

## 6 Methodisch-konzeptionelle Meta-Analyse der Studien

Es gibt, wie weiter oben gezeigt, eine Vielzahl von Untersuchungen, die sich mit dem Thema Gebietsreformen auseinandersetzen. In den Studien kommen zahlreiche unterschiedliche wissenschaftliche Forschungsansätze und Methoden zur Anwendung. Diese zu analysieren ist der Zweck des vorliegenden Abschnitts. Zunächst werden dazu die verschiedenen Ansätze systematisiert, um anschließend die wichtigsten Herausforderungen und Problemstellungen zu reflektieren. Es werden jeweils die allgemeinen methodischen Problemstellungen benannt und dann mit ausgewählten Beispielen belegt. Dadurch wird nachvollziehbar dargestellt, mit welchen Problemen Studien zu Gebietsreformen regelmäßig konfrontiert sind.

Wie für sozialwissenschaftliche Forschung üblich und nicht anders zu erwarten, ist keiner der bisher angewandten Ansätze frei von Limitationen, weshalb die konstruktiv-kritische Auseinandersetzung mit den Studien besonders wichtig ist. Darin liegt keine Schwäche, sondern eine dezidierte Stärke empirischer Verwaltungsforschung. Sie versteht es, komplexe gesellschaftliche Phänomene mit unterschiedlichen Methoden zu untersuchen und darüber zu reflektieren. Der Abschnitt wird sich daher nicht mit der Frage befassen, welche Befunde möglicherweise „richtig“ oder „falsch“, oder schlimmer, welche „gut“ oder „schlecht“ sind.

Dieses Kapitel gliedert sich wie folgt: Zu Beginn wird auf unterschiedliche Untersuchungsdesigns eingegangen, die im Bereich der Forschung zu Gebietsreformen vorliegen. In diesem Zusammenhang werden die verschiedenen Zwecke und Fallstricke der Forschungsansätze kurz besprochen. Daran anschließend geht es um das Problem von Korrelation (d.h. statistisch gemessenen Zusammenhängen) und Kausalität (d.h. existierenden Ursache-Wirkungsbeziehungen). Es folgen Erwägungen zur Datenverfügbarkeit und zur Generalisierbarkeit der in den Studien hervorgebrachten Befunde.

### 6.1 Untersuchungsdesigns

Die Untersuchungen zu Gebietsreformen lassen sich, ähnlich wie in der Evaluationsforschung (vgl. Wollmann 2002), in drei Gruppen untergliedern. So gibt es *ex post*-, begleitende (*ongoing*) und *ex ante*-Untersuchungen. Studien werden also vor, während oder nach der Umsetzung von Gebietsreformen durchgeführt. Dies sagt allerdings wenig über die zur Anwendung kommenden Forschungs-

methoden aus. Sie umfassen die in den jeweiligen Studien angewandten Erhebungs- und Auswertungsinstrumente. Hier kann wiederum zwischen qualitativen, d.h. die Subjektivität und Komplexität zulassende, und quantitativen, d.h. auf Zahlen basierende, Verfahren unterschieden werden (Pierce 2008). Da beide Verfahren mit spezifischen Herausforderungen behaftet sind, hat es sich bewährt, beide Ansätze miteinander zu verbinden und so ihre gegenseitigen Schwächen auszubalancieren. Dabei handelt es sich um sogenannte gemischte Designs bzw. *mixed methods*-Ansätze (Tashakkori/Teddlie 2003; Kelle 2006).

Darüber hinaus spielen die klassischen Perspektiven von Mikro-, Meso- und Makroebene eine wichtige Rolle. Während die *Mikroebene* das Individuum und dessen Handlungen und Wahrnehmungen untersucht (beispielsweise Effekte von Reformen auf Verwaltungspersonal, Bürger, Mandatsträger etc.; vgl. Ems/Greiner 2017, Büchner/Franzke 2001 a, 2001 b, Hesse 2008, Steiner/Kaiser 2016), befasst sich die *Mesoebene*, d.h. die Mittelebene, mit den Effekten von Gebietsreformen auf Organisationen und Organisationseinheiten (vgl. Hesse 2000, Büchner/Franzke 2001 a, b). Das höchste Abstraktionsniveau haben Untersuchungen der *Makroebene*, die Gesellschaften und ihre Teilsysteme untersuchen (Donges 2011, 217 ff.) etwa bei ländervergleichenden Studien unter Berücksichtigung von verschiedenen Aggregatdaten (vgl. Schaap/Karsten 2017; Blume/Blume 2007, Hanes 2015; Rösel 2016 a, b; Blom-Hansen et al. 2016, 2014).

