Evidenzbasierte Entwicklung von Unterricht und Schule im Kontext Neuer Steuerung

Dr. Marcus Pietsch

Juni 2018

Publikationsbasierte Habilitationsschrift

Eingereicht bei der

Fakultät Bildung an der

Leuphana Universität Lüneburg

Evidenzbasierte Entwicklung von Unterricht und Schule im Kontext Neuer Steuerung

1. Einleitung

Seit der Jahrtausendwende sollen Schulen in Deutschland zunehmend dezentralisiert, outputorientiert und evidenzbasiert gesteuert werden. Diese Idee der sogenannten Neuen Steuerung im Bildungssystem setzt dabei konkret auf zwei grundlegende Instrumente (Bellmann 2006): Auf der einen Seite steht die Outputsteuerung (im Englischen: standard-based reforms) durch die Setzung von Bildungsstandards und die externe Evaluation von Schulen und Schulleistungen. Auf der anderen Seite steht die Wettbewerbssteuerung von Schulen (im Englischen: choice policies), worunter die Etablierung von Quasi-Märkten im Bildungssystem durch Dezentralisierung, Schulautonomie und freie Schulwahl verstanden wird.

Schule und deren Kernelement, der Unterricht, werden dabei von ihren Ergebnissen bzw. Wirkungen (Output, Outcome und Impact) her gedacht: Eine gute Schule und ein guter Unterricht sind demnach dadurch gekennzeichnet, dass sie bei gleichen Eingangsbedingungen einen möglichst hohen Lernerfolg aufseiten von Schülerinnen und Schülern ermöglichen (Steffens & Bargel 2016). Damit schließt die Neue Steuerung im deutschen Bildungssystem wiederum an Annahmen und Befunde der School- und Teacher-Effectiveness-Forschung (Cortez, Gayle & Preiss 2006; Teddlie & Reynolds 2000) an, in deren Modellen Bildungserfolge und -erträge die Folge institutionell modifizierter Eingangsbedingungen sind, die unter spezifischen Kontextbedingungen erfolgen (Reynolds et al. 2000).

Zentraler Bestandteil der aktuellen Steuerungsstrategie im deutschen Bildungswesen ist dabei die Verlagerung von Entscheidungskompetenzen auf die Ebene der Einzelschule sowie die zunehmende Gewährung von Handlungs- und Gestaltungsspielräumen für innerschulische Akteure (Altrichter, Rürup & Schuchart 2016; Bonsen 2016; Fuchs 2009). Schulen bzw. schulische Akteure sind daher in zunehmendem Maße verantwortlich für die Qualitätssicherung und Qualitätsentwicklung der Einzelschule und in der Summe des gesamten Bildungswesens (Thiel & Thillmann 2012).

Um sicherzustellen, dass die intendierten Ergebnisziele erreicht werden, wurden in den deutschen Bundesländern wiederum Prüf- und Kontrollmechanismen eingeführt. Hierzu zählen insbesondere (1) Schulinspektionen und andere Verfahren externer Evaluation, (2) zentrale Abschlussprüfungen, (3) Bildungsstandards und, damit verbunden, (4) Lernstandserhebungen sowie (5) Schulprogramme (Fuchs 2009).

Dabei haben diese Mechanismen jedoch häufig eine Doppelfunktion: Auf der einen Seite sollen sie eine Kontrollfunktion ausüben, auf der anderen Seite aber auch Impulse für die evidenzbasierte Weiterentwicklung von Unterricht und Schule liefern (Böttcher & Keune 2012). Es wird daher davon

ausgegangen, dass schulische Akteure empirisch gewonnene Information systematisch und rational nutzen, um die Qualität von Schule und Unterricht und in der Folge die Lernerfolge von Schülerinnen und Schülern wissensbasiert zu optimieren (Altrichter, Moosbrugger & Zuber 2016).

Ob diese Annahmen jedoch zutreffen bzw. die gewählten Mechanismen den ihnen zugeschriebenen Ansprüchen gerecht werden können, ist meist unklar, denn, so Berkemeyer (2010, S. 87), das „*(bildungs-)politische System entscheidet derzeit ohne große Verzögerungen und somit nicht selten ohne über entsprechende wissenschaftliche Expertisen zu verfügen*". Dies liegt nicht zuletzt daran, dass die Evaluation bzw. der Nachweis der Wirksamkeit der Neuen Steuerung und ihrer Einzelmaßnahmen äußerst anspruchsvoll ist. Dies, so Fend (2011), liege erstens daran, dass die Wirkmechanismen sehr komplex seien und eine Isolation einzelner Faktoren, wie normalerweise in kontrafaktischen Analysen nötig, somit nur schwer möglich sei. Zweitens fehlten klare und elaborierte Wirksamkeitsdesigns, die beschreiben, wie die implementierten Mechanismen zu Qualitätssteigerungen führen sollen und können. Und drittens würden entsprechende Maßnahmen ihre Wirkung großflächig meist erst auf lange Sicht entfalten.

Die Folge ist, dass selbst großangelegte Maßnahmen im Bildungssystem ebenso schnell eingeführt wie wieder abgeschafft werden und auch Veränderungen dieser Maßnahmen häufiger weniger empirisch-evidenzbasiert denn als auf Basis politischer Agenden und anekdotischer Evidenz erfolgen. Entsprechend schließt Maag Merki (2018, S. 246f.) mit Blick auf die rezenten Reformen im deutschen Bildungssystem:

> „*Weder die Einführung von zentralen Abschlussprüfungen … noch von Schulinspektionen …, zentrale Reformvorhaben in den letzten Jahren in den deutschsprachigen Ländern, lässt sich eindeutig durch empirische Ergebnisse, die positive Veränderungen für das Bildungswesen annehmen lassen, begründen. Offenbar werden gewisse Entscheide hinsichtlich bestimmter Reformen oftmals einzig begrenzt erfahrensbasiert, sondern eher normativ oder aufgrund gesellschaftlicher Entwicklungen bestimmt; empirische Befunde mögen dabei willkommen sein, werden hinsichtlich des gewünschten Ergebnisses aber bewusst oder unbewusst (um)gedeutet und dienen damit höchstens der Legitimation des bereits normativ gesetzten Entscheides.*"

Die vorliegende Arbeit beschäftigt sich daher mit der Frage, ob eine evidenzbasierte Schul- und Unterrichtsentwicklung im Kontext Neuer Steuerung gelingen kann, wie Wirkungen systematisch evaluiert werden können und welche Mechanismen und Kontextbedingungen ggf. dafür

verantwortlich sein können, dass Wirkungen zu beobachten sind oder nicht. Nach einem kurzen Überblick über zentrale Annahmen im aktuellen Steuerungsparadigma sowie zur theoretischen Fundierung der Schul- und Unterrichtsentwicklungsforschung werden in vier Abschnitten konzeptionelle sowie empirische Arbeiten vorgestellt, die diesen Fragen nachgehen. Abschließend werden die Befunde diskutiert.

2. Evidenzbasierte Schul- und Unterrichtsentwicklung im Kontext neuer Steuerung

Die Neue Steuerung umfasst ein Bündel verwaltungspolitischer Reformbemühungen, die in erster Linie von einer betriebswirtschaftlichen Interpretation des Verwaltungshandelns geleitet werden. Dabei steht die effiziente und effektive Aufgabenwahrnehmung staatlicher Akteure im Vordergrund, wobei hohe Erwartungen an die Eigenverantwortung der Beteiligten gestellt und ebenfalls hohe Erwartungen in die Steuerungskompetenz des Marktes und wettbewerblicher Strukturen gesetzt werden (Schröter & Wollmann 2005). Damit unterscheidet sich dieses Modell deutlich vom bis zum Ende des 20. Jahrhunderts dominierenden bürokratischen Modell der Steuerung, in dem eine perfekt geplante Struktur die maximale Zielerreichung garantieren sollte, und in der Steuerung in erster Linie auf einer aktenmäßigen Verwaltung mit klaren Hierarchien basierte (Jann, 2005; Weber 1980, vgl. Tabelle 1).

