Pietsch, Marcus; Janke, Nike; Mohr, Ingola

Führt Schulinspektion zu besseren Schülerleistungen? Difference-in-Differences-Studien zu Effekten der Schulinspektion Hamburg auf Lernzuwächse und Leistungstrends

Zeitschrift für Pädagogik 60 (2014) 3, S. 446-470

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 3 Mai/Juni 2014

■ *Themenbereiche*

Erziehungswissenschaftliche Bedeutung literarischer Texte

Schule im öffentlichen Diskurs

■ *Allgemeiner Teil*

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 60 – Heft 3
Mai/Juni 2014

Inhaltsverzeichnis

Besprechungen

Dokumentation

Table of Contents

Marcus Pietsch/Nike Janke/Ingola Mohr

Führt Schulinspektion zu besseren Schülerleistungen?

Difference-in-Differences-Studien zu Effekten der Schulinspektion Hamburg auf Lernzuwächse und Leistungstrends

Zusammenfassung: Schulinspektionen sollen zu verbesserten Schülerleistungen auf Einzelschul- und Systemebene führen. Während für Schulinspektionen in Deutschland bislang keine empirischen Befunde zur diesbezüglichen Wirksamkeit vorliegen, zeigen internationale Studien, dass es Schulinspektionen in der Regel nicht gelingt, Leistungssteigerungen herbeizuführen. Jedoch sind diese Befunde aufgrund von Stichprobenproblemen in den Studien meist wenig belastbar. Im vorliegenden Beitrag wird daher am Beispiel der Schulinspektion Hamburg erstmals für eine Schulinspektion in Deutschland mithilfe von Trenddaten des Hamburger Zentralabiturs sowie Längsschnittdaten der Studie Kompetenzen und Einstellungen von Schülerinnen und Schülern (KESS) überprüft, welche Effekte auf Schülerleistungen empirisch nachweisbar sind. Mögliche Stichprobenprobleme werden dabei in den Analysen explizit berücksichtigt, um empirisch belastbare Aussagen zur Wirksamkeit von Schulinspektion auf Schülerleistungen treffen zu können.

Schlagworte: Difference-in-Differences, Schülerleistungen, Schulinspektion, Selektionseffekte, Wirksamkeit

1. Einleitung

In der erziehungswissenschaftlichen Fachliteratur wird Schulinspektion als Intervention auf Ebene der Einzelschule verstanden, deren Wirksamkeit sich daran messen lassen muss, ob es ihr gelingt, zu einer Verbesserung von Schülerleistungen beizutragen (vgl. Ehren & Visscher, 2006). Diejenigen Studien, die sich mit der Frage auseinandersetzen, ob Schulinspektionen diesem Anspruch gerecht werden, weisen jedoch meist nach, dass diese vergleichsweise wirkungslos sind (vgl. Gärtner & Pant, 2011; Husfeldt, 2011). Zwar lassen sich teilweise positive Effekte in Bezug auf bestimmte Gruppierungen, wie z. B. kleinere Mädchenschulen (vgl. Shaw, Newton, Aitkin & Darnell, 2003) oder Schulen mit leistungsschwächeren Schülerinnen und Schülern (vgl. Allen & Burgess, 2012), feststellen. Gleichwohl zeigen all jene Studien, die sich auf Effekte der englischen Schulinspektion OFSTED (The Office for Standards in Education) beziehen, dass bezogen auf das Gesamtsystem keine oder sogar leicht negative Wirkungen auf Schülerleistungen als Folge von Schulinspektionen beobachtet werden können (vgl. de Wolf & Janssens, 2007). Einzig eine aktuelle Studie aus den Niederlanden (vgl. Luginbuhl, Webbink & de Wolf, 2009) zeigt entgegen der allgemeinen Befundlage, dass Schulin-

spektionen auf Systemebene ggf. zu leichten Steigerungen von Schülerleistungen in der Größenordnung von zwei bis drei Prozent einer Standardabweichung in den ersten zwei Jahren nach einer Inspektion führen könnten, wobei dieser Befund, nach Aussage der Autoren, aus methodischen Gründen, jedoch nur eingeschränkt belastbar sei.

Mit Blick auf die Schulinspektionen in Deutschland liegen entsprechende Befunde bislang nicht vor (vgl. Gärtner & Pant, 2011). Dies ist umso problematischer, als sich die Ergebnisse der internationalen Forschung kaum auf deutsche Inspektorate übertragen lassen, da Schulinspektionen in Deutschland weniger als Kontroll- denn als Entwicklungsinstrument betrachtet werden (vgl. Böttcher & Kotthoff, 2010). Hierzulande wird Schulinspektion somit – im Gegensatz zum englischen Paradigma, in dem eine Entwicklung durch Wettbewerb angestrebt wird – vorrangig als eine Dienstleistung für die Einzelschule verstanden, bei der die Bereitstellung der Evaluationsbefunde für Transparenz gegenüber den Schulbeteiligten sorgen und als Basis für einzuleitende Qualitäts- und Schulentwicklungsmaßnahmen dienen soll (vgl. Böttger-Beer & Koch, 2008).

Zwar könnte man Untersuchungen zum Umgang mit Leistungsdaten aus der externen Evaluation und ihrer differenziellen Wirkung nach dem jeweiligen Paradigma als Referenzmaßstab heranziehen. Jedoch lassen sich beide Bereiche nicht ohne Weiteres vergleichen, da im Rahmen Output-orientierter Evaluationsmaßnahmen im Idealfall keine Annahmen über schulische Wirksamkeitsmechanismen getroffen werden und die Prozessanalyse den Schulbeteiligten überlassen wird, wohingegen Schulinspektionen explizite Annahmen über Wirksamkeitsmechanismen von Schulen treffen und Schulbeteiligte angehalten werden, ebendiese Mechanismen zu optimieren. Entsprechend sind Schulinspektionen deutlich anfälliger für die Formulierung fehlerhafter programmtheoretischer Annahmen, die, nach der Devise „Operation gelungen, Patient tot“ (vgl. Scheerens, 1990), flächendeckend zu keiner Optimierung oder ggf. sogar zu paradoxen Entwicklungen schulischer Mechanismen und Prozesse führen können, wenn auf Schulebene versucht wird, mutmaßliche Defizite abzustellen.

Mit Blick auf die Schulinspektionsforschung kommt hinzu, dass die Forschungsdesigns sowie die Methoden, mittels derer die Wirksamkeit von Schulinspektionen bislang erforscht werden, dem Anspruch an eine Wirkungsanalyse in der Regel nicht genügen (vgl. Luginbuhl et al., 2009; de Wolf & Janssens, 2007). Entsprechend resümieren de Wolf und Janssens (2007, S. 391) am Ende eines umfassenden Reviews zu Studien, die sich mit der Wirksamkeit von Schulinspektionen beschäftigen:

> The main conclusion is that the studies do not provide a clear answer to the question of whether school inspections (…) have causal effects. It is not only methodologically difficult to demonstrate causal effects but the methodology used also appears to have a strongly determinative effect on conclusions concerning the extent and direction of the effects.

Folglich ist bislang vollkommen unklar, ob die international beobachtete Nicht-Wirksamkeit von Schulinspektion auf Schülerleistungen einer mutmaßlich unzureichenden Implementierung bzw. einer mangelhaften Durchführungspraxis oder aber den metho-

dischen Unzulänglichkeiten der bislang vorliegenden Untersuchungen geschuldet ist (vgl. Pietsch, Janke & Mohr, 2013).

Im Folgenden wird erstmals die Wirksamkeit einer Schulinspektion in Deutschland auf längsschnittlich (Trend- und Paneldaten) erhobene Schülerleistungen unter Berücksichtigung möglicher Selektionseffekte untersucht. Entsprechend wird zuerst eine Übersicht gegeben zur Wirkungsweise von Schulinspektionen, zur bisher beobachteten Problematik von Selektionseffekten sowie zu den Möglichkeiten, mit diesen umzugehen. Mithilfe einer Kombination von Zufallsstichprobe und statistischer Kontrolle möglicher Selektionseffekte wird anschließend die Wirksamkeit der Schulinspektion Hamburg auf die Entwicklung von Schülerleistungen untersucht. Abschließend werden die berichteten Befunde diskutiert.

2. Theoretischer Hintergrund

2.1 Annahmen zur Wirkungsweise von Schulinspektion

Schulinspektionen sollen die Qualität von schulischen Prozessen evaluieren, um dazu beizutragen, ein ganzheitliches Bild von Schulqualität zu begründen, das über die Erhebung der fachlichen Stärken und Schwächen von Schülerinnen und Schülern mithilfe von Leistungstests hinausgeht. Hierfür werden normative Vorgaben, die in der Regel in landesspezifischen Qualitätsrahmen oder Qualitätstableaus formuliert wurden, durch Schulinspektoren extern an Schulen evaluiert (vgl. van Ackeren & Klemm, 2009). Die Schulinspektion hat dabei vor allem drei Funktionen (vgl. Pietsch, Schnack & Schulze, 2009):

1) Eine Garantiefunktion – die Schulinspektion soll elementare Standards von Bildungsqualität an Schulen gewährleisten.
2) Eine Monitoringfunktion – die Schulinspektion soll für unterschiedliche Akteure im Bildungssystem Informationen bereitstellen.
3) Eine Katalysefunktion – die Schulinspektion soll einen verbesserten Service für das einzelschulische Qualitätsmanagement bieten.

Während der Schwerpunkt der externen Schulevaluation auf internationaler Ebene vor allem den Bereich des Monitorings umfasst (vgl. Husfeldt, 2011), verfolgen Schulinspektionen in Deutschland derzeit vor allem das Ziel, Schul- und Unterrichtsentwicklung mittels der Rückmeldung von Informationen zur extern wahrgenommenen Qualität von Schule und Unterricht zu stimulieren (vgl. Böttcher & Kotthoff, 2010). Wie Pietsch, Schulze, Schnack und Krause (2011) herausstellen, knüpfen die diesbezüglichen Wirksamkeitserwartungen an Schulinspektionen vor allem an die Forschung zum zielorientierten Feedback an (vgl. Kluger & DeNisi, 1996; Visscher & Coe, 2003). Entsprechend werde erwartet, dass das Aufzeigen von Differenzen zwischen normativ vorgegebenen Soll- und empirisch beobachteten Ist-Ständen dazu führe, dass in extern evaluierten

Schulen infolge der Rückmeldung eine Handlungsoptimierung geplant werde, die es ermöglicht, anzustrebende Ziele in Zukunft besser zu erreichen.