All diese Untersuchungen haben ihre Vor- und Nachteile (siehe unten). Als besondere Herausforderung für jegliche Forschungsarbeit kann der Mikro-Meso-Makro-Link bezeichnet werden (Donges 2011; Schimank 2008). Hierbei werden alle drei Ebenen miteinander verbunden. In der Regel ist ein solches Forschungsvorhaben sehr aufwendig und lässt sich kaum im Rahmen eines Buchkapitels oder Forschungsartikels abhandeln. Daher sind umfassende Gutachten und Expertisen (meist als ex ante-Untersuchungen) wichtig, da sich diese umfassender mit der Problematik von Gebietsreformen auseinandersetzen können (bspw. Bogumil 2016, Ragnitz et al. 2010, Seitz 2005, 2007 a, b, u. a. Hesse 2007, 2008).

Gerade in wissenschaftlichen Artikeln sind Forscher dazu angehalten, Schwerpunkte zu setzen. Sie können sich daher oft nur mit einer bestimmten Ebene befassen und sind daher gezwungen, andere Sachverhalte und Zusammenhänge auszublenden. Grundsätzlich ist das nicht problematisch, da empirische Sozialforschung immer einen begrenzten Rahmen für die Gültigkeit der Befunde benötigt. Es wird aber dann zum Problem, wenn deterministische Ansprüche erhoben und allgemeine normative und präskriptive Schlussfolgerungen gezogen werden.

## 6.2 Korrelation und Kausalität

Viele Studien untersuchen vergleichsweise komplexe Sachverhalte. Hinzu kommt, dass Gebietsreformen nicht im luftleeren Raum geschehen. Sie finden vor dem Hintergrund bestimmter Ausgangs- und Rahmenbedingungen statt. Dazu gehören etwa die jeweilige gesamtwirtschaftliche Entwicklung, strukturelle Spezifika, soziale Lage, politische Situation und kulturelle Faktoren (etc.). Diese lassen sich ausdifferenzieren auf die jeweiligen Mikro-, Meso- oder Makroebenen. Eine solch komplexe Gemengelage macht es folglich sehr schwierig, den Untersuchungsgegenstand umfassend zu erörtern, was dementsprechend praktisch nur ausschnittsweise gelingen kann. Damit stellen sich gleichzeitig besondere Herausforderungen an derartige Untersuchungen.

Eine der zentralen Herausforderungen ist die Unterscheidung zwischen Korrelation und Kausalität. Korrelation meint einen statistischen Zusammenhang, der über verschiedene Maßzahlen nachgewiesen werden kann. Kausalität meint dagegen einen faktischen Zusammenhang, der sich an drei Kriterien erkennen lässt. Diese sind (nach Benninghaus 2002: 252):

1. Das Bestehen einer statistischen Beziehung (*association*) zwischen zwei Variablen;
2. das zeitliche Vorgehen einer der beiden Variablen (*causal order* oder *temporal precedence*);
3. das Verbleiben des Zusammenhangs auch bei der Kontrolle von Drittfaktoren (*lack of spuriousness*).

Es liegt letztlich in der Verantwortung des einzelnen Wissenschaftlers, Kausalitäten theoretisch herzuleiten und in den Studien nachzuweisen.

Hierin unterscheiden sich beispielsweise die oben genannten *ex ante*-, die begleitenden und die *ex post*-Studien. Zwar stellen die meisten Studien basierend auf bestehender Forschung Kausalannahmen auf und postulieren Wirkungsbeziehungen, trotzdem ist es nicht die Aufgabe der *ex ante*-Studien, diese Annahmen vorab zu testen. Sie lassen sich bestenfalls argumentativ plausibilisieren. Die begleitende Forschung kann dagegen einen ersten Einblick geben, inwieweit die zu erwartenden Folgen tatsächlich eintreten und die *ex post*-Forschung kann im Nachhinein bestimmte vorab formulierte Kausalannahmen überprüfen.