Bürokratiemodell	**Neue Steuerung**
Ständige Eingriffe ins Tagesgeschäft	Steuerung auf Abstand
Exzessiver Zentralismus	Selbststeuerung dezentraler Einheiten
Organisierte Unverantwortlichkeit (Trennung von Fach- und Ressourcenverantwortung)	Abgestufte, weitgehend delegierte Eigenverantwortung
Übertriebene Arbeitsteilung und Spezialisierung	Re-Integration fragmentierter Aufgabenwahrnehmung
Orientierung an den internen Erfordernissen des Verwaltungsablaufs	Bürger- und Kundenorientierung
Orientierung an arbeitsplatzbezogener Ordnungsmäßigkeit	Umfassende Qualitätsorientierung
Abschottung vom Marktdruck, Monopole	Marktorientierung und Wettbewerb
Präferenz für Eigenherstellung	Konzentration auf Kernkompetenzen
Kameralistische Haushaltsführung	Transparenz von Kosten und Leistungen
Juristische Personalverwaltung	Personalmanagement (Leistungsanreize, Führung, Personalentwicklung)

Tabelle 1: Unterschiede bürokratischer und Neuer Steuerung nach Jann 2005

Zentrale Merkmale des aktuellen Steuerungsverständnisses sind daher grundsätzlich (Hood 2002): a) ein an unternehmerischen Verhaltensweisen orientiertes Management, b) Wettbewerb zwischen den Akteuren, c) die explizite Setzung von Standards, d) eine Output-Überprüfung, e) die Betonung von Wirtschaftlichkeit, f) die Dezentralisierung von Entscheidungsbefugnissen auf niedere Verwaltungseinrichtungen sowie g) ein Kontraktmanagement. Mit Blick auf die Reformen im deutschen Bildungssystem ist die Einführung der Neuen Steuerung aktuell wiederum durch drei Schwerpunkte charakterisiert (Altrichter, Rürup & Schuchart 2016): a) Schulautonomie bzw. die Erhöhung einzelschulischer Gestaltungsspielräume, b) die Verbetrieblichung der Einzelschule sowie c) eine evidenzbasierte Bildungspolitik und Schulentwicklung.

Eine zentrale Annahme im Rahmen des aktuellen Steuerungsparadigmas im deutschen Bildungssystem lautet daher, dass durch die systematische Erhebung von Informationen über schulische Prozesse und Ergebnisse und deren Rückmeldung an schulische Akteure eine Entwicklungsspirale in Gang gesetzt werden kann, die in einer verbesserten Bildungsqualität sowie gesteigerten Schülerleistungen mündet (Dedering 2012; Demski 2017; Fend 2011; Thiel, Cortina & Pant 2014). Die Idee, dass durch Rückmeldungen Entwicklungen ausgelöst werden können, basiert dabei auf Annahmen des klassischen Feedbackverständnisses zur Zielorientierung (Kluger & DeNisi 1996; Ramaprasad 1983), in denen davon ausgegangen wird, dass Ziele anzustrebende Sollgrößen darstellen und empiriebasierte Rückmeldungen den Ist-Stand der Zielerreichung wiedergeben. Die damit verbundene Erwartung lautet: Bei Abweichungen des Ist- vom Sollstand erfolgt eine Anpassung an das Ziel; je konkreter die Ziele vorgegeben sind und je spezifischer eine Rückmeldung gegeben wird, desto genauer kann die Fehleranalyse bei Abweichungen erfolgen und umso besser kann eine Handlungsoptimierung geplant werden, die es ermöglicht, die angestrebten Ziele in Zukunft besser zu erreichen.

Entsprechend wurden im Rahmen der Bildungsreformen seit der Jahrtausendwende sowohl Bildungsziele – Bildungsstandards – als auch Prozessziele – Rahmen zur Qualität von Schule und Unterricht – definiert, die zu erreichende Leistungsziele für Schulen und Schulverantwortliche vorgeben und für die der Grad der Zielerreichung wiederum mithilfe verschiedener Verfahren, wie z.B. Schulinspektionen und Lernstandserhebungen, empirisch festgestellt werden soll (Altrichter, Moosbrugger & Zuber 2016; Altrichter, Rürup & Schuchart 2016; Fuchs 2009). Die Messung der jeweiligen Aspekte und die Rückmeldung der entsprechenden Befunde sollen auf diesem Wege vor allem dazu beitragen, den dezentralisierten Entscheidungsträgern auf Ebene der Einzelschule „*dabei zu helfen, ihre (…) Entscheidungen ein wenig rationaler zu treffen, um so die Qualität (…) zu verbessern*.“ (Stufflebeam 1972, S. 135). Eine solche Annahme zur Nutzung von Evaluationsbefunden

unterstellt dann konsequenterweise, dass Entscheidungen rational, auf Basis bereitgestellter Informationen in einem prozessualen Ablauf und somit von den Datennutzerinnen und Datennutzern möglichst so genannte *clear-cut decisions* (Scheerens, Glas & Thomas, 2003) getroffen werden.

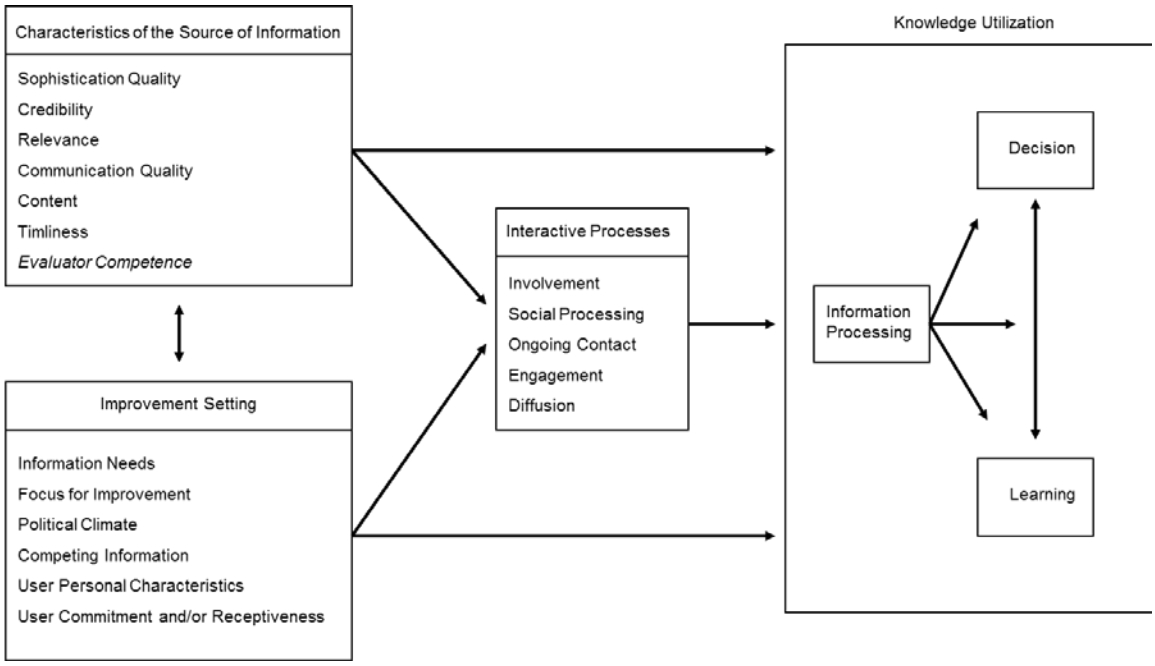


Abbildung 1: Determinanten der Evaluationsnutzung nach Johnson et al. 2009

Dabei ist die Idee rationaler Entscheidungen jedoch eher ein Ideal als eine Beschreibung davon, wie Entscheidungen tatsächlich getroffen werden (vgl. Tarter & Hoy 1997). So werden in der alltäglichen Entscheidungsfindung von Datennutzerinnen und Datennutzern vor allem begrenzt rationale Entscheidungen getroffen, in deren Rahmen einfache Heuristiken zum Einsatz kommen, um bei begrenzten zeitlichen und kognitiven Ressourcen schnell zu zufriedenstellenden Lösungen zu gelangen (Scheerens, Glas & Thomas, 2003). Verschiedene Untersuchungen zeigen dann auch, dass die (zielgerichtete) Nutzung von Daten und eine darauf basierende Qualitätsentwicklung von einer Vielzahl von Faktoren abhängen (Demski 2017; Johnson et al. 2009; Schildkamp, Lai & Earl 2013; Schildkamp et al. 2017). So spielen erstens Charakteristika des Entscheidungssettings, zweitens Charakteristika der Evaluation selber (z.B. Validität und qualitativ hochwertige Informationen) und drittens Kontextmerkmale eine Rolle dabei, ob und, falls ja, wie empirische Informationen zur Qualitätsentwicklung genutzt werden (Johnson et al. 2009, Abb. 1). Hinzu kommen Interaktionseffekte verschiedener Art und ggf. reziproke Zusammenhänge (Chrispeels, Brown & Castillo 2000; Marsh & Farrell 2015).