Rahmenmodelle, die ebenjene pädagogischen Verarbeitungsprozesse in den Blick nehmen, haben beispielsweise Cousins & Leithwood (1993), Helmke und Hosenfeld (2005) oder Reezigt und Creemers (2005) vorgelegt; wobei es sich hierbei weniger um Theorien denn um Zusammenstellungen von hypothetischen und empirisch bekannten Bedingungen und Mechanismen der Informationsverarbeitung handelt. In diesen Modellen werden Rückmeldeinformationen als Impuls verstanden, der Schulentwicklung, im Sinne entscheidungstheoretischer Optimierungsstrategien (vgl. Tarter & Hoy, 1998), stimulieren soll. Eine solche Annahme zur Nutzung von Evaluationsbefunden unterstellt dann konsequenterweise, dass Entscheidungen rational, auf Basis bereitgestellter Informationen in einem prozessualen Ablauf getroffen werden (vgl. Hyyryläinen & Viinamäki, 2008). Abgebildet wird der Prozess der innerschulischen Verarbeitung daher z. B. bei Helmke und Hosenfeld beginnend mit der Rezeption der Ergebnisse, der anschließenden Reflexion der Befunde und den final daraus abgeleiteten Aktionen. D. h. infolge der Übermittlungen und der Auseinandersetzung mit den Inspektionsbefunden werden Erklärungen für Ist-Soll-Unterschiede gesucht – wobei eventuell weitere Datenquellen herangezogen werden –, um darauf aufbauend Maßnahmen zu planen und umzusetzen, die die Verbesserung resp. Optimierung der Schul- und Unterrichtsqualität zum Ziel haben.

Dabei nutzen alle Autoren kontextualisierte, ökologische und vergleichsweise umfassende Modelle, die sowohl schulinterne als auch schulexterne und teilweise sogar Persönlichkeitsmerkmale von Lehrenden und Schulleitungen als moderierende Faktoren mit in den Blick nehmen. Die Modelle unterscheiden sich jedoch in ihrer jeweiligen Reichweite (vgl. Tab. 1). Während Reezigt und Creemers (2005) vor allem innerschulische Aspekte der Schulkultur in den Blick nehmen, weisen Helmke und Hosenfeld (2005) darüber hinaus auch detailliert auf die Relevanz individueller Persönlichkeitsmerkmale von Lehrenden und schulischen Entscheidungsträgern im Verarbeitungsprozess hin. Cousins und Leithwood (1993) wiederum zeigen, dass darüber hinaus auch die soziale Interaktion innerhalb der Schule sowie zwischen Schulbeteiligten und Evaluatoren eine Rolle dabei spielt, inwieweit Evaluationen Wirksamkeit entfalten können.

Weiterhin werden in allen Modellen externe Bedingungen genannt, die den innerschulischen Verarbeitungsprozess von Rückmeldeinformationen beeinflussen. Welche Merkmale die Wirksamkeit speziell von Schulinspektionen en Detail moderieren, haben Ehren und Visscher (2006, vgl. Abb. 1), basierend auf Visscher und Coes (2003) Arbeiten zur Nutzung von Schul-Performance-Feedback-Systemen, herausgearbeitet. In diesem Zusammenhang weisen die Autoren explizit darauf hin, dass Schulinspektionen sowohl erwünschte als auch unerwünschte Wirkungen nach sich ziehen können – ein insbesondere bei Evaluationen im öffentlichen Sektor häufig beobachtetes Phänomen (vgl. z. B. Leeuw & van Thiel, 2002; Smith, 1995). Grundsätzlich gehen jedoch auch Ehren und Visscher davon aus, dass Wirkungen von Schulinspektionen zwar als Folgen einer kausalen Wirkungskette aus (a) Merkmalen des Schulinspektionsprozes-

	Cousins & Leithwood (1993)	Helmke & Hosenfeldt (2005)	Reezigt & Creemers (2005)
Merkmale der Schule/der Schulkultur	Informationsbedürfnis		
	Fokus auf Veränderung	Innovative und explorative Orientierung	Interner Druck für Veränderungen
	Mikropolitisches Klima		
	Widersprüchliche Informationen		
		Ausstattung der Schule	
		Akzeptanz seitens der Eltern und der Schüler	
		Verbindlichkeit durch Verankerung im Schulprogramm	
			Gelebte Schulautonomie
			Geteilte Zukunftsvisionen
			Kollegiale Zusammenarbeit
		Evaluations- und Kooperationsklima	Willen, eine lernende Organisation zu sein/zu werden
			Bisheriger Verlauf der Schulentwicklung
			Eigenständigkeit mit Blick auf die Entwicklung, die Motivation und das Commitment
			Schulleitung
			Stabilität des Kollegiums
			Zeit, Veränderungen umzusetzen
Merkmale des Individuums	Merkmale der Entscheidungsträger	Vorwissen/Expertise	
		Motivation, Emotion, Volition	
		Selbstwirksamkeit	
		Professionelles Selbstverständnis	
		Stabilität von Gewohnheiten	
	Einstellung zur Akzeptanz der Evaluation	Akzeptanz von Evaluationen	
Interaktive Prozesse	Beteiligung der Nutzer		
	Sozialer Austausch		
	Kontakt nach der Evaluation		
	Aktive Einbeziehung		
	Diffusion der Befunde		

Tab. 1: Innerschulische Determinanten der Evaluationsnutzung

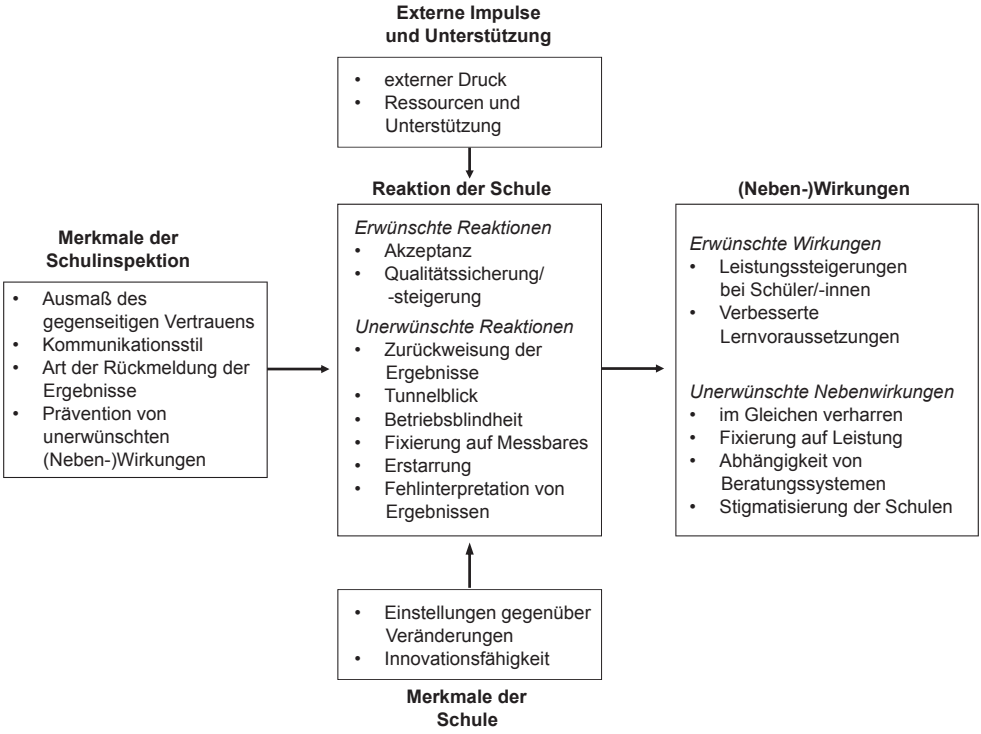


Abb. 1: Modell zur Wirkungsweise von Schulinspektion (vgl. Ehren & Visscher, 2006)

ses und (b) Reaktionen der Schulen auf den Prozess und die Ergebnisse der Inspektion entstehen, wobei neben den Voraussetzungen der Schule auch Unterstützungsmaßnahmen aus dem weiteren Bildungssystem wichtige Determinanten für den Umgang und die Nutzung von Schulinspektionsbefunden durch Schulen als Grundlage für die Schul- und Unterrichtsentwicklung sind.

So ist es einerseits aufseiten der externen Impulsgebung relevant, in welchem Maße finanzielle oder personelle Unterstützungsmaßnahmen für Schulen nach Beendigung der Evaluation bereitgestellt werden oder aber auch ob und wie stark Druck ausgeübt wird, um Veränderungen zu forcieren. Andererseits spielen auch Haltungen und Kompetenzen der Schulverantwortlichen und Lehrerschaft an der evaluierten Schule eine wichtige Rolle dabei, ob Inspektionsbefunde für die Weiterentwicklung von Unterricht und Schule genutzt werden. Diese Faktoren beeinflussen dabei letztlich wiederum, wie Schulen mit den Ergebnissen aus Schulinspektionsverfahren umgehen und ob intendierte Entwicklungen und – vermittelt hierüber – Leistungsziele auch tatsächlich erreicht werden oder ob ggf. sogar unerwünschte Nebenwirkungen oder Performanz-Paradoxa erzielt werden.

Der Wert der vorgestellten Modelle liegt dabei in erster Linie in der Zusammenstellung von Mechanismen und Moderatorvariablen, mit deren Hilfe Wirksamkeitsannah-

men beschreibbar gemacht werden können. Gleichwohl wurde weder herausgearbeitet, wie die individuellen Wahrnehmungs-, Handlungs- und Lernvorgänge von verschiedenen Akteuren im schulischen Mehrebenensystem verknüpft sind, noch welche Handlungsbeiträge die einzelnen Akteure im Rückmeldeprozess liefern und wie die einzelnen Determinanten miteinander zusammenhängen (vgl. Altrichter, 2010). Eine empirische Überprüfung einer komplexen Programmtheorie ist in einem solchen Fall nicht möglich (vgl. Maier, 2008) und würde ggf. dazu führen, dass ad-hoc-Theorien aufgestellt werden, die dem Untersuchungsgegenstand nicht angemessen sind und entsprechend zu Fehlschlüssen führen (vgl. Stuffelbeam & Shinkfield, 2007). Entsprechend betonen Pietsch et al. (2013), dass es – solange keine ausgearbeiteten, validen und empirisch prüfbaren Programmtheorien zur Inspektionswirksamkeit vorliegen – geboten scheint, zur Bestimmung von Inspektionseffekten auf Schülerleistungen Blackbox-Verfahren zu nutzen, die dann jedoch hohen methodischen Anforderungen entsprechen müssen.

2.2 Selektionseffekte als Problem bisheriger Wirksamkeitsanalysen

Insgesamt liegen derzeit sechs solch geforderter Studien zur Wirksamkeit von Schulinspektion auf Schülerleistungen vor (vgl. Cullingford & Daniels, 1999; Luginbuhl et al., 2009; Matthews & Sammons, 2004; Rosenthal, 2004; Shaw et al., 2003; Wilcox & Gray, 1996). Alle diese Studien nutzen *Blackbox*-Evaluationsansätze. D. h. die Komplexität und Kompliziertheit der Intervention Schulinspektion wird in den empirischen Analysen ausgeblendet und die inneren Prozesse der Evaluation sowie die Wirkungsmechanismen werden nicht berücksichtigt (vgl. Scriven, 1994) – es wird einzig und allein untersucht, ob die Intervention Schulinspektion intendierte Effekte nach sich zieht oder nicht.