Aus diesen unterschiedlichen Vorgehensweisen in den Studien lässt sich jedoch keine unterschiedliche Wertigkeit der Untersuchungen ableiten. Zwar mag der erste Eindruck unterstellen, dass *ex post*-Forschung den anderen Untersuchungsdesigns überlegen ist, dies ist jedoch nicht der Fall. *Ex post*-Forschung ist beispielsweise der besonderen Gefahr des Nachrationalisierens ausgesetzt. So besteht das Problem, dass jeder statistische Zusammenhang auch als Kausal-

zusammenhang fehlinterpretiert werden kann. Ob diese Zusammenhänge aber tatsächlich bestehen, ist immer theoretisch-argumentativ zu stützen. Dabei zeigt gerade die Studie von Rösel und Sonnenburg (2016): Je mehr erklärende Variablen in ein komplexes Rechnungsmodell aufgenommen werden, desto mehr verschwindet der vermeintliche Zusammenhang von Gebietsreformen mit der Wahl rechtspopulistischer Parteien. Den Kausalitätskriterien von oben folgend, deutet dies lediglich auf einen statistischen Zusammenhang hin.

Es spricht einiges dafür, dass dieser Zusammenhang – insbesondere für Gebietsreformen – lediglich konstruiert und damit künstlich ist. Überspitzt formuliert, wird die Wahlentscheidung nicht kausal darauf zurückzuführen sein, ob eine Regierungspartei eine Gebietsreform durchgeführt hat oder nicht. Es ist immer eine Gemengelage von Politikinhalten und situativen Faktoren, die als weitaus wichtiger wahrgenommen werden als Gebietsreformen. In diesem Zusammenhang sei gerade auf andere relevante Einflussgrößen hingewiesen, wie etwa die „Flüchtlingskrise“ oder die als defizitär wahrgenommenen Institutionen der EU.

Die Problematik der Kausalität hängt eng zusammen mit der erwähnten Komplexität des Untersuchungsgegenstands. Das bedeutet, dass für das Gelingen von Reformvorhaben eine Vielzahl verschiedener Faktoren zusammenspielt, die sich nicht alle messen und erheben bzw. auswerten lassen. Außerdem sind nicht alle Faktoren direkt beeinflussbar. Je komplexer aber der Untersuchungsgegenstand ist, desto schwieriger sind einzelne Effekte nachzuweisen. Gleichzeitig wird es aber umso wichtiger, diese Effekte zu isolieren, um sie überhaupt nachweisbar zu machen. Das Kausalitätsproblem beschreibt somit einen ganz zentralen Aspekt jeglicher empirischer Forschungsarbeit. Dieses wird aber dann besonders virulent, wenn politische Schlussfolgerungen aus Forschungsarbeiten abgeleitet oder damit legitimiert werden. Das gilt insbesondere dann, wenn der eigentliche Zweck der Forschung zunächst nicht der Wissenstransfer war.

Ein anderes, nicht weniger prominentes Beispiel für die Problematik von Korrelation und Kausalität ist der Zusammenhang zwischen Gebietsreformen und der Qualität von öffentlichen Dienstleistungen. So misst etwa Reingewertz (2012), etwas zugespitzt formuliert, die Bildungsqualität mit Hilfe der Testergebnisse von Matrikulationsexamen oder die Managementqualität mit Hilfe der Steuereinnahmen. Demnach würden Noten schlechter werden und Steuereinnahmen sinken, wenn sich die Qualität jeweils verschlechtert (Reingewertz 2012). Beides jedoch direkt auf Gebietsreformen zurückzuführen, erscheint mehr als fraglich. Sowohl Noten als auch Steuereinnahmen werden auch durch andere Maßnahmen beeinflusst, wie etwa Lehrpläne bzw. Wirtschaftswachstum.

