung untergraben und somit zu einem Selektionsfehler beitragen. So besagt die Grundschulverordnung etwa auch, dass alte Freundschaften bei der Zuweisung zu Klassen berücksichtigt werden können. Im Folgenden werden Simulationen verwendet, um sich mit diesem Problem auseinanderzusetzen. Die Simulationen haben zum Ziel, die eigentlichen, beobachteten Unterschiede in der sozioökonomischen Komposition von Klassen innerhalb einer Schule mit den Unterschieden

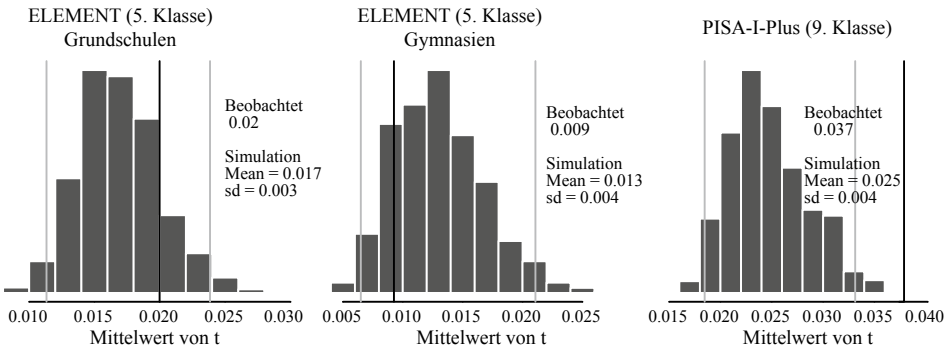


Abb. 4: Simulation eines zufälligen Selektionsprozesses

zwischen Klassen zu vergleichen, die durch einen zufälligen Prozess entstehen. Dafür wird zunächst die Variation der sozioökonomischen Zusammensetzung zwischen Klassen innerhalb einer Schule berechnet. Anschließend werden die Schüler dieser Schule zufällig über die Klassen einer Schule verteilt und die Variationen nach dieser zufälligen Zuweisung erneut berechnet.[14] Die Simulation einer zufälligen Zuweisung wird dann etliche Male wiederholt (in diesem Fall 1000 mal), um schließlich die beobachtete Variation mit der Verteilung möglicher Ergebnisse bei einer zufälligen Zuweisung zu vergleichen. Sollte ein eindeutiger Selektionsprozess am Werk sein, wäre zu erwarten, dass die beobachteten Variationen zwischen Klassen größer sind als die durch reinen Zufall entstandenen Variationen.

Abbildung 4 zeigt die Ergebnisse der Simulationen (Verteilung) zusammen mit den beobachteten Werten (schwarze vertikale Linie) für die Grundschulen und Gymnasien im ELEMENT-Datensatz sowie als Vergleich im PISA-I-Plus Datensatz (9. Klasse). Die Verteilung stellt plausibel Werte unter der Annahme eines zufälligen Selektionsprozesses von Schülern in Klassen innerhalb einer Schule dar. Sowohl für die Grundschulen als auch die Gymnasien im ELEMENT- Datensatz ist klar ersichtlich, dass der beobachtete Wert innerhalb dieser Verteilung liegt und somit konsistent mit einem zufälligen Selektionsprozess ist. Für den PISA-I-Plus-Datensatz zeigt die Abbildung hingegen, dass der beobachtete Wert sehr unwahrscheinlich unter einem zufälligen Selektionsprozess ist. Dieses Ergebnis stimmt mit den bisherigen Erwartungen überein und legt nahe, dass die Selektionsprozesse in der Oberschule aufgrund von Klassenzuweisungen auf Basis von Fremdsprachenwahl sowie der höheren Anzahl von Klassenwiederholern stärker ausgeprägt sind. Somit unterstützen die Simulationen das Argument, dass der Selektionsprozess

14 Als Statistik zum Vergleich der simulierten und beobachteten Werte wird die durchschnittliche quadrierte Abweichung der Klassenmittelwerte vom Schulmittelwert verwendet $t_j = \frac{1}{n_j}\sum_{k=1}^{n_j}(\bar{y}_{jk} - \bar{y}_j)^2$, wobei j Schulen indiziert und k Klassen. Y bezeichnet den sozioökonomischen Status oder eine andere Variable und y_j dementsprechend den Mittelwert der Schule j und y_{jk} den Mittelwert der Klasse k in Schule j. n_j bezeichnet die Anzahl von Klassen in Schule j. Wenn die Anzahl der Schüler in den Klassen gleich ist, handelt es sich um die Varianz der Klassenmittelwerte innerhalb einer Schule.

in Berliner Grundschulen so gut wie zufällig im Hinblick auf den sozioökonomischen Status der Schüler ist.