In einem solch methodisch-orientierten *(method driven)* kausalanalytischen Ansatz kommt der Frage der angewandten Forschungsmethoden eine zentrale Bedeutung zu, können doch Methodenfehler – der Einsatz unzureichender oder unangemessener empirischer Methoden – dazu führen, dass Ergebnisse inakkurat sind (vgl. White, 2010). Das Hauptproblem der bisherigen Wirksamkeitsanalysen besteht dabei darin, dass vollkommen unklar ist, ob die berichteten Effekte auf eine mangelhafte Implementation von Schulinspektionen zurückzuführen oder methodisch-artifizieller Art sind (vgl. de Wolf & Janssens, 2007). Luginbuhl et al. (2009, S. 222) verweisen diesbezüglich darauf, dass insbesondere mögliche Selektionseffekte ein gravierendes Problem in der aktuellen Forschung zur Wirksamkeit von Schulinspektionen darstellen:

> Estimating the effect of school inspections on (…) school performance is difficult because inspectors may not randomly select which schools they inspect. (…) This nonrandom selection can produce an endogeneity bias in the estimates of the impact of school inspections. As a result, an estimated effect could actually be due to correlation between unobserved heterogeneity in the quality of schools and the inspectors' decisions about which schools to inspect.

Generell entsteht bei der Schätzung dieser Art kausaler Effekte das Problem, dass diese auf kontrafaktischen Annahmen beruht (vgl. White, 2010). Für die Untersuchung der Wirkung der Schulinspektion bedeutet dies, dass jede Schule theoretisch erst einmal zwei mögliche Ergebnisse aufweisen kann, je nachdem ob sie inspiziert wurde oder nicht: Wurde an der Schule eine Inspektion durchgeführt, erhält man das Ergebnis y_i^1, wurde an der Schule keine Inspektion durchgeführt, erhält man das Ergebnis y_i^0. Für jede Schule ließe sich dann die Wirkung der Schulinspektion als Differenz zwischen den beiden potenziellen Ergebnissen $y_i^1 - y_i^0$ definieren. Leider können tatsächlich beide Ergebnisse an einer Schule nicht gleichzeitig auftreten – entweder die Schule wurde im entsprechenden Zeitraum inspiziert oder eben nicht. Das jeweils andere Ergebnis gibt es nur als „unbeobachtetes, kontrafaktisches Ergebnis im Sinne einer ‚was-wäre-wenn'-Frage" (Legewie, 2012, S. 127). Die Wirkung der Schulinspektion auf die einzelne Schule lässt sich daher niemals direkt messen.

Ein einfacher Ansatz, diesem Problem zu begegnen, ist, den durchschnittlichen kausalen Effekt als Unterschied zwischen den Ergebnissen der Schulen mit Besuch durch die Schulinspektion und den Ergebnissen von Schulen ohne Besuch durch die Schulinspektion zu interpretieren. Ein Ansatz, der jedoch nicht haltbar ist, wenn die Entscheidung, ob und wann eine Schule inspiziert wird, nicht unabhängig von Merkmalen der Schule erfolgt.

2.3 Möglichkeiten zum Umgang mit Selektionseffekten

Wie Luginbuhl et al. (2009) und auch de Wolf und Janssens (2007) betonen, würde es zur kausalanalytisch-methodisch angemessenen Klärung der Frage, ob Schulinspektion zu einer Verbesserung von Schülerleistungen beiträgt, bereits ausreichen, wenn Schulen im Sinne eines Zufallsexperiments zufällig gezogen würden, um eine Konfundierung von Inspektion und unbeobachteter Heterogenität in Schülerleistungsdaten zu vermeiden. In diesem Fall ließe sich der durchschnittliche kausale Effekt der Schulinspektion durch einen einfachen Mittelwertvergleich zwischen den Schulen mit und ohne Besuch durch die Schulinspektion berechnen.

Stehen hingegen keine experimentellen Daten zur Verfügung – wie es bei der Re-Analyse von Inspektions- im Zusammenhang mit Leistungsdaten üblich ist –, können verschiedene methodische Verfahren angewandt werden, um zufällige Variationen im Treatment-Status zu identifizieren und auf der Grundlage dieser Variationen den kausalen Effekt zu schätzen (vgl. Legewie, 2012; Tab. 2). Dies wird üblicherweise durch Schätzung des Treatment-Effekts nach der Kontrolle beobachtbarer Variablen erreicht, wobei es je nach Datenlage sinnvoll sein kann, einzelne statistische Verfahren miteinander (vgl. Smith & Todd, 2005) oder zur Korrektur von Zuweisungsproblemen ggf. auch mit Zufallsstichproben (vgl. Legewie, 2012) zu kombinieren.

Beim *Standard-Regressionsansatz* wird die Wirkung des Treatments auf die abhängige Variable unter Kontrolle beobachtbarer Variablen, die sowohl mit der abhängigen Variablen als auch mit dem Treatmentstatus zusammenhängen, berechnet. Um je-

Verfahren	Grundgedanke	Voraussetzungen	Probleme	Anwendungsbeispiel
Standard-Regressions-ansatz	Schätzung des Treatmenteffekts unter Kontrolle beobachtbarer Variablen	Alle Variablen, die sowohl auf die Treatment-Zuweisung als auch auf die abhängige Variable wirken können, müssen theoretisch in die Berechnung einbezogen werden.	Im Normalfall ist es nicht gegeben, dass alle beeinflussenden Variablen auch beobachtet und erhoben werden.	Baumert et al. (2009) schätzen mithilfe eines Regressionsmodells die Leistungsentwicklung von Schülerinnen und Schülern an Berliner Schulen.
Matching	Schätzung des Treatmenteffekts durch Vergleich von Paaren, bei denen sich beobachtbare Variablen ähneln	Wie beim Standard-Regressionsansatz müssen auch hier theoretisch alle beeinflussenden Variablen in die Berechnung einfließen.	Auch hier ist es möglich, dass die Selektion auf unbeobachteten Variablen beruht.	Crosnoe (2009) schätzt mithilfe von Matching-Verfahren den Effekt der sozialen Zusammensetzung der Schülerschaft auf verschiedene abhängige Variablen.
Fixed-Effect-Modelle	Schätzung des Treatmenteffekts unter Kontrolle nicht-beobachtbarer Merkmale durch Vergleich von Treatmenteinheiten innerhalb von Gruppen	Verteilung der Einheiten über die Treatment- und Kontrollgruppe ist zufällig, d. h. das Treatment ist unabhängig von den anderen Beobachtungen der Gruppe/des Individuums.	Aussagen können nur über die Individuen/Gruppen getroffen werden, die sich im Treatmentstatus innerhalb der beobachteten Gruppe unterscheiden, im Beispiel etwa nur über Frauen, die zwischen den Beobachtungszeiträumen ein Kind bekommen haben.	Budig und England (2001) untersuchen die Auswirkungen von Mutterschaft auf den Stundenlohn anhand von Paneldaten, indem nur Änderungen im Stundenlohn einer Frau vor und nach der Geburt, aber nicht zwischen unterschiedlichen Frauen mit und ohne Kind verwendet werden.
Difference-in-Differences-Ansatz	Schätzung des Treatmenteffekts durch Kontrolle nicht-beobachtbarer Merkmale unter Konstanthaltung von Merkmalen und Kontrolle globaler Effekte im zeitlichen Verlauf	Es liegen Daten zu mindestens zwei Messzeitpunkten vor, die Treatment-Zuweisung ist von außen gesetzt bzw. erfolgt zufällig.	Es wird angenommen, dass sich der Effekt in der Treatment- und Kontrollgruppe ähnlich entwickelt hätte, wenn kein Treatment erfolgt wäre. Diese kontrafaktische Annahme lässt sich nicht überprüfen, ist aber häufig plausibel.	Helbig et al. (2012) untersuchen mit dem DiD-Ansatz die Auswirkung der Einführung von Studiengebühren auf die Studierneigung in Deutschland.

Tab. 2: Übersicht über verschiedene Verfahren zur Kontrolle zusätzlicher beobachtbarer und unbeobachtbarer Kovariaten

doch wirklich den Einfluss des Treatments auf die abhängige Variable auf diese Weise korrekt schätzen zu können, ist es notwendig, dass alle Variablen, die sowohl auf die Treatment-Zuweisung als auch auf die abhängige Variable wirken können, in die Berechnung einbezogen werden. Nur dann wäre gewährleistet, dass die Verteilung der untersuchten Einheiten auf die Treatment- und Kontrollgruppe so gut wie zufällig ausfällt. Im Normalfall ist es jedoch nicht unbedingt gegeben, dass alle beeinflussenden Variablen auch beobachtet und erhoben werden. Ein geringer Schätzfehler „ist nur dann zu erwarten, wenn dem Sozialforscher reichhaltige pre-treatment Variablen zur Verfügung stehen, die weit über standarddemografische Merkmale hinausgehen, unmittel-

bare Relevanz für den Selektionsprozess haben und präzise gemessen werden“ (Legewie, 2012, S. 131).

Auch beim *Matching* werden kausale Effekte nach der Kontrolle von Variablen, die sowohl die abhängige Variable als auch den Treatmentstatus beeinflussen, geschätzt. Hier werden in einem ersten Schritt Paare von beobachteten Einheiten gebildet, die sich hinsichtlich der kontrollierten Variablen möglichst ähnlich sind, aber unterschiedlichen Treatmentgruppen angehören. Diese gepaarten Einheiten werden dann im zweiten Schritt zur Schätzung der Wirkung des Treatments verwendet. So werden nur Einheiten miteinander verglichen, die sich zwar im Treatmentstatus unterscheiden, aber in Hinblick auf möglicherweise beeinflussende Variablen möglichst ähnlich sind. Wie beim Standard-Regressionsansatz liegt auch hier das Problem darin, dass bedeutsame unbeobachtete Variablen nicht mit einbezogen werden können. Die Validität der Ergebnisse hängt somit auch bei diesem Ansatz vor allem von der Auswahl der kontrollierenden Variablen ab (vgl. Gangl & DiPrete, 2004).