Derartige Kausalzuschreibungen sind problematisch, nicht zuletzt da Qualität auch kein eindimensionales Konzept und als solches schwer messbar ist. Wenn empirische Sachverhalte aber nicht messbar sind oder nicht angemessen differenziert und operationalisiert werden, lassen sich auch Kausalaussagen nicht überzeugend begründen. Ähnlich verhält es sich auch mit anderen mehrdeutigen Konstrukten wie Effizienz oder Wirksamkeit (Callanan et al. 2014). Genaue Zuschreibungen von Ursache und Effekt sind daher sehr schwierig und mit einer gesunden Skepsis zu betrachten.

Dabei gilt grundsätzlich, dass längere und mehrdeutige Kausalketten die sozialwissenschaftliche Forschung immer vor erheblich größere Herausforderungen stellen als kürzere und eindeutige. Das Problem wird auch dadurch nicht reduziert, dass man relevante Einflussfaktoren einfach ignoriert und „wegdefiniert“. Mit zunehmender Länge der Kausalketten, wird es immer schwieriger, die Effekte – etwa von Reformen – sauber zu messen und in Sinnzusammenhänge zu setzen. Von Gebietsreformen auf Politikverdrossenheit, Servicequalität oder Effizienz zu schließen, impliziert zweifellos vergleichsweise lange Kausalketten, die umfangreiche Differenzierungen, Prüfungen und Diagnoseverfahren notwendig machen.

### 6.3 Datenverfügbarkeit und Messbarkeit

Ein weiterer wichtiger Aspekt ist die Verfügbarkeit und die Nutzung von Daten. In der heutigen Zeit ist die Verfügbarkeit von Daten in entwickelten Ländern, zumindest was offizielle Daten und Statistiken betrifft, kaum mehr ein Problem. Vielmehr haben sich die Güte, die Auswahl der „richtigen“ bzw. „passenden“ Daten und deren Vergleichbarkeit als besonders relevant herausgestellt.

So verweisen beispielsweise Blume und Blume (2007) darauf, dass viele Untersuchungen schon deshalb nicht verlässlich durchgeführt werden können, weil die zur Verfügung stehenden Daten nicht für die Beantwortung der Forschungsfrage geeignet sind. Konkret werden dabei Untersuchungen zum Wirtschaftswachstum angeführt, das oft gar nicht in der notwendigen regionalen Untergliederung erhoben wird. Im Vergleich dazu sind Daten über den Schulstand oder über Ausgaben viel besser verfügbar (Blume/Blume 2007).

Generell werden in der Forschung verschiedene Arten von Daten verwendet. Zu unterscheiden sind zunächst Primär- und Sekundärdaten. Während erstere ausschließlich für den Untersuchungszweck erhoben werden, sind letztere zweitgenutzte Daten. Sie wurden nicht für den originären Untersuchungszweck erhoben und können beispielsweise aus unterschiedlichen Datenquellen stammen. Diese Datenquellen lassen sich wiederum unterteilen in wissenschaftliche Daten, offizielle Daten statistischer Ämter und Organisationen sowie Verwal-

tungsdaten. Während die Erstgenannten im Idealfall nach dezidiert wissenschaftlichen Kriterien erhoben werden, sind für die beiden letztgenannten andere Zwecke von Bedeutung. Für die Daten statistischer Ämter spielt beispielsweise Vergleichbarkeit eine große Rolle (etwa zwischen Regionen, Ländern, etc.), wohingegen für Verwaltungsdaten vor allem die Verwendbarkeit der Informationen für administrative Prozesse von Bedeutung ist (z. B. Daten aus dem internen Rechnungswesen).

Während also für die wissenschaftlichen Daten die Herausforderung in der Operationalisierung – d.h. im Messbarmachen – von empirischen Phänomenen liegt, ist für die Verwaltungsdaten vor allem deren Übertragbarkeit und die Nutzbarkeit ein entscheidendes Kriterium. Nur weil Datenbestände umfassend sind, bedeutet das nicht zwangsläufig, dass alle Daten auch ohne weiteres nutzbar sind. Wenig überraschend verweist gerade die internationale Spitzenforschung immer wieder auf sogenannte *proxies* (also quasi Stellvertreter-Variablen; vgl. Callanan et al. 2014; Allers/Geertsema 2016), die zwar Näherungsmaße für bestimmte Sachverhalte darstellen sollen, aber nicht dazu in der Lage sind, diese eindeutig abzubilden.