Die Ergebnisse der Simulationen sind zwar informativ, geben aber keinerlei Auskunft über den eigentlichen Ablauf des Selektionsprozesses. Somit ist es weiterhin denkbar, dass ein Selektionsprozess am Werk ist, der eine Komposition von Klassen hervorbringt, die im Ergebnis der Simulation eines zufälligen Zuweisungsprozess ähnelt. Um ein tiefgreifenderes Verständnis über den eigentlichen Selektionsprozess zu entwickeln, habe ich daher neun qualitative Telefoninterviews mit Schulleitern an Berliner Grundschulen durchgeführt. Die Schulen wurden vornehmlich zufällig ausgewählt, aber die Stichprobe wurde so ergänzt, dass Schulen aus verschiedenen Stadtteilen und mit verschiedenen ethnischer Zusammensetzung befragt wurden. Der Schwerpunkt der etwa 15–20 minütigen Interviews lag auf dem eigentlich Verfahren, über das Schüler Klassen zugewiesen werden, den Kriterien, die dabei eine Rolle spielen, der Art und Weise, wie Eltern versuchen, auf den Prozess Einfluss zu nehmen, wie mit Schülern umgegangen wird, die eine Klasse wiederholen müssen und wie Lehrer Klassen zugewiesen werden (der Interviewleitfaden ist vom Autor erhältlich).

Die Ergebnisse der Interviews zeigen, dass unterschiedliche Verfahren bei der Zuweisung von Schülern zu Klassen verwendet werden. Alle Schulleiter haben aber berichtet, dass Schüler nicht direkt auf Basis ihres Familienhintergrunds Klassen zugewiesen werden und dass Elternwünschen nur in Ausnahmefällen umgesetzt werden. Allerdings verweisen die Interviews auch auf mögliche Selektionsprozesse: Erstens haben die Schulleiter in allen Interviews betont, dass Schüler, die eine Klasse wiederholen zwar meist der kleineren Klasse zugewiesen werden, aber unter Umständen auch Überlegungen über die sozialen Dynamiken im Klassenzimmer eine Rolle spielen. Zweitens haben einige Schulen betont, dass Kinder, die den gleichen Kindergarten besucht haben oder die sich auf anderem Wege kennen, der gleichen Klassen zugewiesen werden. Drittens haben zwei Schulen berichtet, dass Schüler, die nicht über ausreichend Deutschkenntnisse verfügen, einer Klasse zugewiesen werden. Diese drei Kriterien bei der Zuweisung von Schülern zu Klassen stellen mögliche Selektionsprozesse dar, die zu Problemen bei der Schätzung der relevanten Effekte führen könnten. Allerdings ist davon auszugehen, dass diese Kriterien nur eine untergeordnete Rolle spielen. Zum einen sprechen die Ergebnisse der Simulationen dagegen, dass starke Selektionsprozesse am Werk sind und zum anderen haben eine Reihe von Schulleitern auch betont, dass Zufall eine wichtige Rolle bei der Zuweisung von Schülern zu Klassen spielt.

Abschließend lässt sich auf Basis dieser ausführlichen Auseinandersetzung mit dem Selektionsprozess sagen, dass die Zuweisung von Schülern zu Klassen zwar keinem klaren Zufallsprozess folgt, aber dennoch davon auszugehen ist, dass weitaus weniger Selektionsprozesse am Werke sind als bei der üblichen Analyse auf der Schulebene. Außerdem lässt sich dieses Wissen über den Selektionsprozess dazu verwenden, zielgerichtete Sensitivitätsanalysen durchzuführen, die es uns erlauben zu untersuchen, inwiefern die Ergebnisse von den drei möglichen Selektionsprozessen beeinflusst werden.

6.2.2 Analyseverfahren

Unter der Annahme, dass Schüler so gut wie zufällig Klassen zugewiesen werden, lässt sich der kausale Effekt der sozioökonomischen Zusammensetzung der Schülerschaft auf Klassenebene durch Schul-Fixed-Effekt Modelle berechnen:

$$y_{iks} = \alpha_s + \theta(SES\ Komp)_k + X_i\beta + U_k\lambda + \varepsilon_{iks} \quad (1)$$

wobei es sich bei i, k und s um die Indizes für Schüler, Klassen und Schulen handelt. Der entscheidende Koeffizient θ schätzt den kausalen Effekt der sozio-ökonomischen Zusammensetzung auf der Klassenebene. Die Matrix X bezieht sich auf Kontrollvariablen auf der Individualebene und U auf Kontrollvariablen auf der Klassenebene. Sie verbessern die Balance zwischen der Kontroll- and Treatmentgruppe weiter; die entsprechenden Koeffizienten in den Vektoren β und λ lassen sich allerdings nicht kausal interpretieren. Desweitern werden Instrumentelle Variablen-FE Modelle verwendet, um sich genauer mit den möglichen Selektionsprozessen auseinanderzusetzen (Details weiter unten). Die Ergebnisse der Schul-FE-Modelle werden mit Schätzern sowohl auf Basis von OLS Regressionen als auch auf Basis von Mehrebenenmodellen mit 3 Ebenen (Schüler, Klasse und Schule) verglichen.