Wie in einem solchen komplexen Kontext der evidenzbasierten Steuerung Schul- und Unterrichtsentwicklung dann tatsächlich gelingen kann, haben Bryk et al. (2010, Abb. 2) gezeigt und, in der Tradition der School-Effectiveness- und School-Improvement-Forschung, ein empirisch validiertes (logisches) Schulentwicklungsmodell erarbeitet, das zeigt, welche Faktoren nachweislich eine nachhaltige Schul- und Unterrichtsentwicklung unterstützen. Auf Ebene der Einzelschule sind dies a) die Beziehungen zwischen schulischem Personal, Eltern und (Schul-)Gemeinde (parent und community ties), b) die professionellen Kapazitäten schulischer Mitarbeiterinnen und Mitarbeiter (professional capacity), c) ein Lernklima, das auf das Lernen von Schülerinnen und Schülern fokussiert (student-centered learning climate), d) klare Regeln und Verabredungen mit Blick auf das Unterrichten (coherent instructional guidance system) sowie e) Schulleitungshandeln als essentieller Treiber für Veränderung und Entwicklung (leadership as driver for change). Dabei zeigen die Autoren aber auch, dass nur im Zusammenspiel aller Faktoren Schulentwicklung stattfinden kann – bereits das Nichtvorhandensein eines Aspektes kann den Entwicklungsprozess stören – und dass die Wirksamkeit dieser Faktoren wiederum stark von den Kontextbedingungen einer Schule sowie dem Verhältnis und dem Vertrauen zwischen den verschiedenen am Entwicklungsprozess beteiligten Akteuren (z.B. Behörden, Lehrkräfte, Schulaufsichten, Schulleitungen etc.) abhängt.

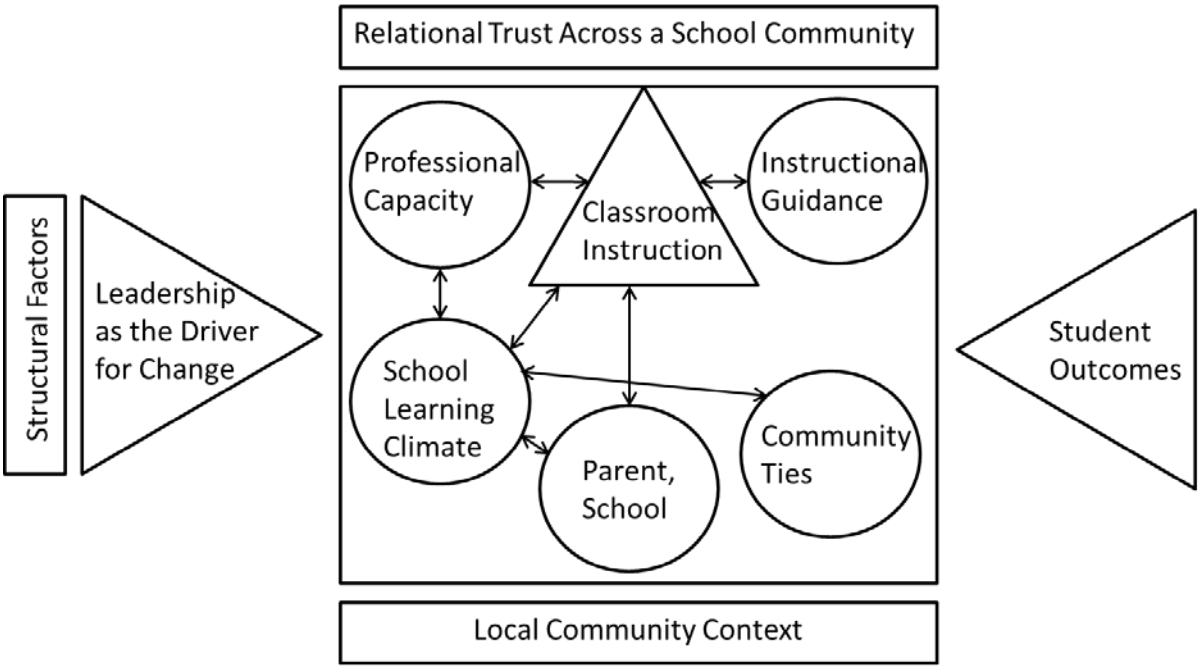


Abbildung 2: Faktoren gelingender Schul- und Unterrichtsentwicklung nach Bryk et al. 2010

Ein belastbares theoriegeleitetes Modell der evidenzbasierten Schulentwicklung, das Begründungen dafür liefert, dass und wie diese Faktoren wirksam werden, liegt bislang hingegen nicht vor. Mit Blick auf den deutschsprachigen Raum konstatieren darüber hinaus Klieme und Steinert (2008, S. 223),

dass „*die empirische Befundlage der Schulentwicklungsforschung auffallend dünn*“ sei. Ein Fakt, der sich auch 10 Jahre später kaum geändert hat (Berkemeyer & Hermstein 2018). Einig ist man sich derzeit jedoch anscheinend darin (Berkemeyer & Hermstein 2018; Demski 2017; Dedering 2012), dass erstens Schulen als pädagogische Handlungseinheiten (Fend 1986) betrachtet werden können und ihnen somit die Fähigkeit eigenständigen Handelns zuzusprechen ist. Zweitens, dass die zentralen Zielkriterien von Schul- und Unterrichtsentwicklung die Qualität von Schule und Unterricht sind (Dedering 2012). Drittens, dass sowohl bildungspolitische und administrative Vorgaben als auch empirische Befunde durch die jeweilig beteiligten Akteure auf den ihnen eigenen Handlungsebenen des schulischen Mehrebenensystems ständig rekontextualisiert werden müssen, damit Entwicklungsimpulse wirksam werden, also eine Übersetzung von Regulationen, normativen Vorgaben und empirischen Informationen in spezifische berufliche und institutionelle Handlungskontexte stattfinden muss (Fend 2008, 2011). Und viertens, dass Kontextbedingungen sowie das Setzen von Anreiz- und Unterstützungselementen den Schul- und Unterrichtsentwicklungsprozess im Rahmen des aktuellen Steuerungsparadigmas zentral beeinflussen (Altrichter & Maag Merki 2016). Wie sich letztere Pole (*Pressure* und *Support*) jedoch zueinander verhalten bzw. verhalten sollen, ist wiederum nicht abschließend geklärt (Thiel, Cortina & Pant 2014).

Diejenigen wenigen eher theoriegeleiteten Modelle, die im deutschsprachigen Raum derzeit diskutiert werden, greifen dann auch mehr oder weniger eklektisch auf Annahmen aus Pädagogik, Didaktik, Psychologie, Soziologie und weiteren Disziplinen zurück und kombinieren Teilaspekte daraus in Theorieverbünden (Rahm 2005) oder bezugstheorieorientierten Architekturen (Maag Merki 2008) der Schulentwicklung, die wiederum mit schultheoretischen und/oder lerntheoretischen und/oder organisationstheoretischen Aspekten angereichert werden. Entsprechend resümiert Reinbacher (2016, S. 295) mit Blick auf die deutschsprachige Schulentwicklungsforschung:

> „*Es handelt sich um eine über weite Strecken theoriefreie Zone zwar nicht in dem Sinne, dass Schulentwicklung auf die Anwendung von Theorien unterschiedlicher Herkunft verzichtet, jedoch insofern, als es jenseits solcher eklektizistischer Bricolage an theoriegeleiteten Modellen zur Integration einzelner Befunde mangelt.*“

Zusammenfassend bleibt daher viererlei festzuhalten: Erstens wird mit Blick auf die Neue Steuerung im deutschen Bildungssystem derzeit angenommen, dass zunehmend autonome schulische Akteure in einem dezentralisierten und wettbewerbsorientierten Kontext, auf Basis valider sowie qualitativ hochwertiger Rückmeldeinformationen wissensbasierte Entscheidungen treffen, die, vermittelt über

eine Qualitätssteigerung von Unterricht und Schule, in verbesserten Schülerleistungen münden sollen. Zweitens haben sich auf internationaler Ebene empirisch sowohl Faktoren herauskristallisiert, die die Nutzung empirisch gewonnener Informationen für die Schul- und Unterrichtsentwicklung befördern, als auch solche Faktoren, die essentiell für eine gelingende Entwicklung von Schule und Unterricht selbst sind. Drittens fehlen dennoch kohärent-schlüssige theoriegeleitete Modelle, die es ermöglichen, Entwicklungen zu definieren, zu beschreiben, zu analysieren und empirische Befunde einzuordnen. Und viertens liegen für den deutschsprachigen Raum derzeit kaum belastbare empirische Befunde dazu vor, ob und wie eine evidenzbasierte Schul- und Unterrichtsentwicklung infolge neuer Steuerungsmechanismen stattfindet und zu einer Theoriebildung beitragen könnte.

3. Beschreibung und Zusammenfassung der vorgelegten Studien

3.1. Systematische Evaluation von Schulinspektionsverfahren

Vor diesem Hintergrund befassen sich die ersten drei Arbeiten inhaltlich und konzeptionell mit der Frage, wie Effekte von Schulinspektionen evaluiert und bewertet werden können. Dahinter steht die Annahme, dass Schulinspektionen, als Instrument der Neuen Steuerung, in der Konsequenz zu einer evidenzbasierten Entwicklung von Schule und Unterricht und in der Folge zu verbesserten Schülerleistungen führen sollen. In den drei folgenden Arbeiten wurden daher einerseits Schemata erarbeitet, die es ermöglichen sollen, die Effekte von Schulinspektionsverfahren systematisch zu evaluieren und die Verfahren zu validieren. Andererseits werden diese Schemata angewandt, um die bislang vorliegenden Studien zum Thema zu systematisieren und Forschungsdesiderata aufzuzeigen.