Die zahlreichen Untersuchungen zu Gebietsreformen belegen eine große Reichhaltigkeit an Daten und Datenquellen (siehe oben). Dabei fällt jedoch auf, dass sich ein Großteil der Datenbestände aus Sekundärdaten zusammensetzt (siehe Tabelle 12 unten). Die Verwendung der Daten statistischer Ämter ist zwar per se nicht problematisch, allerdings ist es eine der grundlegenden Aufgaben wissenschaftlicher Forschung, deutlich zu machen, warum die verwendeten Daten für die genannten Untersuchungszwecke relevant sein könnten. Nicht selten kommt es zu Erhebungen und Untersuchungen, die bei genauerem Hinsehen Fragen aufwerfen, ob mit den verwendeten Daten auch tatsächlich jene Fragen beantwortet werden können, die sich die Forscher eingangs gestellt haben. Dazu seien hier nur einige Beispiele aufgelistet, um die Tragweite dieses Problems etwas zu verdeutlichen.

- Sehr oft wird in der Forschung beispielsweise eine sogenannte quasi-experimentelle Methode angewandt, die suggerieren soll, dass eine große Nähe zur Naturwissenschaft besteht und damit eine hohe Belastbarkeit der Untersuchungen einhergeht. Schon diese Grundprämisse ist aber angesichts der Komplexität von Gebietsreformen höchst fragwürdig, da keinesfalls die gesamte soziale Umwelt kontrollierbar ist. Ein weiteres Problem ist, dass oft unterstellt wird, dass das Ereignis der Gebietsreform unabhängig von anderen Sachverhalten, wie etwa Wirtschaftswachstum, Ausgabendruck und Bevölkerungswachstum wäre. Gerade die internationale Spitzenforschung ist jedoch dazu in der Lage, diese Annahme hinreichend kritisch zu hinterfragen und zu würdigen. In bestimmten Kontexten beeinflusst der Problem-

druck aber eindeutig die Entscheidung zu Gebietsreformen und hier sind Rahmenbedingungen und Reformentscheidung eben nicht unabhängig voneinander – und müssen daher hinreichend geprüft werden (Lüchinger/Stutzer 2002; Swianiewicz/Lukomska 2017; Reingewertz 2012).

- Ferner spielen Variationen in der Datenqualität oder der Verfügbarkeit von Daten eine Rolle. Im internationalen Raum ist es beispielsweise bedeutsam, ob Gebietsreformen im Rahmen einer Freiwilligkeitsphase realisiert werden. Abgesehen von lückenhaften Informationen aus den Medien gibt es dazu aber oft keine systematischen Erhebungen (Blom-Hansen et al. 2014). Nicht ohne Grund fordert eine Übersichtsstudie Qualitätsverbesserungen in der Forschung über Gebietsreformen, etwa durch Methodenkombinationen (Schaap/Karsten 2017).
- Oft liegen auch keine Informationen darüber vor, wie sich zusammengesessene oder fusionierte Gebietskörperschaften entwickelt hätten, wenn sie nicht fusioniert worden wären (d.h. die Gegenprobe), da die Daten konsequenterweise nicht mehr in separater Form erhoben werden (vgl. Kettunen 2014). Folglich wird streng genommen nur in eine Richtung gemessen. Ohnehin gibt es kaum Forschung dazu, ob Gebietsteilungen entgegengesetzte Effekte von Gebietsfusionen haben (Swianiewicz/Lukomska 2017).
- Es fehlen oft Wirkungsindikatoren und aussagekräftige Kennzahlen zur Wirtschaftlichkeit (z. B. Ist-Kosten je Outputeinheit). Die Systeme der Kosten- und Leistungsrechnung sind oft unausgereift und werden (naturgemäß) zu internen Zwecken verwendet, daher stehen diesbezügliche Daten nach außen (z. B. für wissenschaftliche Untersuchungen) nicht oder nur eingeschränkt zur Verfügung. Im Übrigen leiden Daten der kommunalen Kassenstatistik unter der Umstellung auf das kaufmännische Rechnungswesen, da die Statistik tendenziell noch auf Daten im Sinne der Kameralistik basiert.
- Die Problematik der Datenverfügbarkeit führt zu einer Verengung der Forschungsansätze. Das zeigt sich daran, dass es eigentlich nur verwaltungs- und vor allem finanzwissenschaftliche Untersuchungen gibt, jedoch keine betriebswirtschaftlichen Analysen. Untersuchungen finden häufig auf der Makroebene mit wenig aussagekräftigen Aggregatdaten statt – die Meso- (Produkt-/Fachbereiche) und Mikroebene (individuelle Faktoren, z.B. Handlungsorientierungen bzw. Einstellungen von Führungskräften, etc.) werden hingegen praktisch nicht oder nur sehr selten berücksichtigt.
- Zielkonflikte werden etwa aufgrund methodischer Annahmen (wie etwa die Linearität von Zusammenhängen) nicht angemessen abgebildet und modelliert. Allgemeine institutionelle oder ökonomische Vorteile (Makroebene) können beispielsweise als individuell nachteilige Veränderungen (Mikroebene) wahrgenommen werden. Die Perspektive der Beschäftigten auf Einsparungen und Effizienzverbesserungen ist beispielsweise zurückhaltend