6.3 Analyseergebnisse

Abbildung 5 zeigt den Effekt der soziökonomischen Zusammensetzung der Schule/Klasse auf das Lese- (5a) sowie Mathematikverständnis (5b) für 7 verschiedene Modellspezifikationen zusammen mit 95 % Konfidenzintervallen. Der Effekt wird in Standardabweichungen (SD) dargestellt. Dementsprechend bedeutet ein Wert von 0,2, dass das Leseverständnis um 0,2 SD steigt, wenn sich der durchschnittliche soziökonomische Status in der Schule/Klasse um eine SD erhöht.

Die Abbildungen zeigen, dass der Effekt der soziökonomischen Zusammensetzung durchweg positiv und signifikant ist. Allerdings zeigen sich auch starke Variationen in der Größe des Effektes zwischen den verschiedenen Modellspezifikationen. In den OLS und MLM Modellen 1 bis 3 liegt der Effekt bei ungefähr 0,3 SD. Dabei wird in allen drei Modellen eine Bandbreite von Kontrollvariablen sowohl auf der individuellen Ebene als auch auf der Schul/Klassen-Ebene verwendet (s. Tab. 1). Modell 4 zeigt schließlich den Effekt der soziökonomischen Zusammensetzung auf Klassenebene des Schul-FE-Modells, sodass nach allen beobachteten und unbeobachteten Schulmerkmalen sowie zusätzlichen Kontrollvariablen auf individueller und Klassenebene kontrolliert wird. Im Vergleich zu den ersten drei Modellen ist der Effekt auf das Leseverständnis von Schülern von 0,20 SD etwa 33 % kleiner und damit von der Größe her vergleichbar mit den Ergebnissen einiger der weiter oben diskutierten US-amerikanischen Studien.

Bei den drei verbleibenden Modellen handelt es sich schließlich um Sensitivitätsanalysen, die sich mit der Frage auseinandersetzen, inwieweit die durch die qualitativen Interviews dokumentierten Selektionsprozesse das Ergebnis von Modell 4 beeinflussen. In Modell 5 und 6 wurde jeweils eine instrumentelle Variable verwendet, die über den durchschnittlichen sozioökonomischen Status einer Untergruppe von Schülern definiert ist. In Modell 5 handelt es sich um alle Schüler, die keine Klasse wiederholt haben. Damit