Fixed-Effect (FE) -Modelle vermeiden das Problem der oben genannten Methoden, dass die Selektion möglicherweise auch mit unbeobachteten Merkmalen zusammenhängt, indem hier mehrere Beobachtungen innerhalb von Gruppen (also z. B. von Klassen innerhalb einer Schule) oder Individuen (z. B. Paneldaten) verwendet werden, um nach unbeobachteten Merkmalen auf der Gruppen- oder Individualebene zu kontrollieren (vgl. Allison, 2009). Es wird also nur ein bestimmter Teil der Variation verwendet (nur die Variation zwischen Klassen einer Schule und nicht die zwischen den Schulen oder nur die Variation zwischen zwei Zeitpunkten eines Längsschnitts, aber nicht zwischen den Personen mit verschiedenen Merkmalsausprägungen), sodass die Variation zwischen den Gruppen (z. B. Schulen oder Individuen) genutzt wird, um auch nach unbeobachteten Variablen auf Gruppenebene zu kontrollieren.

Der *Difference-in-Differences*-Ansatz ist eine spezielle Art von FE-Modell, bei dem auf Gruppenebene aggregierte Daten genutzt werden (vgl. Angrist & Pischke, 2009). Es wird der kausale Effekt einer Intervention – hier des Besuchs durch die Schulinspektion – durch den Vergleich der Trends der Gruppe mit und ohne Intervention geschätzt. Man vergleicht daher den Trend in den Schulen, die im entsprechenden Zeitraum von der Schulinspektion besucht wurden, mit dem Trend in den Schulen, in denen keine Schulinspektion stattgefunden hat. Es wird dabei einerseits für Fixed Effects (eine Schulinspektion hat stattgefunden oder nicht) sowie für Effekte, die alle Schulen gleichermaßen betreffen (z. B. die Klassengröße an allen Gymnasien beträgt maximal 28 Schülerinnen und Schüler), kontrolliert. Effekte, die auf das Treatment Schulinspektion zurückgeführt werden, werden aus der intertemporalen Variation zwischen Treatment- und Kontrollgruppe abgeleitet. Der Difference-in-Differences-Ansatz ist jedoch nur bei einer zufälligen Auswahl der Treatmentgruppe anwendbar.

3. Forschungsdesiderata und Fragestellung

Betrachtet man nun, inwieweit in den bislang vorliegenden Studien Zufallsstichproben und/oder methodische Verfahren zum Umgang mit potenziellen Selektionseffekten genutzt wurden (vgl. Tab. 3), wird sichtbar, dass nur in den zwei Studien von Rosenthal (2004) und Lunginbuhl et al. (2009) ein entsprechendes Design angewendet wurde.

Entsprechend muss hervorgehoben werden, dass in den letzten Jahren trotz intensiver Bestrebungen in nahezu keiner Studie die (Nicht-)Wirksamkeit auf Schülerleistungen oder gar die Entstehung von Performanz-Paradoxa durch Schulinspektion empirisch verlässlich nachgewiesen werden konnte und generalisierte Aussagen hierzu zumindest fragwürdig erscheinen.

Problematisch ist vor allem, dass die bisher dokumentierten Feststellungen zur (Nicht-)Wirksamkeit von Schulinspektion auf Schülerleistungen systematisch mit den eingesetzten Methoden variieren und daher unklar ist, ob die Durchführung von Schulinspektionen nachweisbare – ob positive oder negative sei dahingestellt – Effekte nach sich zieht. Zu klären ist daher, ob Effekte von Schulinspektionen auf Schülerleistungen nachweisbar sind, wenn adäquate Methoden der empirischen Kausalanalyse eingesetzt werden, die es ermöglichen, mit potenziellen Selektionseffekten umzugehen.

Studie	Nation	Zufallsstich-probe	statistische Korrektur	Methode	Effekte
Wilcox & Gray, 1996	England	Nein	Nein	–	negativ
Cullingford & Daniels, 1999	England	Nein	Nein	–	negativ
Shaw et al., 2003	England	Nein	Nein	–	negativ
Matthews & Sammons, 2004	England	Nein	Nein	–	positiv
Rosenthal, 2004	England	Nein	Ja	Regression	negativ
Luginbuhl et al., 2009	Niederlande	Nein/Ja	Ja	Fixed Effects	positiv/kein

Tab. 3: Studien zur Inspektionswirksamkeit unter Berücksichtigung des Umgangs mit möglichen Selektionseffekten

4. Evaluation von Inspektionseffekten am Beispiel der Schulinspektion Hamburg

Die Schulinspektion der Hansestadt inspiziert seit dem Jahr 2006 jährlich bis zu 80 Schulen, die nach dem Zufallsprinzip ausgewählt wurden. Ziel der Inspektion ist es, Mindeststandards schulischer Qualität zu sichern, empirische Erkenntnisse zu gewinnen und bereitzustellen sowie Schulentwicklung zu stimulieren. Die Berichte werden nicht veröffentlicht und nur der Schulöffentlichkeit zur Verfügung gestellt. Jede

Hamburger Schule wird dabei im Sinne einer Full-Inspection, mit allen zur Verfügung stehenden Methoden und Verfahren, extern evaluiert. Als Datengrundlage für die Berichterstellung und die Schulrückmeldung dienen Onlinebefragungen und teilstandardisierte Interviews aller Schulbeteiligten, Dokumentenanalysen sowie systematische Unterrichtsbeobachtungen.

Das Rückmeldeverfahren der Schulinspektion Hamburg besteht dabei aus sechs Elementen: (a) einem Feedbackgespräch zwischen Inspektionsteam und Schulleitung am letzten Tag des Schulbesuchs, (b) einer Präsentation des fertiggestellten Inspektionsberichts gegenüber der Schulleitung, ca. zwei bis drei Wochen nach dem Schulbesuch, (c) einer Präsentation gegenüber der Schulöffentlichkeit (auf Wunsch der Schulleitung), (d) der Übergabe des Inspektionsberichts, (e) der Übergabe von (quantitativen) Daten auf CD-ROM und (f) einem Response seitens der evaluierten Schule gegenüber ihrer zuständigen Schulaufsicht, wobei der letzte Teil (Response) nicht mehr in den Aufgabenbereich der Schulinspektion Hamburg fällt, sondern in den Aufgabenbereich der Hamburger Schulaufsicht (vgl. Pietsch, 2011a).

Die Auswahl der zu inspizierenden Schulen erfolgt zufällig, wobei die Schulinspektion in ihrer jährlichen Gesamtstichprobe zwischen Kern- und Ergänzungsstichproben unterscheidet. Während die Kernstichprobe eine Substichprobe der jährlichen Gesamtstichprobe darstellt, anhand derer Berichte auf Systemebene (z. B. Jahresbericht der Schulinspektion Hamburg und Bildungsbericht der Freien und Hansestadt Hamburg) verfasst werden, dient die Ergänzungsstichprobe als zweiter Teil der Gesamtstichprobe dazu, die administrativen Leistungsvorgaben der Inspektion zu erfüllen (vgl. Leist, Pietsch & Vaccaro, 2009). Praktisch ermittelt die Schulinspektion Hamburg diese jährliche Gesamtstichprobe als mehrstufige Zufallsauswahl, die sich an den Merkmalen Schulform und soziale Zusammensetzung der Schülerschaft der Schule orientiert (vgl. Pietsch, 2011b). Als Grundlage für die Stichprobenziehung dient eine schuljährlich aktualisierte Schulliste aller Hamburger staatlichen Schulen ($N_{Schulen2006}$ = 402, $N_{Schüler2006}$ = 164 378), die seitens der Hamburger Schulstatistik bereitgestellt wird und aus der Jahr für Jahr die bereits inspizierten Schulen als nicht stichprobenrelevant im Sinne der jährlichen Grundgesamtheit herausgenommen werden.[1]

Während die Kernstichprobe Jahr für Jahr abgearbeitet werden muss – es sich somit immer um eine echte Zufallsstichprobe von Schulen handelt, die das System hinsichtlich der Merkmale Schulform und soziale Zusammensetzung der Schülerschaft repräsentiert –, kann die Inspektion von Schulen der Ergänzungsstichprobe ggf. zwischen Jahren verschoben werden, wenn besondere Umstände (z. B. Einarbeitung einer neuen

1 Wichtig zu beachten ist hierbei, dass die Stichprobengröße der jährlichen Gesamtstichproben insbesondere aufgrund von strukturellen Änderungen auf Systemebene, wie z. B. Schulschließungen oder -zusammenlegungen (so nahm die Anzahl der staatlichen Schulen in Hamburg zwischen den Jahren 2006 und 2009 z. B. von 402 auf 392 ab), sowie in Abhängigkeit vom verfügbaren Inspektionspersonal (da die Schulinspektion Hamburg nur über 13 hauptamtliche Schulinspektorinnen und -inspektoren verfügt, haben Vakanzen einen erheblichen Einfluss auf die Anzahl der praktisch durchführbaren Inspektionen pro Jahr) zwischen einzelnen Jahren variieren kann.

Schulleitung etc.) vorliegen, was zur Folge hat, dass die Zufälligkeit für diesen Teil der Schul-Gesamtstichprobe nicht sichergestellt werden kann. Entsprechend sind nur die jährlichen Kernstichproben als Analysegrundlage sinnvoll nutzbar.

4.1 Methodisches Vorgehen

Grundsätzlich könnte man daher die Schulen der jährlichen Kernstichprobe mit zufällig gezogenen Schulen vergleichen, die nicht inspiziert wurden. Da jedoch unbeobachtete Zuweisungsmechanismen von Schülerinnen und Schülern zu Schulen – die ggf. zeitdynamische Effekte auf die gemessenen Schülerleistungen nach sich ziehen – in einem solchen Ansatz zu Fehlschätzungen führen können (vgl. Baumert, Becker, Neumann & Nikolova, 2009), ist es sinnvoll, darüber hinaus die oben dargestellten Verfahren einzusetzen, die einen Umgang mit dieser Problematik ermöglichen (vgl. Angrist & Pischke, 2009).

Für das weitere Vorgehen wird der Difference-in-Differences-Ansatz genutzt, da Schülerleistungsdaten zu mehreren Messzeitpunkten miteinander verglichen werden sollen. Im Rahmen des Difference-in-Differences-Ansatzes wird der kausale Effekt einer Intervention geschätzt, indem der Trend innerhalb der Gruppe der Schulen mit Schulinspektion mit dem Trend innerhalb der Gruppe der Schulen ohne Schulinspektion verglichen wird. Der Trend der nicht-inspizierten Schulen wird im Rahmen dieses Ansatzes entsprechend als kontrafaktischer „was-wäre-wenn"-Trend verwendet. Dabei wird der kausale Effekt der Schulinspektion als Differenz der Differenzen der Mittelwerte zu den jeweiligen Messzeitpunkten geschätzt.