bis ablehnend. Nicht ganz zu Unrecht werden im Zusammenhang mit Gebietsreformen Personalabbau in Verbindung mit individueller Aufgabenverdichtung, Überalterung durch Einstellungsstopps, verringerte Aufstiegsmöglichkeiten wegen „verstopfter“ Führungsstrukturen nach der Fusion, längere Arbeitswege usw. befürchtet.

Diese Beispiele sollen keinesfalls davon abhalten, Sekundärdaten zu verwenden. Sie sollen aber verdeutlichen, dass es besonders wichtig ist, die Verwendung und die Zwecke von Sekundärdaten ausführlich darzulegen und plausibel zu machen. Dazu gehört auch, stärker hervorzuheben, welche Datenquellen überhaupt zur Anwendung gekommen sind (offizielle Statistiken, Verwaltungsdaten, etc.). Denn hierdurch ändert sich die Interpretation der Daten hinsichtlich ihrer Belastbarkeit und Güte. So werden etwa Verwaltungsdaten mit anderen Zwecken erhoben und gepflegt als beispielsweise Daten, die zu rein wissenschaftlichen Zwecken erhoben wurden.

Es gibt aber auch Untersuchungen mit eigenen qualitativen (z. B. Interviews) und quantitativen (z. B. schriftlichen Befragungen) Erhebungen, die unter anderem im Rahmen von Gutachten und Auftragsforschung durchgeführt wurden (vgl. Hesse 2008) und einen hohen Forschungsaufwand mit sich bringen. Der Vorteil dieser Daten liegt in der spezifischen, punktgenauen Erhebung von Informationen bezogen auf den jeweiligen Untersuchungszweck. Dies bedeutet zwar, dass bestimmte Sachverhalte nur indirekt gemessen werden können, weil es keine direkten Maße für jedes Erkenntnisinteresse gibt.

Streng genommen müssten die Datenerhebungen für die *ex ante*-, *ex post*- oder Begleitforschung von politischen Reformvorhaben, jeweils vor den Reformen als sogenannte Grundlinien und Basisdaten definiert werden (Studerus 2016). Dazu fehlen jedoch oftmals die Zeit und die notwendigen Ressourcen. Zudem ist fraglich, ob für jedes Reformvorhaben Begleitforschung und Evaluation notwendig sind. Dennoch führt dies gerade bei *ex post*-Betrachtungen dazu, dass die Datenbestände nicht direkt auf den Reformfokus und dessen Ziele passen. Dagegen ist nichts einzuwenden. Dennoch bedeutet es erhöhte Anforderungen hinsichtlich der Interpretierbarkeit der Ergebnisse. Wenn Ziele jedoch klar definiert und mit Messgrößen hinterlegt werden, dann kann im Zeitverlauf näherungsweise untersucht werden, ob eine Reform in Sachen Zielerreichung erfolgreich war oder nicht. Andernfalls sind lediglich plausible Schätzungen möglich. Diese sollten aber grundsätzlich als solche klar kenntlich gemacht und kritisch gewürdigt werden.