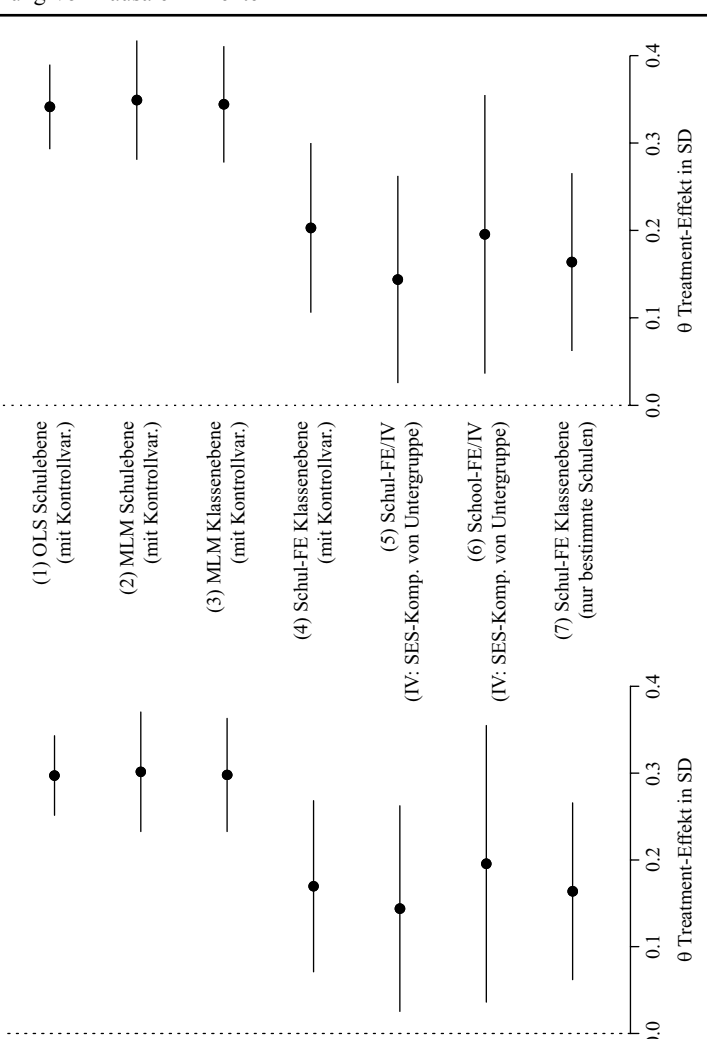


Abb. 5: Effekt der Sozioökonomischen Zusammensetzung der Schule/Klasse

bezieht sich der geschätzte Effekt in Modell 5 nur auf den Anteil der Variation in der sozioökonomischen Zusammensetzung, der auf Schüler zurückzuführen ist, die keine Klasse wiederholt haben. Dieser Teil der Variation ist eindeutig nicht von dem möglichen Selektionsprozess betroffen über den Schüler, die eine Klasse wiederholen müssen, bestimmten Klassen zugewiesen werden. Das Ergebnis dieses IV-FE Schätzers unterstützt die bisherigen Ergebnisse und deutet darauf hin, dass der Selektionsprozess keinen Einfluss auf das Ergebnis hat.

In Modell 6 basiert die Definition des Instruments hingegen auf den Schülern, die keine KITA (Kindertagesstätte) besucht haben, eine Klasse übersprungen oder von einer anderen Schule gewechselt haben. Diese Untergruppe von Schülern ist nicht von dem möglichen Selektionsprozess betroffen, dass manche Schulen Schüler von einer KITA der gleichen Klassen zuweisen. Wie in Modell 5 bezieht sich der geschätzte Effekt auf den Anteil der Varianz in der Zusammensetzung der 5 Klasse, der sich durch den durchschnittlichen sozioökonomische Status dieser Untergruppe von Schülern erklären lässt. Das Ergebnis zeigt, dass auch dieser Prozess den geschätzten Effekt nicht zu beeinflussen scheint.

Schließlich setzt sich das 7. Modell mit Selektion auf Basis von Sprachfähigkeit und ethnischer Herkunft auseinander. Es handelt sich um den gleichen FE-Schätzer wie in Modell 4, jedoch mit dem Unterschied, dass nur eine Untergruppe von Schulen zur Schätzung des Effektes verwendet wurde. In dem Modell wurden Schulen ausgeschlossen, in welchem die Zusammensetzung der Klassen darauf hindeutet, dass Schüler auf Basis der ethnischen Herkunft Klassen zugewiesen werden.[15] Das Ergebnis auf Basis dieses FE-Modells zeigt, dass auch dieser mögliche Selektionsprozess keinen Einfluss auf das Ergebnis zu haben scheint.

7 Schlussfolgerung

Der vorliegende Beitrag hat sich mit dem Problem der Kausalität und verschiedenen Verfahren zur Schätzung von kausalen Effekten anhand der Auswirkungen der sozialen Zusammensetzung der Mitschüler auf die Leistungen von Schülern auseinandergesetzt und dabei eine kritische Perspektive auf die übliche Praxis geworfen. Inhaltlich zeigen die Ergebnisse, dass auch in Deutschland die soziale Zusammensetzung der Schule/Klasse einen Einfluss auf die Leistungen von Schülern zu nehmen scheint. Die Größe der geschätzten Effekte variiert allerdings stark mit dem verwendeten Analyseverfahren, was darauf hinweist, dass die einfach Kontrolle nach beobachteten Variablen zu einer deutlichen Überschätzung der Effekte führt. Dieses Ergebnis ist gerade im Hinblick auf die Diskussion zum dreigliedrigen Schulsystem relevant. Es legt nahe, dass eine frühe Verteilung von Schülern auf verschiedene Schularten zur weiteren Benachteiligung von Schülern aus prekären Familienverhältnissen beiträgt. Des Weiteren veranschaulicht das

15 Um diese Schulen zu identifizieren, wurde ein einfacher z-Test verwendet, der den Anteil der Schüler mit Migrationshintergrund zwischen den Klassen vergleicht. Alle Schulen mit einem p-Wert von unter 0,1 – also einem signifikanten Unterschied im Anteil der Schüler mit Migrationshintergrund zwischen den Klassen – wurden von der Analyse ausgeschlossen.

Beispiel eine ausführliche Auseinandersetzung mit dem Selektionsprozess anhand von einer Untersuchung der offiziellen Regulierungen, einer Simulation sowie qualitativer Interviews mit den entscheidenden Akteuren. Diese Auseinandersetzung trägt zu einem allgemeinen Verständnis über die Plausibilität der Annahmen bei und führt somit zu mehr Sicherheit über die geschätzten Effekte. Sie erlaubt es auch, gezielte Sensitivitätsanalysen durchzuführen, die es ermöglichen zu untersuchen, inwiefern die Ergebnisse durch mögliche Selektionsprozesse beeinflusst werden. Die gewonnene Sicherheit ist sowohl für den innerwissenschaftlichen Dialog als auch für die politischen Implikationen auf Basis der Ergebnisse entscheidend.