Dabei steckt der erste Beitrag den Rahmen für alle weiteren Beiträge dieser Arbeit, indem vorgeschlagen wird, mit Blick auf Wirkungsanalysen dem Verständnis der mechanismenbasierten Evaluation (Chen 1990; Pawson & Tilley 1997) zu folgen und in einem kleinschrittigen, iterativen Prozess aus empirischen Analysen und programmtheoretischer Theoriebildung den Fragen nachzugehen: „What works?“ und „What works for whom in what circumstances and why?“ (Astbury & Leeuw 2010). Zentral sind in diesem Ansatz so genannte Mechanismen, die durch Programme bzw. Interventionen ausgelöst werden und die je nach Kontext zu unterschiedlichen Ergebnissen führen können (Weber 2006). Dabei sind Mechanismen zu verstehen als “the causal forces, powers, processes or interactions that generate change, combining the use of resources and reasoning that people make – including the choices, reasoning, and decisions that people make as a result of the resources provided by the programme” (Punton, Vogel & Lloyd 2016, S. 2).

Beitrag 1: 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, N. (Hrsg.), *Schul- und Bildungsforschung. Diskussionen, Befunde und Perspektiven* (S. 167-185). Münster: Waxmann.

Zentrales Anliegen des ersten Beitrages ist es daher, herauszuarbeiten, warum eine Vielzahl empirischer Studien zum dem Schluss kommt, dass Schulinspektionen mit Blick auf die Lernentwicklung von Schülerinnen und Schülern in der Regel wirkungslos bleiben. Basierend auf einem programmtheoretischen Ansatz werden diejenigen empirischen Studien, die zum Thema vorliegen, mit Blick auf Theorie-, Programm- und Methodenfehler untersucht. Dabei werden Schulinspektionen als Interventionen auf Schulebene verstanden, die einen Effekt nach sich ziehen sollen. Während ein Programmfehler vorhanden ist, wenn es mittels einer Intervention nicht gelingt, eine intendierte Wirkung nachweisbar zu erzeugen, liegt ein Theoriefehler vor, sofern ein theoretisch postulierter Wirkungszusammenhang nicht ausreichend valide begründet wurde. Ein Methodenfehler liegt wiederum vor, wenn es im Rahmen der empirischen Kausalstudie aufgrund methodologischer Unzulänglichkeiten nicht gelingt, den angenommenen Wirkungszusammenhang empirisch verlässlich zu überprüfen.

Mit Blick auf Theoriefehler zeigen die Befunde, dass ausgearbeitete Programmtheorien zur Wirksamkeit von Schulinspektionen bislang nur vereinzelt vorliegen und die Modelle in der Regel nicht generalisierbar sind, also nur für den jeweils spezifischen Kontext anwendbar sind. Desweiteren wird in den aktuell vorliegenden Modellen sowohl die Komplexität der Intervention als auch deren Kompliziertheit kaum berücksichtigt. Denn einerseits handelt es sich bei schulinspektionsbasierten Interventionen aus programmtheoretischer Perspektive um komplizierte Interventionen, die aus mehreren Komponenten bestehen und z. B. multidimensional angelegt sind und/oder nur im Zusammenspiel mit anderen Interventionen Wirkung zeigen. Und andererseits sind diese Interventionen auch komplex, da sie durch verschiedene Akteure oder Organisationen implementiert werden, die in Abhängigkeit von den gegebenen Voraussetzungen unterschiedliche Wirkungen entfalten können und daher sowohl abhängig von den Ausgangs- als auch den Kontextbedingungen sind, unter denen die Intervention stattfindet, und sich darüber hinaus adaptiv verhalten. Es ist daher möglich, dass berichtete Effekte bzw. Nicht-Wirksamkeiten darauf zurückzuführen sind, dass relevante Aspekte außer Acht gelassen wurden.

Hinsichtlich von Programmfehlern machen die Befunde deutlich, dass zur Evaluation der Wirksamkeit von Schulinspektion derzeit vor allem Blackbox-Verfahren eingesetzt werden, also Studiendesigns,

die es ermöglichen herauszufinden, ob eine intendierte Wirkung infolge einer Intervention eintritt oder nicht, ohne dabei die Wirkmechanismen zu berücksichtigen. In den vorhandenen empirischen Studien wird daher geprüft, ob die Intervention Schulinspektion einen direkten Einfluss auf bestimmte Effektvariablen wie z.B. Schülerleistungen hat. Untersucht werden zumeist Effekte auf Einstellungen und Handlungsabsichten von Schulbeteiligten, direkte Reaktionen auf Ebene der Einzelschule auf eine Schulinspektion und die Veränderung von Schülerleistungen. So werden mit Blick auf die Entwicklung von Schülerleistungen zumeist Testleistungen reanalysiert. Die bislang vorliegenden Studien präsentieren ein heterogenes Bild, und es werden sowohl negative als auch positive und Null-Effekte von Schulinspektionsverfahren auf Schülerleistungen berichtet. Es ist insofern unklar, ob die berichteten Effekte eine Folge nicht gut oder nicht korrekt durchgeführter Schulinspektionen sind.

Bezogen auf Methodenfehler wiederum zeigt die Auseinandersetzung, dass bislang nur eine Studie untersucht hat, ob die empirischen Methoden, mit deren Hilfe in den letzten Jahren auf die Wirksamkeit von Schulinspektion geschlossen wurde, überhaupt geeignet sind, um kausale Zusammenhänge zu modellieren. Dabei kamen die Autoren zu dem Schluss, dass es in allen bis zum damaligen Zeitpunkt durchgeführten Studien problematisch ist, dass die genutzten Daten nicht auf Zufallsstichproben von Schulen basieren. Entsprechend konnte in diesen Untersuchungen nicht abgesichert werden, dass Befunde nicht arbiträr und dem Vorhandensein unzureichender Stichproben geschuldet waren. Darüber hinaus wurden auch nahezu nirgends statistische Verfahren zur Schätzung des Treatment-Effekts nach der Kontrolle beobachtbarer Variablen eingesetzt. Es ist daher nicht auszuschließen, dass die berichteten Effekte auf ein unzureichendes methodisches Vorgehen zurückzuführen sind.

Die Probleme der bislang vorliegenden Studien liegen unseren Analysen zufolge daher vor allem im Bereich der Theorien und der angewandten Methoden: a) die aktuell genutzten Programmtheorien sind zu unterkomplex und berücksichtigen darüber hinaus kaum die Kompliziertheit der Intervention, b) die eingesetzten Forschungsmethoden wiederum sind kausalanalytischen Fragestellungen zumeist nicht angemessen und berücksichtigen bekannte methodologische Fallstricke nicht oder nur ungenügend. Dies wiederum hat zur Folge, dass Aussagen zu Programmfehlern nur schwer möglich sind und die bislang vorliegenden Befunde wenig verlässliche Aussagen zur Wirksamkeit von Schulinspektionsinterventionen auf Schülerleistungen ermöglichen.

Mit Blick auf die weitere Forschung im Bereich der Wirkungen von Schulinspektionen wird entsprechend empfohlen, diese auf zweierlei Art zu evaluieren: Erstens theoriebasiert mithilfe der

General-Elimination-Methode (GEM), auch Modus-Operandi-Methode genannt, in deren Rahmen man die Ursachen für Wirkungen ableitet, indem man Ereignisse, Prozesse oder Eigenschaften untersucht, die mit den Effekten zu tun haben (können); und zweitens mithilfe von Method-Driven-Evaluationsansätzen, in deren Rahmen mithilfe angemessener kausalanalytischer Verfahren Black-Box-Evaluationen zur Wirksamkeit von Schulinspektionen durchgeführt werden.

Beitrag 2: Pietsch, M., van den Ham, A.-K. & Köller, O. (2015). Wirkung von Schulinspektion: Ein Rahmen zur theoriegeleiteten Analyse von Schulinspektionseffekten. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), Schulinspektion in Hamburg. Der erste Zyklus der 2007 – 2013: Grundlagen, Befunde, Perspektiven (S. 117-136). Münster: Waxmann.