Bezeichnet man z. B. den Mittelwert der Schülerleistungen der inspizierten Schulen *vor* der Schulinspektion als TB (Treatmentgruppe – before) und den Mittelwert der Schülerleistungen der inspizierten Schulen *nach* der Schulinspektion mit TA (Treatmentgruppe – after) sowie entsprechend den Mittelwert der Schülerleistungen der nicht-inspizierten Schulen *vor* dem Zeitpunkt der Schulinspektion als CB (Kontrollgruppe – before) und den Mittelwert der Schülerleistungen der nicht-inspizierten Schulen *nach* dem Zeitpunkt der Schulinspektion mit CA (Kontrollgruppe – after), also:

	Treatment Group	**Control Group**
Before	TB	CB
After	TA	CA

Tab. 4: Darstellung einer beispielhaften Difference-in-Differences-Vierfeldertafel

So kann man den kausalen Effekt der Schulinspektion Δ_{DiD} als Differenz der Differenzen der Mittelwerte schätzen: „Difference-in-Differences" = (TA − TB) − (CA − CB)

(vgl. Morgan & Winship, 2007). Sinnvoll ist es häufig, diesen Differenzwert mithilfe eines linear-gepoolten Regressionsmodells und unter Ermittlung robuster Standardfehler, die es ermöglichen, die statistische Signifikanz des Effektes zu prüfen, zu berechnen (vgl. z. B. Helbig, Baier & Kroth, 2012).

Praktisch wurden in der Software SPSS auf Schulebene aggregierte Daten mithilfe einer Regression, analog der Vorschläge von Buckley und Shang (2003), analysiert und robuste Standardfehler ermittelt, die der Mehrebenenstruktur der Daten Rechnung tragen. Damit eine solche Analyse jedoch nicht zu ökologischen Fehlschlüssen führt, ist es notwendig, vorab zu prüfen, ob der Mittelwert der Schülerleistungen an einer Schule ein zuverlässiges Maß für die Leistung dieser Schülerschaft ist. Dies ermöglicht die Berechnung von Intraklassenkorrelationen (ICC2), die sich mittels der Variation in den Schülerleistungen zwischen und innerhalb von Schulen berechnen lassen und Werte zwischen 0 und 1 einnehmen können (vgl. Bliese, 2000). In der Regel werden Werte von 0.7 als Mindestmaß für die Aggregation von Individualdaten auf höheren Ebenen eingefordert (vgl. Lüdtke, Trautwein, Kunter & Baumert, 2006).

Die Bildung der Kontrollgruppen erfolgt in den Analysen im Sinne eines *Random-Program-Start*-Ansatzes (vgl. Sianesi, 2004). Anders als im klassischen Ansatz der Kontrollgruppenbildung, bei dem die Kontrollgruppe aus allen vorliegenden Fällen ohne Treatment ermittelt wird, wird in einem solchen Vorgehen berücksichtigt, dass die Wahl der Treatmentgruppe einem zeitdynamischen, stochastischen Prozess unterliegt und alle Schulen der Gesamtstichprobe irgendwann einmal einen Besuch durch die Schulinspektion erhalten, dies jedoch zu unterschiedlichen Messzeitpunkten. Dieses Vorgehen wurde gewählt, da die Identifizierung kausaler Effekte nur dann verlässlich möglich ist, wenn die Annahme der konditionalen Unabhängigkeit *(conditional independence assumption, CIA)* von Treatmentstatus und Ergebnisvariablen gegeben ist (vgl. Rubin, 1977). Schulen müssen entsprechend unabhängig von ihren jeweiligen Schülerleistungen entweder der Treatment- oder der Kontrollgruppe zugeordnet werden können. Diesem Problem wird zwar einerseits mithilfe der jährlichen Stichprobenziehung der Schulinspektion Hamburg begegnet, jedoch ergibt sich andererseits durch die zeitlich versetzte Inspektion von Schulen das Problem, dass Schulen im zeitlichen Verlauf eine zunehmend höhere Wahrscheinlichkeit haben, Teil der jährlichen Inspektionsstichprobe zu werden. Die Antizipation dieser Tatsache könnte beispielsweise Schulen, die in den ersten Jahren der Inspektion noch nicht inspiziert wurden, dazu bringen, dass sie mit zunehmender Zeitdauer in Erwartung eines Schulinspektionsbesuches präventiv Maßnahmen einleiten, die zu Steigerungen der Schülerleistungen führen, welche jedoch grundsätzlich unabhängig davon sind, ob eine Inspektion tatsächlich stattgefunden hat oder nicht. In einem solchen Fall hinge der Treatmentstatus einer Schule mit zukünftigen, potenziellen Ergebnissen zusammen, was eine erneute Verletzung der *CIA* nach sich ziehen und zu fehlerbehafteten Ergebnissen in der Kausalanalyse führen würde (vgl. Sianesi, 2004).

Für die vorliegenden Analysen werden daher die Lernentwicklungen und Leistungstrends an zufällig gezogenen Schulen miteinander verglichen, die zu unterschiedlichen Zeitpunkten durch die Schulinspektion Hamburg evaluiert wurden. Relevant für den

Treatmentstatus ist demnach nicht, ob eine Schule inspiziert wurde oder nicht, sondern ob zu einem bestimmten Zeitpunkt oder später.

4.2 Studie 1: Analyse von Trenddaten des Hamburger Zentralabiturs

Abschlussprüfungen mit zentralen Elementen werden in der Hansestadt Hamburg seit dem Schuljahr 2004/2005 durchgeführt. Dabei findet eine zentrale Aufgabenstellung seit dem Jahr 2010 nur in den Kernfächern Deutsch, Mathematik und fortgeführte Fremdsprache (unterteilt in die Sprachen: Englisch, Französisch, Latein, Polnisch, Russisch, Spanisch, Türkisch) und auch dort nur für die schriftlichen Abiturarbeiten statt. In allen anderen Fächern existiert eine dezentrale Aufgabenstellung. Die Abituraufgaben werden jährlich von Lehrkräften entworfen und seitens einer Kommission der Hamburger Behörde für Schule und Berufsbildung geprüft und ausgewählt. Die Aufgaben werden dabei so gestellt,

> „dass sie nicht nur den Unterricht eines Halbjahres berücksichtigen und dass sie Leistungen in den folgenden drei Anforderungsbereichen ermöglichen:
>
> - Anforderungsbereich I umfasst das Wiedergeben von Sachverhalten und Kenntnissen im gelernten Zusammenhang sowie das Beschreiben und Anwenden geübter Arbeitstechniken und Verfahren in einem wiederholenden Zusammenhang.
> - Anforderungsbereich II umfasst das selbständige Auswählen, Anordnen, Verarbeiten und Darstellen bekannter Sachverhalte unter vorgegebenen Gesichtspunkten in einem durch Übung bekannten Zusammenhang und das selbständige Übertragen und Anwenden des Gelernten auf vergleichbare neue Zusammenhänge und Sachverhalte.
> - Anforderungsbereich III umfasst das zielgerichtete Verarbeiten komplexer Sachverhalte mit dem Ziel, zu selbständigen Lösungen, Gestaltungen oder Deutungen, Folgerungen, Begründungen und Wertungen zu gelangen. Dabei wählen die Schülerinnen und Schüler aus den gelernten Arbeitstechniken und Verfahren die zur Bewältigung der Aufgabe geeigneten selbständig aus, wenden sie in einer neuen Problemstellung an und beurteilen das eigene Vorgehen kritisch."
> (vgl. Behörde für Schule und Berufsbildung, 2011, S. 3–4)

Die rechtlichen Regelungen zur Durchführung der zentralen Aufgabenstellung in Abiturarbeiten an Schulen in Hamburg wurden dabei erstmalig in der Ausbildungs- und Prüfungsordnung zum Erwerb der Allgemeinen Hochschulreife (APO-AH) vom 25. März 2008 in der Änderungsfassung vom 18. März 2009 zusammengefasst, sodass Standards für die Punktevergabe in zentralen Abiturarbeiten erstmals im Jahr 2010 angewandt wurden.

Diesen Standards zufolge bewertet die an der Schule für das Fach zuständige Lehrkraft jede Arbeit mit einer Punktzahl von 0 bis 15. Anschließend wird jede Arbeit von

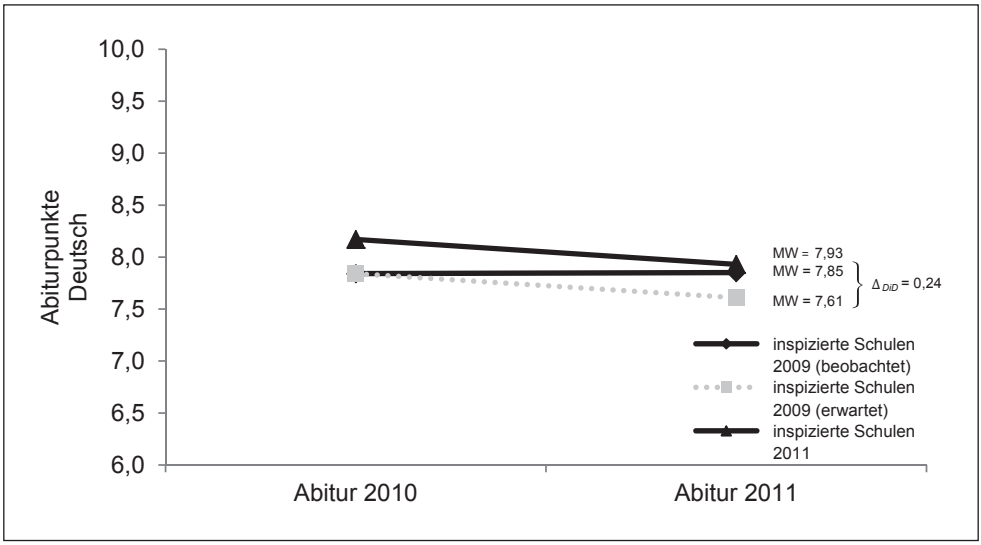
einer zweiten Fachlehrkraft durchgesehen, die sich entweder der ersten Bewertung anschließt oder ein ergänzendes Gutachten mit Bewertung anfertigt. Abschließend legt die oder der Vorsitzende des Prüfungsausschusses die endgültige Punktzahl der Abiturarbeit fest. Beträgt die Differenz der im Erstgutachten und im ergänzenden Gutachten erteilten Punktzahlen nicht mehr als drei Punkte, bildet sie oder er den Mittelwert beider Punktzahlen. Liegt der Mittelwert zwischen zwei Punktzahlen, wird zur nächsten vollen Punktzahl aufgerundet. Darüber hinaus kann in begründeten Fällen (bei einem Punktunterschied von mehr als drei Punkten zwischen Erst- und Zweitgutachten) ein Drittgutachten veranlasst werden.