Neben diesen inhaltlichen Ergebnissen hat der Beitrag in verschiedene Methoden der kausalen Inferenz eingeführt und eine kritische Perspektive auf die aktuelle Forschungspraxis geworfen. Zur Einführung in die Kausalanalyse wurde zunächst das Konzept der kontrafaktischen Kausalität sowie das fundamentale Problem der Kausalanalyse vorgestellt. Anschließend wurden Experimente, der in der Praxis dominierende Regressionsansatz sowie eine Reihe von alternativen Verfahren und deren implizite Annahmen diskutiert. Standard-Regressionsmodelle erlauben es nach konfundierenden Variablen zu kontrollieren, um somit eine Vergleichbarkeit der Kontroll- und Treatmentgruppe zu erreichen. Dabei beruht eine kausale Interpretation der Koeffizienten auf der Annahme, dass die Verteilung der Einheiten über die Treatment- und Kontrollgruppe im Hinblick auf die abhängige Variable nach der Kontrolle von X so gut wie zufällig ist (conditional independence assumption oder selection on observables). Matchingverfahren lassen sich als wichtige Ergänzung dieses Ansatzes verstehen, deren Vorteil insbesondere darin besteht, dass der Anwender zu einer expliziten Auseinandersetzung mit dem Selektionsprozess angeregt wird und die Evaluation des Bias zwischen der Treatment und Kontrollgruppe ein wichtiger Bestandteil des Vorgehens ist. Die alternativen Verfahren gehen über die einfache Kontrolle nach konfundierenden Variablen hinaus und verwenden zur Schätzung des kausalen Effektes nur einen bestimmten Anteil der Variation im Treatmentstatus. Die Grundidee dieser Ansätze besteht darin, einen bestimmten Teil der Gesamtvariation im Treatmentstatus zu identifizieren und möglichst überzeugend zu zeigen, dass dieser Anteil der Variation so gut wie zufällig ist. Das hier diskutierte Anwendungsbeispiel basiert etwa auf dem Argument, dass die Zusammensetzung von Klassen innerhalb von Schulen so gut wie zufällig ist und greift auf die offiziellen Regulierungen, statistische Simulationen und qualitative Interviews zurück, um dies überzeugend darzulegen.

Die diskutierten alternativen Verfahren erweitern den Fundus von statistischen Methoden zur Schätzung von kausalen Effekten. Die Ergebnisse auf Basis dieser Verfahren sind unter Umständen weitaus plausibler als die einfache Kontrolle nach beobachteten Variablen mit Hilfe von Regressionen oder Matchingverfahren. Entscheidend ist allerdings, unabhängig vom verwendeten Verfahren, die explizite Auseinandersetzung mit dem Selektionsprozess und den Annahmen, die mit einer kausalen Interpretation der Ergebnisse einhergehen.

Der Fokus des Beitrags lag auf dem Schätzen von kausalen Effekten, also den empirischen Aspekten der Auseinandersetzung mit Kausalität. Ebenso wichtig sind die theoretischen Aspekte, um die beobachteten Effekte durch zugrunde liegende Mechanismen oder Prozesse zu erklären. Aktuelle Beiträge, die betonen, wie wichtig diese Mechanismen

sind und auch dazu aufrufen, diese empirisch zu untersuchen, finden sich bei Hedström (2005) und Hedström und Bearman (2009).

Danksagung: Ich danke Tom DiPrete, Martin Ehlert, Anette Fasang, Nicolas Legewie und Merlin Schaeffer für hilfreiche Kommentare.

Anhang A – Analyseverfahren

In diesem Anhang werden weitere Details zu den verschiedenen Analyseverfahren vorgestellt. In Bezug auf Matchingverfahren wird das genaue Vorgehen in vier Schritten ausführlicher beschrieben und bei den anderen Verfahren erste, sehr verkürzte Hinweise auf die statistische Umsetzung der Verfahren gegeben.