## 6.4 Generalisierbarkeit von Befunden

Ein weiterer wichtiger Aspekt ist die Übertragbarkeit von Befunden. Zwar sind die Regeln für die Generalisierbarkeit von empirischen Befunden vergleichsweise klar (etwa Repräsentativität, Validität, d.h. Gültigkeit, und Reliabilität, d.h. Verlässlichkeit von Messungen, etc.), dennoch handelt es sich insbesondere in der empirischen Sozialforschung um ein zentrales Problem. In der vergleichenden Forschung wird dies oft als das „Äpfel- und Birnen-Problem“ bezeichnet, welches auf eine eingeschränkte Vergleichbarkeit von ausgewählten Fällen hindeutet. Dies gilt oft schon für die Studien innerhalb eines Landes, die mit unterschiedlichen Methoden, Foki, Zielgruppen, Zeitpunkten, (etc.) erstellt worden sind (Schaap/Karsten 2017).

In der Tat werfen zahlreiche Untersuchungen zu Gebietsreformen die Frage auf, inwieweit die erhobenen Befunde auf andere Länder oder Verwaltungseinheiten übertragbar sind. Problematisch ist in diesem Zusammenhang oft der Vorwurf der Nicht-Vergleichbarkeit, der Veränderungsbemühungen in Hinblick auf Reformen in anderen Ländern unmittelbar diskreditiert. Ohne Zweifel ist der Blick in andere Länder hilf- und lehrreich, wenngleich nicht jedes Vorhaben eins zu eins übernommen werden kann. Andere Verwaltungsstrukturen und -kulturen, politische Gegebenheiten, sozialstrukturelle Zusammensetzungen, territoriale Besonderheiten (u.v.a.m.) machen jedoch differenzierte Betrachtungen und Bewertungen notwendig. Außerdem ist die gegenwärtig noch geringe Anzahl an internationalen Vergleichsstudien zu Gebietsreformen letztlich ein Ausdruck der schwierigen Übertragbarkeit von Befunden (De Ceuninck et al. 2010; Swianiewicz/Lukomska 2017).

Von gewichtiger Bedeutung sind beispielsweise Sachverhalte wie Verwaltungskulturen und -traditionen, die insbesondere bei der Leistungserbringung oder bei der Durchführung und Umsetzung von Reformen eine wichtige Rolle spielen (De Ceuninck et al. 2010). Daher erweisen sich gerade Zeitreihenvergleiche als besonders wichtig. Durch vorher-nachher Untersuchungen lassen sich bei verlässlicher Datenbasis belastbare Erkenntnisse gewinnen (Schaap/Karsten 2017) – vorausgesetzt, die Daten der von Gebietsreformen betroffenen Fälle sind sowohl vor als auch nach der Gebietsreform verfügbar.

Wichtig für die Übertragbarkeit von Befunden ist die Definition eines *tertium comparationis*, d.h. eines Vergleichskriteriums, das unabhängig von Rahmenbedingungen (externer Varianz) gültig ist. In diesem Falle lassen sich die Funktionsbedingungen eingrenzen, unter denen bestimmte Reformvorhaben bestimmte Effekte nach sich ziehen. Bisher erfüllen jedoch nur wenige vergleichende Untersuchungen diese Kriterien, die in der Komparatistik (d.h. der vergleichenden Politikwissenschaft) einen sehr hohen Stellenwert genießen.

Viele der bisher vorgenommenen Untersuchungen sind im Grunde Fallstudien, die mehr oder weniger nebeneinander stehen. Für die deutschen Bundesländer (oder zumindest für Teilmengen davon) kann zwar eher auf eine Vergleichbarkeit geschlossen werden, dennoch wäre es auch hier notwendig, die unterschiedlichen Rahmenbedingungen zu kontrollieren, bevor diese Forschungsarbeiten direkt verglichen werden könnten. Beispielhaft sei nur auf unterschiedliche Bevölkerungsstrukturen, soziale Problemlagen oder fiskalische Ausgangssituationen verwiesen.