Der zweite Beitrag greift die Befunde des ersten Beitrages auf und entwickelt auf dieser Basis ein theoriebasiertes, dreischrittiges Verfahren zur systematischen Evaluation von Schulinspektionseffekten. Hierbei wird insbesondere dafür plädiert, dass im Rahmen theoriegeleiteter Evaluation von Schulinspektionseffekten die Komplexität und die Kompliziertheit der Intervention stärker berücksichtigt werden als bisher und die im Rahmen programmtheoretisch fundierter Analysen generierten Befunde mittels des Ansatzes des „Interpretation / Use Argument“ nach Kane validiert werden, um so die Belastbarkeit der Studien und der daraus abzuleitenden Konsequenzen für die Weiterentwicklung von Inspektionsverfahren zu erhöhen. Dabei knüpft das Verständnis von Validität an das Konzept der Informellen Argumente an und empfiehlt daher Validitätsnachweise in Form von Validitätsargumenten zu organisieren, wie z.B. im Argument Based Approach oder dem Evidence-Centered Design üblich.

Zunächst wird im Rahmen des Beitrages herausgearbeitet, dass die zentrale Funktion von Schulinspektionen in Deutschland vor allem in ihrer Schulentwicklungsfunktion liegt, um anschließend darauf einzugehen, welche Möglichkeiten Schulbeteiligte haben, um mit Befunden aus Inspektionen umzugehen. Diesbezüglich können grundsätzlich drei Nutzungssysteme voneinander unterschieden werden:

> 1) Instrumentelle Nutzung (Instrumental Use): Die instrumentelle Nutzung von Informationen aus Evaluationen bezieht sich darauf, konkrete Probleme zu lösen bzw. konkrete Entscheidungen zu treffen, die den Evaluationsgegenstand betreffen.

2) Konzeptionelle Nutzung (Conceptual Use): Die konzeptionelle Nutzung von Informationen aus Evaluationen findet statt, wenn diese indirekt genutzt werden, um das Wissen bzgl. des Evaluationsgegenstandes zu erweitern.

3) Symbolische Nutzung (Symbolic Use): Die symbolische Nutzung von Informationen aus Evaluationen findet statt, wenn Befunde zum Evaluationsgegenstand eingesetzt werden, um bereits getroffene Entscheidungen gegenüber Dritten zu legitimieren.

Der vorgestellte konzeptionelle Rahmen zur Analyse von Schulinspektionseffekten schlägt im Weiteren ein dreischrittiges Vorgehen vor. So wird zuerst ein dreidimensionales Schema (Zeit, Intention, Quelle) zur Beschreibung von erwarteten Inspektionswirkungen erarbeitet, wobei der Begriff Wirkung zugunsten des Begriffs Einfluss (Influence) verworfen wird, um sowohl intendierte als auch nicht-intendierte Effekte in künftigen Analysen mithilfe eines kohärenten Designs analysieren zu können. Der Vorteil eines solchen Schemas liegt dabei vor allem darin, dass er über den Begriff der Nutzung hinaus geht und es vor allem auch erlaubt, nicht-intendierte Wirkungen in die Analysen mit einzubeziehen und darüber hinaus Effekte resp. Veränderungen – und somit Wirkungen, nicht die Nutzung von Evaluationen – in den Mittelpunkt stellt. Weiterhin ermöglicht dieses Schema zu berücksichtigen, dass Schulinspektionen, anders als z. B. Vergleichsarbeiten und Lernstandserhebungen, auch einen Einfluss durch das Wirken von Inspektorinnen und Inspektoren vor Ort haben und nicht ausschließlich Effekte durch die Rückmeldung von Ergebnissen erzielen. Untersucht werden kann in einem solchen Design somit: a) ob intendierte Effekte durch die Evaluation erzielt wurden oder nicht (Intention), b) ob diese durch die Ergebnisrückmeldung oder den Evaluationsprozess zustande gekommen sind (Quelle) und c) ob die Effekte unmittelbar, zum Ende eines Zyklus oder aber erst langfristig nachweisbar sind (Zeit).

In einem weiteren Schritt geht es im Rahmen einer theoriegeleiteten Evaluation zu Wirkungen von Schulinspektion dann darum zu beschreiben, mithilfe welcher Mechanismen, auf welche Art und Weise die definierten Wirkungen durch ein Schulinspektionsverfahren generiert werden sollen. Von einfach-linearen Modellen wird abgeraten, da diese für die Analyse von Schulinspektionswirkungen zu unterkomplex sind (siehe auch Beitrag 1). Es wird dabei in Anlehnung an Vorarbeiten aus der programmtheoretischen Forschung dafür plädiert, logische Modelle für die Evaluation von Inspektionswirkungen zu nutzen, die sowohl ein Veränderungs- als auch ein Handlungsmodell beinhalten und in der Lage sind, reziproke Effekte zu berücksichtigen.

In einem dritten Schritt ist dann das Ziel, die so modellierten Ergebnisse sinnvoll zusammenzufassen und zu validieren. Diesbezüglich wird auf die Standards für Pädagogisches und Psychologisches Testen von 1999 und 2014 verwiesen und Validität als das Ausmaß definiert, in dem empirische Nachweise und Theorien die Interpretation von Testergebnissen für eine bestimmte Nutzung stützen. Vorgeschlagen wird in diesem Zusammenhang daher eine argumentbasierte Validierung von Inspektionseffekten in Anlehnung an Toulin sowie die Anwendung des „Interpretation/ Use-Arguments“ nach Kane. Dieses Modell kann unserer Ansicht nach zur Evaluation der Wirksamkeit von Schulinspektionen genutzt werden und Programm-, Theorie- und Methodenfehler aufdecken bzw. diesen vorbeugen. Dabei wird die Argumentationskette zweimal durchlaufen. In der ersten Phase wird die Maßnahme, in diesem Fall die Schulinspektion, evaluiert. Bei der Bewertung kann überprüft werden, ob mit den Instrumenten ermittelte Ergebnisse das Verhalten der Schule tatsächlich repräsentieren oder ob beispielsweise die Bewertung der Fragebögen, Interviews, Schulbegehungen etc. subjektiv oder fehlerbehaftet sind. Bei der Generalisierung kann untersucht werden, ob sich die Schulinspektion auf andere Facetten und Kontexte der Durchführung übertragen lässt. Hier stellt sich unter anderem die Frage, ob eine Inspektion mit anderen Inspektoren, zu einem anderen Zeitpunkt, in anderen Unterrichtsstunden oder bei anderen Lehrkräften ein gleichwertiges Ergebnis produziert hätte. Bei der Extrapolation ist es möglich zu analysieren, ob das Konstrukt einer „guten Schule“ empirisch bestätigt werden kann und ob die Erfüllung dieser Normen tatsächlich eine „gute Schule“ in der realen Welt bedingt. Lassen sich keine Argumente oder sogar Gegenargumente dafür finden, dass das Ergebnis der Schulinspektion die reale Schulwirklichkeit beschreibt, kann dies keinen oder einen unerwünschten Prozess zur Folge haben. Gründe können ein Programm- oder ein Methodenfehler in der Schulinspektion sein. Können jedoch Argumente gefunden werden, dass das Ergebnis der Schulinspektion die reale Schulwirklichkeit beschreibt, so liegt eine solide Basis für die Schulentwicklung vor. Auf dieser Basis können dann Entscheidungen getroffen und ein gewünschter Prozess initiiert werden. Der Prozess führt wiederum zu einem eventuell veränderten beobachtbaren Verhalten der Schule. Dieses Verhalten kann erneut mithilfe von Instrumenten gemessen werden, womit das Modell von Kane ein zweites Mal durchlaufen werden kann.

Beitrag 3: Ehren, M. & Pietsch, M. (2016). Validation of Inspection Frameworks and Methods. In M. Ehren (Hrsg.), Methods and Modalities of effective School Inspections (S. 47-68). London: Springer.

Der dritte Beitrag greift wiederum den zweiten Beitrag auf und überträgt diesen in einen internationalen Kontext. Konkret wird in einem ersten Schritt ein argumentbasiertes Validierungsschema für den Bereich der Schulinspektion erarbeitet und vorgestellt und dieses in

einem zweiten Schritt genutzt, um vorliegende Befunde zur Reliabilität und Validität von Schulinspektionsverfahren einzuordnen und analysieren. Beim Vorgehen wird dabei auf die AERA-Standards for Educational and Psychological Testing verwiesen und vorliegende Studien aus Deutschland, Großbritannien und den Niederlanden entsprechend anhand der dort definierten Validitätsdimensionen – Validity Evidence Based on Test Content, Validity Evidence Based on Relations to Other Variables, Validity Evidence Based on Internal Structur, Validity Evidence Based on Response Processes und Validity Evidence Based on Consequences of Testing – klassifiziert.

Die Befunde zeigen grundsätzlich, dass auf internationaler Ebene zu allen genannten Validitätsdimensionen erste empirische und konzeptionelle Arbeiten vorliegen. Gleichwohl sind Studien, die die Validität von Inspektionen anhand von Außenkriterien prüfen, Studien, die die Inhaltsvalidität (Test Content) von Orientierungsrahmen in den Blick nehmen sowie Studien, die den Impact von Schulinspektionen analysieren, auf internationaler Ebene äußert rar. Wie bereits Beitrag 1 zeigt diese Studie insofern, dass erheblicher Forschungsbedarf im Bereich der Schulinspektionswirksamkeitsforschung besteht.