Matchingverfahren

Stuart (2010) beschreibt das Vorgehen bei der Umsetzung von Matchingverfahren anschaulich in vier Schritten (weitere Details finden sich auch bei Rosenbaum 2009 oder Morgan und Harding 2006). Demnach wird im ersten Schritt zunächst ein Unterschiedsmaß definiert, auf deren Basis die Beobachtungseinheiten verglichen werden. Unterscheiden lässt sich grundsätzlich zwischen einer exakten Maßzahl auf der einen Seite und eindimensionalen Unterschiedsmaßen auf der anderen Seite. Beim exakten Matching werden nur Einheiten gepaart, die in allen verwendeten Kovariaten übereinstimmen. Gerade bei einer hohen Anzahl von Kovariaten und kontinuierlichen Variablen ist dieser Ansatz allerdings aufgrund zu kleiner Zellgrößen in der Regel nicht durchführbar. Als Alternative werden verschiedene eindimensionale Unterschiedsmaße verwendet. Darunter fällt auch der prominente Ansatz des propensity score (p-score) matching. Unter der wahren propensity score versteht man die (unbekannte) Wahrscheinlichkeit, dass eine bestimmte Einheit das Treatment erhält. In der Regel ist die wahre propensity score allerdings unbekannt, sodass die geschätzte Wahrscheinlichkeit des Treatments bei bestimmten Kovariaten verwendet wird. Diese lässt sich unter anderem durch eine direkte Modellierung des Selektionsprozesses mit Hilfe von logistischen Regressionen oder ähnlichen Verfahren als $P(T_i = 1|X)$ berechnen. Entscheidend im ersten Schritt ist auch die Frage, welche Variablen in die Berechnung der Distanz zwischen Beobachtungen einbezogen werden. Grundsätzlich gelten die gleichen, bereits diskutierten Kriterien wie auch bei Regressionsmodellen. Da in diesem Schritt allerdings das vornehmliche Ziel in der Vorhersage der propensity scores besteht, lassen sich weitgehend problemlos eine große Anzahl von Variablen verwenden, solange diese Variablen nicht selbst vom Treatment beeinflusst werden (pre-treatment-Variablen).

Im zweiten Schritt wird das eigentlich Matching auf Basis der zuvor definierten Distanz durchgeführt, wobei unterschiedliche Methoden wie etwa „nearest neighbor matching“ oder „full matching“ in der Literatur zur Anwendung kommen. Diese Methoden unterscheiden sich im Wesentlich dadurch, wie viele Einheiten nach dem Matching erhal-

ten bleiben und welches Gewicht den verschiedenen Einheiten zugesprochen wird. Beim einfachsten Verfahren, dem 1:1 „nearest neighbor matching“, wird etwa jede Einheit aus der Treatmentgruppe mit der Einheit aus der Kontrollgruppe gepaart, die ihr im Hinblick auf die zuvor definierte Distanz am ähnlichsten ist.[16] Weitere Methoden werden etwa von Stuart (2010, S. 7–10) diskutiert.

Im dritten Schritt wird die Qualität der gematchten Stichprobe durch eine Evaluation des Bias zwischen der Treatment- und Kontrollgruppe analysiert. Eigentliches Ziel ist es, dass die multivariate Verteilung aller beobachten Kovariaten zwischen den beiden Gruppen gleich ist. Ein Vergleich von multivariaten Verteilungen ist allerdings mit Schwierigkeiten verbunden. Daher werden in der Praxis normalerweise Statistiken auf Basis von einzelnen Variablen verwendet, wie etwa die standardisierte Differenz zwischen den Mittelwerten. Sollte die gematchte Stichprobe die Unterschiede zwischen den Gruppen im Vergleich zu der Ausgangsstichprobe nicht eindeutig reduzieren oder weiterhin einen klaren Bias aufweisen, wird der erste bis dritte Schritt wiederholt, solange bis die gematchte Stichprobe möglichst kleine Unterschiede zwischen den beiden Gruppen aufweist.

Schließlich wird im vierten Schritt zum ersten Mal die abhängige Variable einbezogen und der kausale Effekt auf Basis der gematchten Stichprobe geschätzt. Ho et al. (2007) empfehlen, das gleiche Regressionsmodell zu verwenden, dass auch ohne Matching zur Anwendung gekommen wäre. Abschließend sollten auch Sensitivitätsanalysen durchgeführt werden, wie in der Literatur zu Matchingverfahren besprochen (Gangl und DiPrete 2004).

Fixed-Effekt-Modelle

Statistisch lassen sich FE-Modelle durch einfache Regressionsmodelle mit einer zusätzlichen Dummy-Variable für jede Gruppe/jedes Individuum schätzen, was identisch mit einer getrennten Konstante für jede Gruppe ist:

$$y_{ij} = \alpha_j + \theta T_{ij} + X_{ij}\beta + \varepsilon_{ij}$$

wobei sich j auf die Gruppe und i auf die Beobachtung innerhalb der Gruppe bezieht. Der Index j für die Konstante α bedeutet, dass eine gruppenspezifische Konstante geschätzt wird. Diese Spezifikation ist allerdings bei einer hohen Anzahl von Gruppen mit Problemen verbunden, da für jede Gruppe ein getrennter Koeffizient geschätzt werden muss. Dementsprechend werden FE-Modelle üblicherweise nach der Subtraktion der gruppenspezifischen Mittelwerte geschätzt. Dies führt zu identischen Koeffizienten, aber macht es unnötig, getrennte gruppenspezifische Konstanten zu schätzen:

$$(y_{ij} - \bar{y}_j) = \theta(T_{ij} - \bar{T}_j) + (X_{ij} - \bar{X}_j)\beta + (\varepsilon_{ij} - \bar{\varepsilon}_j)$$

16 Alternativ lassen sich auch jeder Einheit in der Treatmentgruppe mehrere Einheiten aus der Kontrollgruppe zuweisen (k:1 nearest neighbor matching).