Aufgrund der erst 2010 eingeführten Durchführungsstandards und der damit verbundenen Vergleichbarkeit von Abiturnoten können für die weiteren Analysen nur Abiturdaten aus den Jahren 2010 und 2011[2] genutzt werden. Im Jahr 2010 wurden von 16 101 Zentralabiturarbeiten 39 Prozent im Fach Deutsch, 27 Prozent im Fach Mathematik und 34 Prozent in den oben genannten Fremdsprachen geschrieben. Im Jahr 2011 entfielen, bei 13 209 auswertbaren Zentralabiturarbeiten, 32 Prozent auf das Fach Deutsch, 20 Prozent auf das Fach Mathematik und 48 Prozent auf die fortgeführten Fremdsprachen. Um systematische Unterschiede zwischen Schulen durch mögliche Unterschiede in den Aufgabenschwierigkeiten zwischen einzelnen Fremdsprachen zu vermeiden, werden im weiteren Verlauf jedoch nur die Fächer Deutsch und Mathematik für die Difference-in-Differences-Analysen genutzt.

Betrachtet wird daher der diesbezügliche Fächertrend im Zentralabitur der Jahre 2010 und 2011 für Sekundarschulen, die im Kalenderjahr vor dem ersten Messzeitpunkt inspiziert wurden, verglichen mit dem Trend an Schulen, die im Kalenderjahr des zweiten Messzeitpunktes inspiziert wurden. Schulen, die mehr als ein Kalenderjahr vor dem ersten Messzeitpunkt inspiziert wurden, wurden aus der Analyse ebenso ausgeschlossen wie Schulen, die nach dem zweiten Messzeitpunkt oder noch nie durch die Schulinspektion Hamburg inspiziert wurden. Insgesamt können Daten für 21 Schulen (N der Abiturarbeiten im Fach Deutsch zu t0 = 1164 und zu t1 = 1226; N der Abiturarbeiten im Fach Mathematik zu t0 = 869 und zu t1 = 795) als Treatmentgruppe genutzt werden. Die Vergleichsgruppe hingegen umfasst 17 Schulen (N der Abiturarbeiten im Fach Deutsch zu t0 = 878 und zu t1 = 894; N der Abiturarbeiten im Fach Mathematik zu t0 = 523 und zu t1 = 611). Die Intraklassenkorrelationen lagen dabei ausreichend hoch, sodass eine Analyse der Abiturleistungen auf Schulebene zulässig ist (ICC2 im Fach Deutsch zu t0 = 0.802 und zu t1 = 0.844, ICC2 im Fach Mathematik zu t0 = 0.840 und zu t1 = 0.859).

Wie den Abbildungen 2 und 3 zu entnehmen ist, lassen sich sowohl im Zentralabitur für das Fach Deutsch als auch für das Fach Mathematik leichte Treatmenteffekte nachweisen. Im Fach Deutsch liegen die Abiturnoten an inspizierten Schulen im Jahr 2011 statistisch nachweisbar um 0.24 Punkte (Δ_{DiD}) höher, als zu erwarten gewesen wäre ($p < 0.001$). Auffällig ist dabei, dass der Trend in den Abiturarbeiten an inspizierten Schulen von 2010 zu 2011 entgegen dem allgemeinen Trend nicht negativ ausgeprägt ist, sondern die Schülerschaft an den inspizierten Schulen das Niveau im Deutsch-

2 Abiturnoten für das Jahr 2012 lagen zum Zeitpunkt der Analyse noch nicht vor.



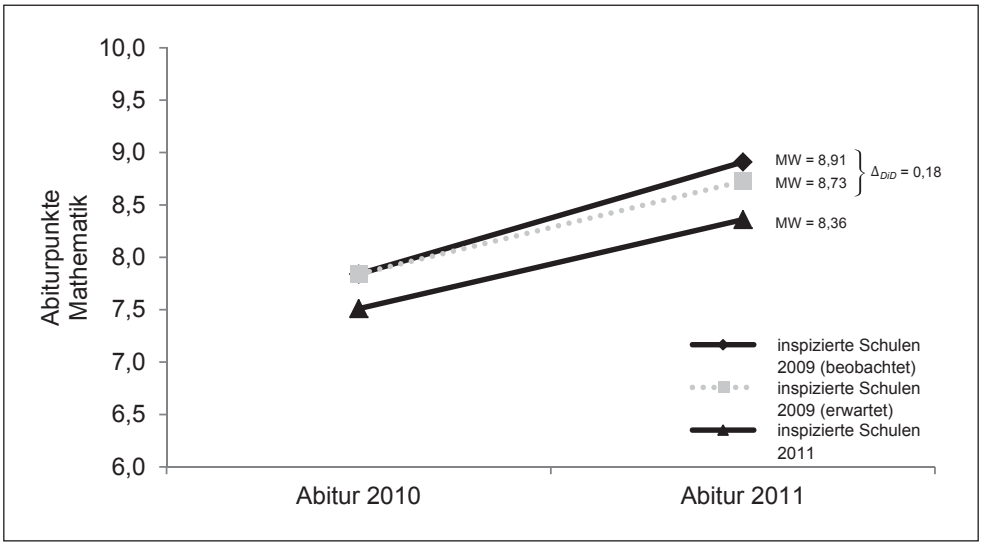


Abb. 2 und 3: Trends im Zentralabitur Deutsch (Abb. 2) und Mathematik (Abb. 3) für die Treatment- (beobachtet und erwartet) und Vergleichsgruppe (beobachtet)

Zentralabitur halten konnte. Für das Fach Mathematik lässt sich Ähnliches, wenn auch nicht im vergleichbaren Ausmaß, feststellen: Die mittlere Abiturleistung an inspizierten Schulen liegt im Jahr 2011 um 0.18 Punkte (Δ_{DiD}) höher, als ohne Intervention zu erwarten gewesen wäre ($p < 0.050$).

4.3 Studie 2: Analyse von Längsschnittdaten der Hamburger KESS-Studie[3]

Seit dem Jahr 2003 werden in Hamburg die Kompetenzen und Einstellungen von Schülerinnen und Schülern flächendeckend im Rahmen der längsschnittlich angelegten KESS-Studie (vgl. Bos, Bonsen & Gröhlich, 2009; Bos & Gröhlich, 2010; Bos & Pietsch, 2006; Vieluf, Ivanov & Nikolova, 2011) erhoben. Ausgehend von einer Basiserhebung zum Ende der Grundschulzeit werden im Rahmen der Studie ca. alle zwei Jahre die Kompetenzen einer vollständigen Schülerkohorte gemessen und Lernentwicklungen berichtet. Die KESS-Untersuchungen wurden von der Hamburger Behörde für Bildung und Sport in Auftrag gegeben und in wechselnden Kooperationsverbünden durchgeführt. Zu den Projektpartnern gehören die Universität Hamburg, das in Dortmund ansässige Institut für Schulentwicklungsforschung (IFS) sowie das Hamburger Landesinstitut für Lehrerbildung und Schulentwicklung.

Die einzelnen Datenerhebungen wurden zu folgenden Messzeitpunkten durchgeführt: im Juni 2003 am Ende der Grundschulzeit, im September 2005 zu Beginn der Jahrgangsstufe 7, im Juni 2007 am Ende der Jahrgangsstufe 8, im Juni bzw. September 2009 am Ende der Sekundarstufe I bzw. zu Beginn der gymnasialen Oberstufe, im Mai 2011 am Ende der Studienstufe der zweijährigen Oberstufe des achtstufigen Gymnasiums und im Mai 2012 am Ende der Studienstufe der dreijährigen Oberstufe an Gesamtschulen, Aufbaugymnasien und Beruflichen Gymnasien.[4] Die Testdurchführung wurde zu allen Messzeitpunkten durch speziell geschulte externe Testleiterinnen und Testleiter realisiert. Die Teilnahme an den Leistungstests war für alle Schülerinnen und Schüler der Testkohorte zu allen Messzeitpunkten der Studie verpflichtend, sodass in der Regel Teilnahmequoten von über 95 Prozent[5] erzielt wurden.

Die Testinstrumente der KESS-Studie orientieren sich an dem für internationale Schulvergleichsuntersuchungen gängigen Literacy-Ansatz und beinhalten neben eigens entwickelten Aufgaben vor allem Items aus Studien wie dem *Programme for International Student Assessment* (PISA), der *Trends in International Mathematics and Science Study* (TIMSS) und den *Internationalen Grundschul-Lese-Untersuchungen* (IGLU). Die Kompetenztests erfassen dabei verschiedene Domänen, wobei jedoch nur die Do-

3 Die Autoren danken Herrn Stanislav Ivanov vom Hamburger Institut für Bildungsmonitoring und Qualitätsentwicklung für die Aufbereitung und Bereitstellung der Daten aus der KESS-Untersuchung.

4 Zum Zeitpunkt der hier durchgeführten Untersuchung lagen Daten bis einschließlich Mai 2011 zur Re-Analyse vor.

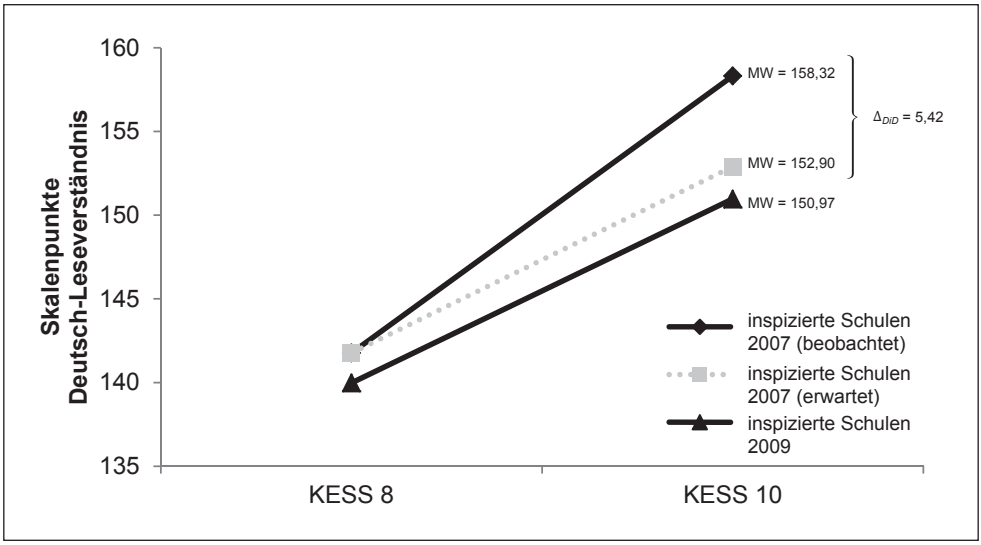
5 Eine Ausnahme bildet die Erhebung KESS 10/11, in der die Teilnahmequoten – trotz Teilnahmepflicht – nur im Bereich von 83 bis 86 Prozent lagen.

mänen Leseverstehen und Mathematik zu allen Messzeitpunkten von allen Schülerinnen und Schülern bearbeitet wurden. Die Tests der KESS-Untersuchung waren als rotiertes Multi-Matrix-Design angelegt, sodass einzelne Schülerinnen und Schüler nur jeweils eine Teilmenge von Aufgaben der einzelnen Tests bearbeiten mussten. Entsprechend wurde für die Modellierung von Leistungswerten die Item-Response-Theorie (IRT) genutzt, in der Aufgabenschwierigkeiten und Personenfähigkeiten auf einer gemeinsamen Metrik abgebildet werden können und Marginal-Maximum-Likelihood-Schätzer den Umgang mit geplantem Datenausfall erlauben. Um Längsschnittanalysen zu ermöglichen, wurden einzelne Aufgaben, sogenannte Anker-Items, zu verschiedenen Messzeitpunkten eingesetzt. Die Skalen, auf denen die Schülerleistungen berichtet werden, haben einen Ausgangsmittelwert (in KESS 4) von 100 und eine Standardabweichung von 30 Punkten.