Die Befunde machen darüber hinaus deutlich, dass die Standardisierung von Annahmen zur Qualität von Schule und Unterricht im Rahmen von Orientierungsrahmen für Schulqualität aus verschiedenen Gründen problematisch sein kann. So kann eine solche Standardisierung beispielsweise dazu führen, dass festgelegte Standards von Schulinspektorinnen und Schulinspektoren im Rahmen der Inspektionspraxis von Schule zu Schule (und von Inspektor zu Inspektor) unterschiedlich interpretiert und gehandhabt werden – die seitens der Bildungsadministration gewünschte und für die Kontrollfunktion von Schulinspektionen notwendige Objektivität also nicht gegeben ist. Daher wird dafür plädiert, Orientierungsrahmen so zu gestalten, dass sie es in der Inspektionspraxis ermöglichen kontextspezifisch zu inspizieren. Dies, so schließen wir, erfordert ein entsprechend wissenschaftlich fundiertes Standard-Setting-Procedere mit Blick auf kontextualisierte Assessments.

Einordnung der unter 3.1. berichteten Beiträge

Zusammengenommen machen die hier berichteten Beiträge deutlich, dass die Forschung zu Inspektionswirkungen sowie zu deren Wirkmechanismen derzeit nicht sehr weit entwickelt ist. Entscheidungen zu entsprechenden Verfahren im Rahmen von Bildungspolitik, Bildungsadministration und Inspektionspraxis erfolgen insofern meist ohne wissenschaftliche

Expertise und, anders als im Rahmen der Neuen Steuerung im Bildungssystem gefordert, überwiegend nicht evidenzbasiert.

Die drei Beiträge sind darüber hinaus für den Diskurs bedeutsam, da sie deutlich machen, wo entsprechende Lücken vorhanden sind, welche Forschungsdesidarate zukünftig bearbeitet werden müssen und mithilfe welcher konkreten Ansätze die Wirksamkeit von Schulinspektionsverfahren definiert und analysiert werden kann. Auch zeigen die Befunde, dass im überwiegenden Teil aktueller Studien von unterkomplexen Modellen ausgegangen wird und viele Studien nicht dem methodischen State-of-the-Art entsprechen.

Insbesondere schließen die Arbeiten dabei die bislang eklatante Lücke zwischen Evaluationsforschung, empirischer Bildungsforschung und Schulinspektionsforschung, indem aktuelle Modelle, Annahmen und Methoden der Evaluationsforschung und der empirischen Bildungsforschung erstmals mit der Schulinspektionsforschung in Verbindung gebracht werden.

3.2. Wirkung(en) von Schulinspektionsverfahren

Wie bereits in Abschnitt 3.1. dargestellt, sollen Schulinspektionen in Deutschland zu einer evidenzbasierten (Weiter-)Entwicklung von Schule und Unterrichts führen und vermittelt darüber in gesteigerten Schülerleistungen münden. Die folgenden drei Beiträge beschäftigen sich daher mit den Wirkungen von Schulinspektionen auf ebenjene abhängigen Variablen. Anhand dreier unterschiedlicher Datensätze wird untersucht, welche Effekte Schulinspektionen auf die zentralen Variablen Unterricht und Schülerleistung haben. Inhaltlich handelt es sich um Black-Box- sowie Grey-Box-Evaluationen. So werden einerseits die Effekte auf Schülerleistungen evaluiert, ohne die Wirkungsmechanismen und Wirkungsbedingungen in den Blick zu nehmen. Andererseits werden Interventions- und Wirkungszusammenhänge in logischen Modellen betrachtet, ohne jedoch im Detail darauf einzugehen, wie diese konkret funktionieren. Hierbei kommen methodisch sowohl rein quantitative Ansätze als auch ein Mixed-Method-Ansatz zum Einsatz.

Beitrag 4: Pietsch, M., Janke, N. & Mohr, I. (2014). Führt Schulinspektion zu besseren Schülerleistungen? Difference-in-Differences-Studien zu Effekten der Schulinspektion Hamburg auf Lernzuwächse und Leistungstrends. Zeitschrift für Pädagogik, 60(3), 446-470.

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, wie in Beitrag 1 gezeigt, 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.

Im Rahmen des Beitrages wird davon ausgegangen, dass Schulinspektion infolge der Einzelschulevaluation einen Effekt auf die Lernergebnisse von Schülerinnen und Schülern haben sollte. Um die antizipierten Wirkmechanismen herauszuarbeiten, werden dafür in einem ersten Schritt vorhandene Modelle verglichen, die die pädagogischen Verarbeitungsprozesse von Rückmeldungen aus Evaluationen in den Blick nehmen. Deutlich wird, dass alle Autoren kontextualisierte, ökologische und vergleichsweise umfassende Modelle entworfen haben, die sowohl schulinterne als auch schulexterne und teilweise sogar Persönlichkeitsmerkmale von Lehrenden und Schulleitungen als moderierende Faktoren mit in den Blick nehmen, die sich jedoch in der spezifischen Reichweite voneinander unterscheiden. Dabei liegt der Wert der vorgestellten Modelle unserer Meinung nach in erster Linie in der Zusammenstellung von Mechanismen und Moderatorvariablen, mit deren Hilfe Wirksamkeitsannahmen 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. Eine empirische Überprüfung einer komplexen Programmtheorie ist in einem solchen Fall nicht möglich und würde ggf. dazu führen, dass Ad-hoc-Theorien aufgestellt werden, die dem Untersuchungsgegenstand nicht angemessen sind und entsprechend zu Fehlschlüssen führen könnten. Entsprechend plädieren wir, unter Bezugnahme auf Beitrag 1, dafür, zur Bestimmung von Inspektionseffekten auf Schülerleistungen Blackbox-Verfahren zu nutzen, die dann jedoch hohen methodischen Anforderungen entsprechen müssen.

Im Folgenden stellen wir daher verschiedene Methoden vor, mit deren Hilfe untersucht werden kann, ob die Intervention Schulinspektion intendierte Effekte nach sich zieht oder nicht. Dabei wird vertieft auf das Problem von Selektionseffekten in empirischen Studien eingegangen und dargestellt, welche Methoden zur Klärung kausalanalytischer Fragestellungen angemessen sind. Anschließend wird gezeigt, dass in allen bislang vorgelegten Studien, in denen Schulinspektionsverfahren in Beziehung zu Schülerleistungsdaten gesetzt wurden, keine Zufallsstichproben von Schulen genutzt wurden und auch nur in einem Drittel aller Studien (33,3%) Methoden genutzt wurden, um einen hieraus resultierenden Selektionsbias statistisch zu kontrollieren. Es zeigt sich somit, dass fast alle 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.

Im Folgenden werden dann – erstmals für eine Schulinspektion in Deutschland – Effekte der Schulinspektion Hamburg auf Schülerleistungen untersucht. Hierfür werden einerseits Querschnittsdaten aus dem Zentralabitur der Stadt Hamburg für zwei Messzeitpunkte und andererseits längsschnittlich erhobene Leistungsdaten aus der Studie Kompetenzen und Einstellungen von Schülerinnen und Schülern (KESS) genutzt. Für das konkrete Vorgehen wurde ein ökonometrischer Difference-in-Differences-Ansatz eingesetzt, da Schülerleistungsdaten zu mehreren Messzeitpunkten miteinander verglichen werden sollten. 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.

Die Ergebnisse der Studien zeigen, dass bei Einsatz maßgeschneiderter kausalanalytischer Verfahren sowohl Effekte auf Lernzuwächse als auch Leistungstrends von Schülerinnen und Schülern in Hamburg nachgewiesen werden können. In beiden Studien finden sich positive Zusammenhänge mit der Deutschleistung, im Rahmen des Zentralabiturs auch zum Fach Mathematik. Damit widersprechen die vorgelegten Analysen klar den meisten bislang vorliegenden und insbesondere den in England generierten Befunden zur Wirksamkeit von Schulinspektion auf Schülerleistungen, die überwiegend mithilfe fragwürdiger kausalanalytischer Methoden generiert wurden, und decken sich

eher mit denjenigen Befunden aus einer niederländischen Studie, für die festgestellt werden konnte, dass dort ein möglicher Selektionsbias mit statistisch angemessenen Verfahren korrigiert wurde.

Beitrag 5: Pietsch, M., Feldhoff, T. & Petersen, L. S. (2016). Schulentwicklung durch Inspektion? Welche Rolle spielen innerschulische Verarbeitungskapazitäten? In AG Schulinspektion (Hrsg.), Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten (S. 227-262). Wiesbaden: VS.