Difference-in-Difference-Ansatz

Statistisch lässt sich der kausale Effekt in DD-Modellen durch einfache Regressionen mit einer gruppenspezifischen Konstante wie bei FE Modellen sowie einem Interaktionsterm berechnen:

$$y_{ijt} = \alpha_j \gamma_t + \theta T_{ijt} + \varepsilon_{ijt}$$

wobei α_j den zeitkonstanten, gruppenspezifischen Effekt darstellt, γ_t den zeitspezifischen Effekt, der den Gruppen gemein ist und T_{ijt} einen Interaktionsterm der Zeitpunkte und Gruppen, wo die Intervention aufgetreten ist (in Abb. 1 die Beobachtungen der oberen Trendlinie zu den Zeitpunkten 6 und 7). Dementsprechend stellt θ den Treatmenteffekt dar. Diese Modelle lassen sich einfach erweitern, etwa durch Kontrollvariablen in der Form von $X_{ijt}\beta$ (Meyer 1995).

IV-Modelle

Statistisch lassen sich IV-Modelle durch two-stage least square (2SLS) Verfahren berechnen. Dabei wird in einem ersten Schritt der Effekt des Instruments Z auf den Treatmentstatus T geschätzt (Formel 2) und in einem zweiten Schritt der Effekt des aus dem ersten Schritt vorhergesagten Treatmentstatus T auf die abhängige Variable Y (Formel 3). Diese first und second stage Regressionen lassen sich darstellen als

$$\hat{T}_i = \alpha_1 + \delta Z_i + X_i \beta_1 \tag{2}$$

$$y_i = \alpha_2 + \theta \hat{T}_t + X_i \beta_2 + \varepsilon_i \tag{3}$$

wobei es sich bei $\hat{T}_i$ um den endogenen Treatmentstatus handelt, bei Z_i um ein exogenes Instrument für den Treatmentstatus und bei X_i um zusätzliche exogene Kovariaten nach denen kontrolliert wird. Angrist und Pischke (2008, 121 ff.) diskutieren das 2SLS Verfahren im Detail sowie alternative Schätzverfahren. Eine gute Einführung findet sich auch bei Angrist und Krueger (2001).

Anhang B – Kontrollvariablen

Tab. 1: Beschreibung der Kontrollvariablen

Unabhängige Variable	
Individuelle Ebene	
Alter	kontinuierlich, in Monaten
Geschlecht	0 – männlich, 1 – weiblich
Familienhintergrund	ISEI Skala
Migrationshintergrund	Kategorische Variable:
	1 – beide Eltern in Deutschland geboren
	2 – ein Elternteil in Deutschland geboren
	3 – kein Elternteil aber Kind in Deutschland geborgen
	4 – Kind im Ausland geboren
Klasse wiederholt	0 – keine Klasse wiederholt, 1 – Klasse wiederholt
Schul-/Klassenebene	
Größe der Klasse/Schule	Anzahl von Schülern in Schule/Klasse
Anteil von Jungen	Anteil von Jungen auf Schul-/Klassenebene

Literatur

Allison, Paul D. 1994. Using panel data to estimate the effects of events. *Sociological Methods & Research* 23:174–199.

Allison, Paul D. 2009. *Fixed effects regression models*. Sage.

Ammermueller, Andreas, und Jörn-Steffen Pischke. 2009. Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics* 27:315–348.

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

Angrist, Joshua D., und Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Baumert, Jürgen, Petra Stanat und Rainer Watermann. 2006. Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000*, Hrsg. Jürgen Baumert, Petra Stanat und Rainer Watermann, 99–185. Wiesbaden: VS Verlag für Sozialwissenschaften.

Boozer, Michael A., und Stephen E. Cacciola. 2001. *Inside the „Black Box" of Project STAR: Estimation of peer effects using experimental data*. New Haven: Yale University.

Budig, Michelle J., und Paula England. 2001. The wage penalty for motherhood. *American Sociological Review* 66:204–225.

Card, David, und Alan B. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review* 84:772–793.

Card, David, und Alan B. Krueger. 2000. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. *American Economic Review* 90:1397–1420.

Coleman, James. 1966. *Equality of educational opportunity*. Washington: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Cook, Thomas D., und Vivian C. Wong. 2008. Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique* 91:127–150.

Crosnoe, Robert. 2009. Low-income students and the socioeconomic composition of public high schools. *American Sociological Review* 74:709–730.