Für die kommenden Analysen werden KESS-Daten aus den Erhebungen KESS 8 (Erhebung im Juni 2007 am Ende der Jahrgangsstufe 8) und KESS 10 (Erhebung im Juni 2009 am Ende der Sekundarstufe I) genutzt. Analog zur Analyse der Zentralabiturdaten werden auch hier nur Daten aus dem Bereich Deutsch (Leseverständnis) und Mathematik betrachtet, da diese Testdomänen durch alle Schülerinnen und Schüler an allen teilnehmenden Schulen flächendeckend bearbeitet wurden. Insgesamt liegen zum Messzeitpunkt t0 (KESS 8) für 14 180 Schülerinnen und Schüler Daten zu diesen beiden Testdomänen vor. Zum Messzeitpunkt t1 (KESS 10) liegen vergleichbare Daten für insgesamt 13 328 Schülerinnen und Schüler vor.

Mithilfe des Difference-in-Differences-Ansatzes wird nachfolgend die Lernentwicklung in den beiden Testdomänen Leseverständnis und Mathematik von Schülerinnen und Schülern an Schulen, die im Jahr 2007 inspiziert wurden, mit der Lernentwicklung von Schülerinnen und Schülern an Schulen, die im Jahr 2009 inspiziert wurden, verglichen. Insgesamt können Daten aus 11 Schulen mit Inspektionsjahr 2007 (N der Schülertests im Leseverständnis zu t0 und t1 = 681; N der Schülertests in Mathematik zu t0 und t1 = 683) und Daten aus 23 Schulen mit Inspektionsjahr 2009 (N der Schülertests im Leseverständnis zu t0 und t1 = 1209; N der Schülertests in Mathematik zu t0 und t1 = 1207) berücksichtigt werden. Die Intraklassenkorrelationen lagen dabei sehr hoch, sodass eine Analyse der Leistungen auf Schulebene zulässig ist (ICC2 im Test Mathematik zu t0 = 0.983 und zu t1 = 0.968, ICC2 im Test Leseverständnis zu t0 = 0.976 und zu t1 = 0.968).

Wie die Abbildungen 4 und 5 zeigen, lassen sich statistisch belastbare Treatmenteffekte in den KESS-Längsschnittdaten nur für die Leseleistungen der Schülerinnen und Schüler nachweisen ($p < 0.001$). Die Schülerleistungen im Leseverständnis in KESS 10 liegen an den im Jahr 2007 inspizierten Schulen rund 5.4 Punkte (Δ_{DiD}) oder fast 20 Prozent einer Standardabweichung auf der KESS-Metrik über dem Erwartungswert. Da der mittlere Lernzuwachs im Lesen von KESS 8 zu KESS 10 rund 13.4 Skalenpunkte beträgt, beläuft sich der Inspektionseffekt im Lesen für Schülerinnen und Schüler in diesen Schulstufen auf zusätzlich ca. 80 Prozent eines Lernjahres. Für die Testleistung in Mathematik lässt sich jedoch ein ähnlicher Treatmenteffekt nicht feststellen ($p > 0.100$). Die beobachtete Testleistung entspricht hier der erwarteten Testleistung ($\Delta_{DiD} = 0.31$).



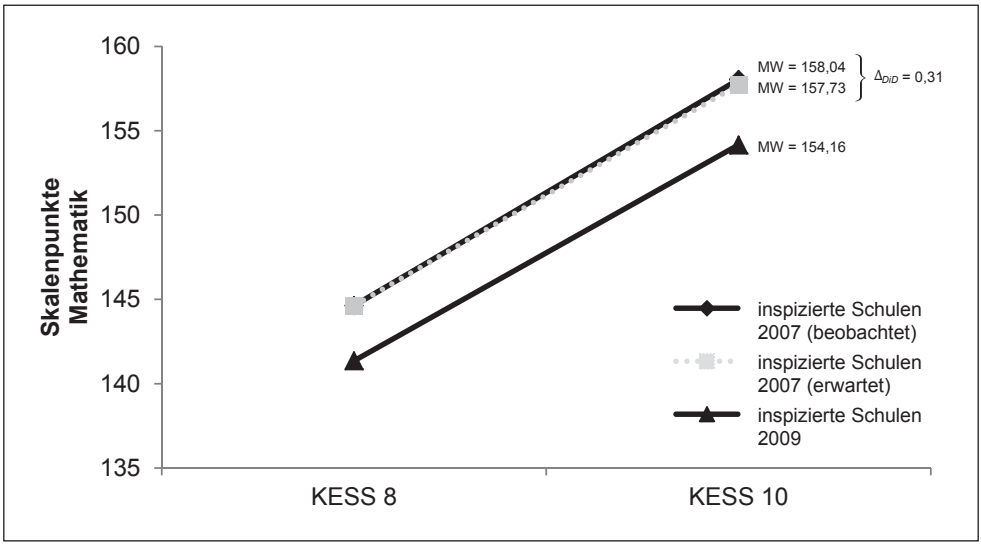


Abb. 4 und 5: Lernentwicklungen von Schülerinnen und Schülern im Deutsch-Leseverständnis (Abb. 4) und in Mathematik (Abb. 5) für die Treatment- (beobachtet und erwartet) und die Vergleichsgruppe (beobachtet)

Schulinspektion, verstanden als Intervention auf Schulebene, hat demnach keinen Einfluss auf die Lernentwicklung von Schülerinnen und Schülern im Fach Mathematik.

5. Zusammenfassung und Diskussion

Zur Wirksamkeit von Schulinspektionen auf Schülerleistungen liegt bislang kaum verlässliche empirische Evidenz vor, da diesbezügliche Studien häufig methodische Schwächen aufweisen. Daher wurden im Beitrag Verfahren, die empirisch robuste Kausalanalysen ermöglichen, vorgestellt und deren Einsatzmöglichkeiten am Beispiel der Schulinspektion Hamburg exemplarisch demonstriert. Die Befunde zeigen, dass bei Einsatz maßgeschneiderter kausalanalytischer Verfahren sowohl Effekte auf Lernzuwächse als auch Leistungstrends von Schülerinnen und Schülern in Hamburg nachgewiesen werden können. Damit widersprechen die vorgelegten Analysen klar den meisten bislang vorliegenden und insbesondere den in England generierten Befunden zur Wirksamkeit von Schulinspektion auf Schülerleistungen und decken sich eher mit den Analyseergebnissen von Luginbuhl et al. (2009), wobei die hier berichteten Ergebnisse noch deutlich positiver ausfallen als die Ergebnisse der niederländischen Studie.

Da die niederländische Schulinspektion zum Zeitpunkt der vorgelegten Analysen ebenso wie die Schulinspektion Hamburg dem Paradigma folgte, Schulentwicklung durch Bereitstellung von Informationen zu stimulieren, verdichten sich somit die empirischen Hinweise darauf, dass Schulinspektion mit Blick auf Schülerleistungen, zumindest unter diesem Paradigma, nicht schadet und ggf. sogar positive Effekte nach sich ziehen kann. Für das in England etablierte Wettbewerbsmodell lassen sich solche Aussagen hingegen nicht zuverlässig treffen, da weiterhin unklar ist, ob die dort berichteten Effekte Programm-, Theorie- oder Methodenfehlern geschuldet sind.

Bei den hier dargestellten Ergebnissen ist jedoch auffällig, dass in beiden Studien der Einfluss der Inspektion auf die Leistung von Schülerinnen und Schülern im Fach Deutsch resp. im Leseverständnis höher ausfiel als auf die Leistungen im Fach Mathematik. Da die durchgeführten Studien sich wechselseitig validieren, scheint es so, dass die Hamburger Schulinspektion differenzielle Effekte nach sich zieht und entsprechend nicht auf alle Schülerleistungen gleichermaßen wirkt. Diesbezüglich werfen die Ergebnisse die Frage auf, über welche Mechanismen und Prozesse Schulinspektion Einfluss ausübt. Weitere Untersuchungen deuten diesbezüglich darauf hin, dass diese Faktoren vor allem im Bereich der innerschulischen Informationsverarbeitung, aber auch in den Kontextbedingungen, unter denen Schulen mit den Ergebnissen der Schulinspektion umgehen müssen, zu suchen sind (vgl. Pietsch, 2011a).

Abschließend muss jedoch auch auf die Einschränkungen der vorliegenden Studie hingewiesen werden: So kann erstens nicht geklärt werden, wie nachhaltig die beobachteten Effekte sind, da jeweils nur zwei Messzeitpunkte für die Analysen vorlagen. Zweitens ist nicht nachweisbar, dass allein die Einführung der Schulinspektion oder die Ankündigung, dass eine Inspektion an der jeweiligen Schule durchgeführt wird, bereits zu Veränderungen geführt hat, Rückmeldungen somit keine Rolle als Grundlage für die

Schulentwicklung spielen. Und drittens ist es nicht möglich zu zeigen, ob eine andere Form der schulbezogenen Intervention – z. B. eine begleitete Selbstevaluation – nicht auch zu vergleichbaren Effekten geführt hätte. Mit Blick auf die genutzten Daten ist darüber hinaus die Analyse der KESS-Daten als die deutlich stärkere zu werten, da hier einerseits Paneldaten verwendet werden und andererseits – aufgrund der externen Testdurchführung und -auswertung – ausgeschlossen werden kann, dass Inspektionseffekte sich z. B. auf die Praxis der Leistungsbewertung an den Schulen auswirken, sich jedoch nicht in den Schülerleistungen selbst niedergeschlagen haben.

Vor diesem Hintergrund ergeben sich aus unserer Sicht zwei zentrale Forschungsdesiderata: (1) Weitere Untersuchungen müssen die vorgelegten Befunde mithilfe der vorgestellten kausalanalytischen Verfahren replizieren, um auf diesem Wege weitere belastbare empirische Evidenz zu erzeugen, die es ermöglicht, Aussagen zur Wirksamkeit von Schulinspektionen – möglichst anhand mehrerer Messzeitpunkte – zu generalisieren sowie zu validieren, und (2) es müssen elaborierte logische Modelle entwickelt werden, die es ermöglichen, den Evaluationsgegenstand angemessen auszuleuchten und neben reinen Blackbox- auch programmtheoretisch-orientierte Evaluationen zu ermöglichen, die es z. B. gestatten, differenzielle Wirkungsweisen von Schulinspektionen oder die Modellierung nicht-linearer Wirkungsmechanismen in den Blick zu nehmen.