Ein solcher Vergleich würde daher voraussetzen, dass die Bundesländer in ähnlichen Untersuchungen betrachtet worden wären, um die Ergebnisse einander gegenüberstellen zu können. Denkbar wäre dies beispielsweise für verschiedene Untersuchungen von Hesse (1999, 2000, 2004, 2007, u. a. m.), die jedoch auch verschiedene Nuancen enthalten. Unabhängig davon ist das Konsolidieren von verschiedenen Forschungsarbeiten nur dann möglich, wenn Daten und Untersuchungskonzepte weitestgehend identisch sind.

*Tabelle 11: Forschungsansätze, Methoden und Erhebungsinstrumente*

Design	Methoden	Vertreter
Ex ante	Deskriptiv und Analytisch	Bogumil 2016; Bogumil et al. 2014; Färber 2014; Hesse 2000; Hesse 2007; Junkernheinrich et al. 2011; Kuhlmann et al. 2012; SRH 2009; Ragnitz et al. 2010; Ziekow et al. 2013; Seitz 2005, 2007 a, b
	Mixed Methods Ansatz	Hesse 2008
Ongoing	Beobachtung	Hesse 2014, 2015
Ex post	Quasi-experimentell u. a. <i>difference-in-difference, matching</i> und <i>diachrone Vergleiche, synthetic control method</i>	Blom-Hansen et al. 2016, 2014; Kjær et al. 2010; Hansen et al. 2004; Hansen 2015, 2012; Foged 2016; Saarimaa 2015; Blesse/Baskaran 2013; Hinnerich 2009; Allers/Geertsema 2015; Rösel 2015, 2016 a, b; Reingewertz 2012; Allers/Geertsema 2016; Jordahl 2010; Steiner/Kaiser 2016; Moiso/Uusitalo 2013; De Ceuninck/Valcke 2010; Swianiewicz/Lukomska 2017; Steiner/Kaiser 2016; Studerus 2016
	Nicht experimentelle Ansätze	Blume/Blume 2007; Hanes 2015; Terlouw 2016; Schaap/Karsten 2015; Kraaykamp et al. 2001; Büchner/Franzke 2001 b; Kettunen 2014; Callanan et al. 2014; Tavares/Rodrigues 2015; Lüchinger/Stutzer 2002; Aulich et al. 2014; Bikker/Van der Linde 2016; Ems/Greiner 2017
Reviews		Dollery/Fleming 2006; Schaap/Karsten 2017

Wenn die Übertragbarkeit von empirischen Befunden ein verständlicherweise großes Problem ist, dann bedeutet dies aber auch, dass die Generalisierbarkeit von Befunden nur in Ausschnitten gegeben ist. Wenn beispielsweise behauptet wird, dass Gebietsreformen zu geringeren Kosten oder Ausgaben führen (Tava-

res/Rodrigues 2015), diese erhöhen (Bikker/Van der Linde 2016) oder kein Zusammenhang festgestellt werden kann (teilweise Seitz 2005), dann gilt dies zunächst nur im Kontext der jeweiligen Studie und der dort ausgewerteten Datenbestände. Daraus jedoch die Schlussfolgerung abzuleiten, dass Gebietsreformen auch unter anderen Umständen solche Effekte nach sich ziehen, widerspricht schon dem gegenwärtig vorliegenden Erkenntnisstand, der offenkundig alles andere als eindeutig ist. Insofern hilft hier nur die weiter oben bereits beschriebene differenzierte Betrachtungsweise des Themas. Alles andere wären unseriöse und unzulässige Generalisierungen und Pauschalisierungen, die einem solch komplexen Untersuchungsgegenstand nicht gerecht werden.

Dies setzt schlussendlich hohe Anforderungen an die Betrachtung der Befunde. Sind diese nicht eindeutig, erlaubt dies, für sehr unterschiedliche Sichtweisen immer die geeignete wissenschaftliche Untersuchung zu zitieren, ohne jeweils auf deren genaue Gültigkeitsbedingungen zu achten. Dadurch kann letztlich eine selektive Wahrnehmung bzw. Nutzung von Forschungsergebnissen befördert werden, die aber für das Verständnis des Untersuchungsproblems bzw. für das Begreifen sozialer Realität nur wenig hilfreich ist.