Diekmann, Andreas. 2008. Soziologie und Ökonomie: Der Beitrag experimenteller Wirtschaftsforschung zur Sozialtheorie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 60:528–550.

DiPrete, Thomas A., und Markus Gangl. 2004. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology* 34:271–310.

Finn, Jeremy D., und Charles M. Achilles. 1990. Answers and questions about class size: A statewide experiment. *American Educational Research Journal* 27:557–577.

Gangl, Markus. 2010. Causal inference in sociological research. *Annual Review of Sociology* 36:21–47.

Gangl, Markus, und Thomas A. DiPrete. 2004. Kausalanalyse durch Matchingverfahren. In *Kölner Zeitschrift für Soziologie und Sozialpsychologie: Methoden der Sozialforschung*, Hrsg. Andreas Diekmann, 396–420. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gelman, Andrew, und Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Halaby, Charles N. 2004. Panel models in sociological research: Theory into practice. *Annual Review of Sociology* 30:507–544.

Hedström, Peter. 2005. *Dissecting the social: On the principles of analytical sociology*. Cambridge: Cambridge University Press.

Hedström, Peter, und Peter Bearman. 2009. *The Oxford handbook of analytical sociology*. Oxford: Oxford University Press.

Helbig, Marcel. 2010. Sind Lehrerinnen für den geringeren Schulerfolg von Jungen verantwortlich? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62:93–111.

Ho, Daniel E., Kosuke Imai, Gary King und Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

Imbens, Guido, und Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142:615–635.

Imberman, Scott, Adriana Kugler und Bruce Sacerdote. 2009. *Katrina's children: A natural experiment in peer effects from hurricane evacuees.* NBER Working Paper No. 15291.

Legewie, Joscha, und Thomas A. DiPrete. 2011. *Gender differences in the effect of peer SES: Evidence from a second quasi-experimental case study.* Working paper, Columbia University.

Legewie, Joscha, und Thomas A. DiPrete. 2012. School context and the gender gap in educational achievement. *American Sociological Review* 77:3.

Lehmann, Rainer, und Jenny Lenkeit. 2008. *ELEMENT. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien*. Berlin: Humboldt Universität zu Berlin.

Meyer, Bruce D. 1995. Natural und quasi-experiments in economics. *Journal of Business & Economic Statistics* 13:151–161.

Morgan, Stephen L., und David J. Harding. 2006. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods Research* 35:3–60.

Morgan, Stephen L., und Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

Pischke, Jörn-Steffen. 2007. The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *The Economic Journal* 117:1216–1242.

Pop-Eleches, Cristian, und Miguel Urquiola. 2008. The consequences of going to a better school. Columbia University.

Rosenbaum, Paul R. 2009. *Design of observational studies*. New York: Springer.

Rumberger, Russell W., und Gregory J. Palardy. 2005. Does segregation still matter? The impact of student composition on academic achievement in High School. *Teachers College Record* 107:1999–2045.

Sacerdote, Bruce. 2011. Peer effects in education: How might they work, how big are they, and how much do we know thus far? In *Handbook of Economics of Education*, vol. 3, Hrsg. Eric A. Hanushek, Stephen Machin und Ludger Woessmann, 249–277. Amsterdam: North-Holland.

Schulze, Alexander, Felix Wolter und Rainer Unger. 2009. Bildungschancen von Grundschülern: Die Bedeutung des Klassen- und Schulkontextes am Übergang auf die Sekundarstufe I. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 61:411–435.

Shadish, William R., M. H. Clark und Peter M. Steiner. 2008. Can nonrandomized experiments yield accurate Answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association* 103:1334–1344.

Smith, Jeffrey A., und Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125:305–353.

Sobel, Michael E. 1996. An introduction to causal inference. *Sociological Methods Research* 24:353–379.

Sobel, Michael E. 2000. Causal inference in the social sciences. *Journal of the American Statistical Association* 95:647–651.

Sørensen, Aage B., und Stephen L. Morgan. 2006. School effects: Theoretical and methodological issues. In *Handbook of the Sociology of Education*, Hrsg. Maureen T. Hallinan, 137–160. New York: Springer.

Steiner, Peter M., Thomas D. Cook und William R. Shadish. 2011. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics* 36:213–236.

Stuart, Elizabeth A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25:1–21.

Joscha Legewie, 1983, PhD Student der Columbia University in New York City. Forschungsinteressen: Bildungssoziologie, soziale Ungleichheit, Peer-Effekte sowie statistische Methoden insbesondere zur Schätzung von kausalen Effekten. Momentanes Projekt: Die Rolle des Schulkontextes für Geschlechtsungleichheiten im Bildungssystem (American Sociological Review 77 (2012)) sowie das Projekt zu den Auswirkungen von Terroranschlägen auf die Wahrnehmung von Ausländern.