Auch dieser Beitrag setzt sich empirisch mit der Wirksamkeit von Schulinspektionsverfahren auseinander. Der Fokus liegt dabei auf innerschulischen Voraussetzungen und es wird der Frage nachgegangen, inwiefern innerschulische Verarbeitungsaktivitäten bzw. –voraussetzungen relevant für Schulinspektionseffekte sind. Hierfür wird das Modell der Kapazitäten organisationalen Lernens genutzt und Daten von 49 Hamburger Schulen, die infolge der Schulinspektion Entwicklungsmaßnahmen ergriffen haben, mithilfe eines Mixed-Method-Ansatzes analysiert.

Nach einer kurzen Einführung zu den Wirkungsweisen von Schulinspektionen wird daher anhand eines theoretischen Schulentwicklungsmodells untersucht, ob Schulen mit unterschiedlichen Schulentwicklungskapazitäten sich in Bezug auf initiierte und umgesetzte Maßnahmen der Schul- und Unterrichtsentwicklung im Anschluss an die Schulinspektion unterscheiden. Damit wird sowohl international als auch für den deutschsprachigen Raum Neuland beschritten. Denn während für den Bereich der schulischen Kontextmerkmale in jüngster Zeit erste vielversprechende Ideen und Erklärungsmodelle eingebracht wurden und sich zeigt, dass insbesondere Entwicklungsdruck dazu beiträgt, dass sich Schulen infolge von Inspektionen entwickeln, standen vergleichbare Modellierungen für innerschulische Merkmale und Prozesse zum damaligen Zeitpunkt noch aus.

Als Basis für die empirischen Analysen dient das Modell der „Kapazitäten organisationalen Lernens“. Dieser Ansatz unterscheidet sich von normativen Ansätzen der Schulentwicklung. Er beruht auf einer organisationstheoretischen Fundierung sowie empirischen und theoretischen Befunden der internationalen Forschung zum Organisationalen Lernen im schulischen Kontext unter Bezugnahme auf allgemeintheoretischr Konzepte zum Organisationalen Lernen. Dabei unterscheidet der Ansatz sieben Dimensionen, die sich Großteils – da sich die Orientierungsrahmen der Schulinspektionen zumeist auf Annahmen der Schuleffektivitätsforschung beziehen – mit den im Rahmen von Schulinspektionen zu evaluierenden Bereichen decken: 1) Organisationsstruktur, 2) Gemeinsame Ziel- und Wertvorstellungen und Kooperation im Kollegium, 3) Wissen und Fertigkeiten, 4) Führung und Management, 5) Qualitätssicherung, Zielüberprüfung und Feedback, 6) Austausch mit der schulischen Umwelt und 7) Partizipation der Lehrkräfte.

Mithilfe dieses Modells wird dann untersucht, inwieweit diese Kapazitäten einen Einfluss darauf haben, ob und, falls ja, wie Schulinspektion wirksam wird. Dabei wird davon ausgegangen, dass Schulen die über eine hohe Kapazität verfügen, die Fähigkeit besitzen, ihr eigenes Handeln und Wissen sowie ihre Normen und Werte kritisch zu hinterfragen und neuen Sichtweisen und Wissen gegenüber offen sind. Solche Schulen sind zu höherwertigem Lernen in der Lage; der Reflektion der eigenen Normen und Werte (double-loop-learning) und zur Reflektion der eigenen Lernprozesse (deutero learning). An Schulen mit geringen Kapazitäten hingegen, ist davon auszugehen, dass sie „lediglich" in der Lage sind, ihr eigenes Handeln innerhalb ihrer Normen und Werte kritisch zu hinterfragen und neuen Sichtweisen und Wissen kritisch oder gar verschlossen gegenüber stehen. Diese Schulen sind überwiegend von defensiven Routinen, Normen und Werten geprägt, die auf den Erhalt den Status Quo und nicht auf eine professionelle Weiterentwicklung ausgerichtet sind. Schulen, die über geringe Kapazitäten Organisationalen Lernens verfügen, dürften daher nur in sehr begrenztem Maße dazu in der Lage sein Rückmeldungen aus Schulinspektionsverfahren so zu verarbeiten, dass daraus sichtbare Impulse für die Schul- und Unterrichtsentwicklung resultieren.

Überprüft wird diese Annahme anhand von Schulentwicklungsberichten, in denen Schulleiterinnen und Schulleiter, deren Schulen einer zweiten Inspektion unterzogen wurden, darlegen, welche Veränderungen es seit der ersten Inspektion gegeben hat und welche Entwicklungsschritte bzw. Maßnahmen die Schulen eingeleitet haben sowie ob und wie diese umgesetzt wurden. Die Berichte sollen auf den zwischen Schule und Schulaufsichten getroffenen Ziel- und Leistungs-Vereinbarungen basieren. Um die schulischen Kapazitäten – also die innerschulische Ausgangslage – zu modellieren, wurden wiederum die Qualitätsberichte von Schulen aus dem ersten Inspektionszyklus zugrunde gelegt. Methodisch wurden in einem ersten Schritt die vorliegenden Schulentwicklungsberichte mithilfe einer quantitativen Inhaltsanalyse der hinsichtlich der umgesetzten Entwicklungsmaßnahmen analysiert. In einem zweiten Schritt wurden diese Daten dann mithilfe von Korrelations- und Wahrscheinlichkeitsanalysen mit den Kapazitäten organisationalen Lernens der Schulen in Zusammenhang gebracht.

Die Ergebnisse zeigen, dass a) Schulen infolge einer Inspektion Entwicklungsmaßnahmen generell vor allem im Bereich des Unterrichts sowie der Schul- und Unterrichtsorganisation ergreifen, wobei innerschulische Voraussetzungen diesbezüglich keinen Unterschied zwischen Schulen machen, b) an Schulen mit guten innerschulischen Voraussetzungen für die Schulentwicklung infolge einer Inspektion vor allem Maßnahmen im Bereich von Schulleitung und Schulmanagement ergriffen werden und dass c) Schulen mit geringen Kapazitäten infolge einer Inspektion vor allem auf möglichst

viele symbolische Maßnahmen setzen, die vor allem dazu beitragen, die Außenwirkung der Schule zu verbessern.

Beitrag 6: Pietsch, M. & Hosenfeld, I. (2017). Von der Schulinspektion zur Unterrichtsentwicklung: Welche Rolle spielt die Schulleitung. Empirische Pädagogik, 31(2), 202-220.

Beitrag 6 greift den Befund des vorhergehenden Beitrages auf und fokussiert auf die Rolle der Schulleitung im Rahmen der inspektionsgestützten Unterrichtsentwicklung. Dabei wird im Beitrag konkret untersucht, welchen Einfluss unterschiedliche Facetten des Führungshandelns (Transformationale, Instruktionale, Kollaborative Führung) von Schulleitungen in Relation zur Innovationskapazität, der Qualität der Kooperation und der Nutzung von Schulinspektionsrückmeldungen auf die Weiterentwicklung der Unterrichtsqualität besitzen.

Nach einer kurzen Darstellung von Führungsstilen sowie zu Annahmen über die Wirksamkeit von Schulleitungen werden dafür Daten aus dem in Niedersachen durchgeführten Projekt „Evaluation der Impulswirkung von Schulinspektionen und Vergleichsarbeiten auf die Qualitätsentwicklung an Schulen“ (EISVQS) reanalysiert. Für die Analyse wird auf ein logisches Modell zurück gegriffen, in dem einerseits angenommen wird, dass das Schulleitungshandeln bzw. Instruktionale, Transformationale und Kollaborative Führung durch Schulleitung direkt sowohl auf die Nutzung von Inspektionsdaten als auch die Unterrichtsentwicklung wirkt. Andererseits wird erwartet, dass auch vermittelte Effekte sichtbar werden, wobei diese sowohl durch die Innovationskapazität der Lehrkräfte als auch die Rahmenbedingungen, unter denen Datennutzung und Unterrichtsentwicklung erfolgen, moderiert werden. Internationalen Befunden zufolge kann das Schulleitungshandeln auf diesem Wege etwa 20 bis 30 Prozent der Variation in der Unterrichtsgestaltung durch Lehrkräfte bzw. der Unterrichtsentwicklung aufklären.

Mithilfe von Daten von n = 933 Lehrkräften aus 82 niedersächsischen Schulen wird dann ein Strukturgleichungsmodell spezifiziert. Die Gütekriterien für das Modell wiesen auf eine insgesamt sehr gute Passung hin. Dabei ließen sich rund 17 Prozent der Varianz in der inspektionsbasierten Unterrichtsentwicklung mithilfe der Modellvariablen erklären, jedoch nur drei Prozent der Unterschiede im Umgang mit Inspektionsergebnissen durch Lehrkräfte. Einen direkten und empirisch nachweisbaren Einfluss darauf, ob Lehrerinnen und Lehrer sich intensiv mit den Ergebnissen aus einer Inspektion auseinandersetzen, hat ausschließlich deren Innovationskapazität, also die Überzeugung, dass sie dazu in der Lage sind, den eigenen Unterricht aus eigener Kraft weiter zu entwickeln. Den größten direkten Einfluss darauf, ob infolge einer Schulinspektion der Unterricht an

einer Schule weiter entwickelt wird, hat die Auseinandersetzung der Lehrkräfte mit den Inspektionsergebnissen. Auch die Instruktionale Führung der Schulleitung hat einen nachweisbar positiv direkten Einfluss auf die Unterrichtsentwicklung, während sich für die Kooperation im Kollegium ein negativer direkter Effekt findet. Hinsichtlich der Entwicklung des Unterrichts lassen sich für alle anderen Modellvariablen keine direkten Effekte nachweisen, die gegen den Zufall abgesichert sind.