Literatur

Allen, R., & Burgess, S. (2012). *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England.* Bristol: Centre for Market and Public Organisation.

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks: Sage.

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter & K. Maag Merki (Hrsg.), *Neue Steuerung im Schulwesen* (S. 219–254). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Baumert, J., Becker, M., Neumann, M., & Nikolova, R. (2009). Frühübergang in ein grundständiges Gymnasium – Übergang in ein privilegiertes Entwicklungsmilieu? Ein Vergleich von Regressionsanalyse und Propensity Score Matching. *Zeitschrift für Erziehungswissenschaft, 12*(2), 189–215.

Behörde für Schule und Berufsbildung (2011). *Abitur 2011: Regelungen für die zentralen schriftlichen Prüfungsaufgaben.* Hamburg: Behörde für Schule und Berufsbildung.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Hrsg.), *Multilevel theory, research, and methods in organizations* (S. 349–381). San Francisco: Jossey-Bass.

Böttcher, W., & Kotthoff, H.-G. (2010). Neue Formen der Schulinspektion: Wirkungshoffnungen und Wirksamkeit im Spiegel empirischer Bildungsforschung. In H. Altrichter & K. Maag Merki (Hrsg.), *Neue Steuerung im Schulwesen* (S. 295–325). Wiesbaden: VS Verlag für Sozialwissenschaften.

Böttger-Beer, M., & Koch, E. (2008). Externe Schulinspektion in Sachsen – ein Dialog zwischen Wissenschaft und Praxis. In W. Böttcher, W. Bos, H. Döbert & H. G. Holtappels (Hrsg.), *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive* (S. 253–265). Münster: Waxmann.

Bos, W., Bonsen, M., & Gröhlich, C. (2009). *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7.* Münster: Waxmann.

Bos, W., & Gröhlich, C. (2010). *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Jahrgangsstufe 8.* Münster: Waxmann.

Bos, W., & Pietsch, M. (2006). *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen.* Münster: Waxmann.

Buckley, J., & Shang, Y. (2003). Estimating Policy and Program Effects with Observational Data: The „Differences-in-Differences" Estimator. *Practical Assessment, Research & Evaluation, 8*(24). http://PAREonline.net/getvn.asp?v=8&n=24 [23. 11. 2012].

Budig, M. J., & England, P. (2001). The wage penalty for motherhood. *American Sociological Review, 66*, 204–225.

Cousins, J. B., & Leithwood, K. A. (1993). Enhancing knowledge utilization as a strategy for school improvement. *Knowledge: Creation, Diffusion, Utilization, 14*(3), 305–333.

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

Cullingford, S., & Daniels, S. (1999). Effects of OFSTED inspections on school performance. In C. Cullingford (Hrsg.), *An inspector calls: OFSTED and its effects on school standards* (S. 59–69). London: Kogan Page.

de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspection and accountability in education: An overview of empirical studies. *Oxford Review of Education, 33*(3), 379–396.

Ehren, M. C. M., & Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *British Journal of Educational Studies, 54*(1), 51–72.

Gangl, M., & DiPrete, T. A. (2004). Kausalanalyse durch Matchingverfahren. In A. Diekmann (Hrsg.), *Methoden der Sozialforschung. 44. Sonderheft der Kölner Zeitschrift für Soziologie und Sozialpsychologie* (S. 396–420). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gärtner, H., & Pant, H. A. (2011). Validierungsstrategien für Verfahren und Ergebnisse von Schulinspektion. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz in empirischer Sicht* (S. 9–32). Münster: Waxmann.

Helbig, M., Baier, T., & Kroth, A. (2012). Die Auswirkung von Studiengebühren auf die Studierneigung in Deutschland. Evidenz aus einem natürlichen Experiment auf Basis der HIS-Studienberechtigtenbefragung. *Zeitschrift für Soziologie, 41*(3), 227–246.

Helmke, A., & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127–151). Bern: Hep.

Husfeldt, V. (2011). Wirkungen und Wirksamkeit der externen Schulevaluation: Überblick und Stand der Forschung. *Zeitschrift für Erziehungswissenschaft, 14*(2), 259–283.

Hyyryläinen, E., & Viinamäki, O.-P. (2008). The implications of the rationality of decision-makers on the utilization of evaluation findings. *International Journal of Public Administration, 31*(10), 1223–1240.

Kluger, A. N., & DeNisi, A. S. (1996). The Effects of Feedback Interventions on Performance: Historical Review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 199*(2), 254–284.

Leeuw, F. L., & van Thiel, S. (2002). The performance paradox in the public sector. *Public Performance & Management Review, 25*(3), 267–281.

Legewie, J. (2012). Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 64*(1), 123–153.

Leist, S., Pietsch, M., & Vaccaro, E. (2009). Grundlagen der Berichterstattung. In Institut für Bildungsmonitoring (Hrsg.), *Jahresbericht der Schulinspektion Hamburg 2008* (S. 6–14). Hamburg: Institut für Bildungsmonitoring.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Analyse von Lernumwelten. Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. *Zeitschrift für Pädagogische Psychologie, 20*(1/2), 85–96.

Luginbuhl, R., Webbink, D., & de Wolf, I. F. (2009). Do inspections improve primary school performance? *Educational Evaluation and Policy Analysis, 31*(3), 221–237.

Maier, U. (2008). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54*(1), 95–117.

Matthews, P., & Sammons, P. (2004). *Improvement through inspection: An evaluation of the impact of OFSTED's work.* London: OFSTED.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles in social research.* Cambridge: Cambridge University Press.

Pietsch, M. (2011a). *Nutzung und Nützlichkeit der Schulinspektion. Befunde der Hamburger Schulleitungsbefragung.* Hamburg: Institut für Bildungsmonitoring.

Pietsch, M. (2011b). Fehlende Daten bei Unterrichtsbeobachtungen: Eine Sensitivitätsanalyse anhand von Daten der Schulinspektion Hamburg. *Empirische Pädagogik, 25*(1), 47–87.

Pietsch, M., Janke, N., & Mohr, I. (2013). Führt Schulinspektion wirklich nicht zu besseren Schülerleistungen? Eine Einschätzung zur Belastbarkeit vorliegender Wirksamkeitsstudien aus programmtheoretischer Perspektive. In K. Schwippert, M. Bonsen & N. Berkemeyer (Hrsg.), *Schul- und Bildungsforschung – Diskussionen, Befunde und Perspektiven* (S. 167–185). Münster: Waxmann.

Pietsch, M., Schnack, J., & Schulze, P. (2009). Unterricht zielgerichtet entwickeln: Die Schulinspektion Hamburg entwickelt ein Stufenmodell für die Qualität von Unterricht. *Pädagogik, 61*(2), 38–43.

Pietsch, M., Schulze, P., Schnack, J., & Krause, M. (2011). Elaborierte Rückmeldungen zur Qualität von Unterricht. Über empirisch abgesicherte Bezugsnormen als Grundlage für die Weiterentwicklung von Unterricht und Schule. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektionen in Deutschland – Eine Zwischenbilanz aus empirischer Sicht* (S. 193–216). Münster: Waxmann.

Reezigt, G. J., & Creemers, B. P. M. (2005). A comprehensive framework for effective school improvement. *School Effectiveness and School Improvement, 16*(4), 407–424.

Rosenthal, L. (2004). Do school inspections improve school quality? OFSTED inspections and school examination results in the UK. *Economics of Education Review, 23*(2), 143–151.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of Covariate. *Journal of Educational Studies, 2*(1), 1–26.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement, 1*(1), 61–80.

Scriven, M. (1994). The fine line between evaluation and explanation. *Evaluation Practice, 15*(1), 75–77.

Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do OFSTED inspections of secondary education make a difference to GCSE results? *British Educational Research Journal, 29*(1), 63–75.

Sianesi, B. (2004). An evaluation of the Swedish system of active labor market programs in the 1990's. *Review of Economics and Statistics, 86*(1), 133–155.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1-2), 305–353.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration, 18*, 277–310.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation Theory, Models, & Applications.* San Francisco: Jossey-Bass.
Tarter, C. J., & Hoy, W. K. (1998). Toward a contingency theory of decision making. *Journal of Educational Administration, 36*(3), 212–228.
van Ackeren, I., & Klemm, K. (2009). *Entstehung, Struktur und Steuerung des deutschen Schulsystems.* Wiesbaden: VS Verlag für Sozialwissenschaften.
Vieluf, U., Ivanov, S., & Nikolova, R. (2011). *KESS 10/11 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe.* Münster: Waxmann.
Visscher, A. J., & Coe, R. (2003). School Performance Feedback Systems: Conceptualisation, Analysis and Reflection. *School Effectiveness and School Improvement, 14*(3), 321–349.
White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation, 16*(2), 153–164.
Wilcox, B., & Gray, J. (1996). *Inspecting schools: Holding schools to account and helping schools to improve.* Buckingham: Open University Press.

Abstract: School inspectorates are meant to improve student performance on the level of both the individual school and the school system. Whereas, in this context, there are no empirical findings on the efficiency of school inspectorates in Germany, international studies show that, as a rule, school inspectorates do not succeed in bringing about an improvement in performance. However, these findings are usually not very reliable due to problems with the random samples chosen for the studies. The present contribution for the first time examines for the German context which effects on student performance are empirically verifiable, using the school inspectorate Hamburg as an example. The investigation is based on trend data on Hamburg's central school leaving exam and on longitudinal data provided by the study "Students' competencies and attitudes" (German abbreviation: KESS). Possible problems with random samples are explicitly taken into account in the analyses in order to be able to come up with empirically reliable findings on the effects of school inspections on student performance.

Keywords: Difference-in-Differences, Student Achievement, School Inspection, Selection Bias, Effectiveness

Anschrift des Autors/der Autorinnen

Dr. Marcus Pietsch, Leuphana-Universität Lüneburg, Scharnhorststraße 1,
21335 Lüneburg, Deutschland
E-Mail: pietsch@leuphana.de

Dr. Nike Janke, Landesinstitut für Schule Bremen, Am Weidedamm 20,
28215 Bremen, Deutschland
E-Mail: njanke@lis.bremen.de

Dr. Ingola Mohr, Landesinstitut für Schulentwicklung Baden-Württemberg,
Heilbronner Straße 172, 70191 Stuttgart, Deutschland
E-Mail: ingola.mohr@ls.kv.bwl.de