Die Befunde machen dann auch deutlich, dass es theoriekonform für die inspektionsgestützte Unterrichtsentwicklung eines Zusammenspiels von Instruktionaler und Transformationaler Führung bedarf, wobei letztere ausschließlich indirekt zum Tragen kommt. Auffällig ist weiterhin, dass die Auseinandersetzung der Lehrkräfte mit Inspektionsergebnissen von zwei Faktoren abhängt: Einen positiven Einfluss darauf hat einerseits, ob die Schulleitung im Sinne der Transformationalen Führung motivierende Leitbilder vorgibt und ob es ihr gelingt, die Entwicklungsbedarfe der Lehrkräfte zu erkennen und deren Potentiale systematisch zu fördern. Andererseits spielt es auch eine Rolle, ob Lehrkräfte sich befähigt fühlen, mit Veränderung und Innovationen umzugehen bzw. diese in den eigenen Unterricht zu implementieren, wenn es darum geht, dass sie sich intensiv mit den Inspektionsergebnissen auseinandersetzen.

Die vorgelegten Befunde machen deutlich, dass Schulinspektionen auf komplexe innerschulische Bedingungsgefüge treffen und eine gelingende Unterrichtsentwicklung infolge einer Inspektion nicht selbstverständlich ist. Dabei ist es vor allem vom Führungshandeln einer Schulleitung abhängig, ob Daten aus einer Inspektion durch Lehrkräfte überhaupt rezipiert werden und in welchem Maße der Unterricht weiterentwickelt wird. Wichtig ist es hierbei, dass Schulleitungen einerseits aktiv steuernd in das Unterrichtsgeschehen an ihrer Schule eingreifen, andererseits aber auch den Lehrkräften eine attraktive Vision und klare Orientierung geben. Dazu müssen sie kommunizieren, was mit Blick auf den Unterricht besser sein wird als bislang und wen sie diesbezüglich wie individuell fordern und fördern, kurz: die Schulleitungen müssen die Lehrerinnen und Lehrer trainieren und coachen und ihnen dadurch bei der Entwicklung und Entfaltung ihrer Potenziale und ihrer individuellen Stärken helfen.

Einordnung der unter 3.2. berichteten Beiträge

Die drei in diesem Abschnitt zusammen gefassten Beiträge fokussieren auf die Wirkungen von Schulinspektionen. Dabei steht die Variable der Schülerleistung im Mittelpunkt, von der

angenommen wird, dass diese über Maßnahmen der inspektionsbasierten Schul- und Unterrichtsentwicklung beeinflusst werden kann.

Die vorgelegten Untersuchungen zeigen, dass Schulinspektionen durchaus, wie intendiert, zu einer Entwicklung von Unterricht und Schule sowie zu verbesserten Schülerleistungen führen können. Jedoch ist die Analyse mutmaßlicher Effekte voraussetzungsreich und bedarf eines angemessenen Forschungsdesigns. Darüber hinaus zeigen die Befunde, erstmals für den deutschsprachigen Raum, dass innerschulische Voraussetzungen einen erheblichen Einfluss darauf haben, ob Schulinspektionen Wirkungen auf die Schul- und Unterrichtsentwicklung haben können.

Deutlich wird dabei, dass insbesondere die Schulleitung eine wichtige Rolle dabei spielt, ob und welche Maßnahmen infolge einer Schulinspektion ergriffen werden. Denn zum einen konnten wir zeigen, dass Schulleitungen infolge einer Inspektion vor allem dazu neigen sich selbst weiter zu professionalisieren. Zum andern ist der Einfluss von Schulleitungen auf die Entwicklung des Unterrichts durch Lehrkräfte hoch. Schulleitungen nehmen also eine zentrale Rolle innerhalb von Schulen ein, wenn es um die evidenzbasierte Entwicklung von Unterricht und Schule geht.

3.3. Schulleitungen als zentrale Akteure evidenzbasierter Schul- und Unterrichtsentwicklung

Die drei Beiträge in diesem Abschnitt fokussieren daher auf die Schulleitung als zentralen innerschulischen Akteur im Rahmen einer evidenzbasierten Schul- und Unterrichtsentwicklung. Zentrale Annahme ist, dass sich die Rolle von Schulleitungen im Rahmen des aktuellen Steuerungsparadigmas gewandelt hat und Schulleitungen daher zunehmend als Manager und Leader und weniger als Verwalter agieren müssen. Untersucht werden hier empirisch der Zusammenhang von Schulleitungshandeln mit der Qualität des Unterrichts sowie der Zusammenhang mit diesbezüglich vermittelnden Variablen. Dabei schließen die Arbeiten an den internationalen Diskurs zu wirksamen Schulleitungen an und gehen davon aus, dass Schulleitungen in erster Linie indirekt auf die Gestaltung des Unterrichts durch Lehrkräfte wirken.

Beitrag 7: Pietsch, M., Lücken, M., Thonke, F., Klitsche, S. & Musekamp, F. (2016). Der Zusammenhang von Schulleitungshandeln, Unterrichtsgestaltung und Lernerfolg: Eine argumentbasierte Validierung zur Interpretier- und Nutzbarkeit von Schulinspektionsergebnissen im Bereich Führung von Schulen. Zeitschrift für Erziehungswissenschaft, 19(3), 527-555.

Beitrag 7 greift die zentralen Befunde der oben angeführten Studien auf, führt die Abschnitte 3.1. und 3.2. zusammen und beschäftigt sich daher mit der Rolle, die Schulleitungen im Rahmen von Schulinspektionsverfahren zugeschrieben wird. Im Zentrum der Arbeit wird dabei der Frage nachgegangen, wie valide Schulinspektionsergebnisse im Bereich Führung von Schulen interpretiert und als Grundlage für eine mögliche Schulentwicklung genutzt werden können. Im Beitrag wird dieser Frage anhand einer Stichprobe von n = 37 Sekundarschulen, n = 1663 Lehrkräften und der Leistungsentwicklung von n = 23.943 Schülerinnen und Schülern nachgegangen. Die Klärung dieser Frage ist auch daher von besonderer Bedeutung, da die Qualitätsentwicklung der Einzelschule zunehmend durch die Ergebnisse externer Evaluationen bestimmt wird. Nicht-valide Inspektionen führen zu fehlerhaften Bewertungen und Urteilen, die wiederum zu fehlerhaften administrativen und/oder politischen Entscheidungen führen, die dann keine oder im schlimmsten Fall gar negative Effekte nach sich ziehen.

Im Rahmen der Arbeit wird dabei davon ausgegangen, dass Schulinspektionsverfahren resp. Qualitätsrahmen in starkem Maße auf Befunde und Annahmen der Schuleffektivitätsforschung rekurrieren. In einem ersten Schritt werden daher zentrale Annahmen zum Thema Schulleitungshandeln und Wirksamkeit von Schulleitungen herausgearbeitet. Schulleitungen, die zum Bildungserfolg der Schülerinnen und Schüler an ihren Schulen beitragen, legen demnach die Schwerpunkte ihrer Arbeit dabei vor allem darauf, 1) den Schulbeteiligten Wege und Ziele vorzugeben, 2) Mitarbeiterinnen und Mitarbeiter (weiter) zu entwickeln, 3) die Schule (neu) zu gestalten und 4) das Lernen und Lehren an der Schule aktiv zu steuern. Die Schuleffektivitätsforschung unterscheidet dabei zwei Konzepte voneinander: pädagogische bzw. instruktionale und transformationale Führung. Instruktionale Führung umfasst dabei vor allem Managementaspekte; sie erfolgt primär aufgaben- und produktorientiert, zielt auf die Optimierung vorhandener Strukturen und Prozesse ab und führt im Idealfall zu einer Verbesserung bereits vorhandener Prozesse und Mechanismen. Die Schulleitung kontrolliert und koordiniert entsprechend gezielt Aspekte des Schul- und Unterrichtsgeschehens, die den Lernfortschritt der Schülerinnen und Schüler betreffen, und nimmt direkten Einfluss auf den Unterricht und das Curriculum, z. B. durch die aktive Anleitung von Lehrkräften mittels Zielvorgaben, abgestimmte Fortbildungsmaßnahmen und der Evaluation von Schülerleistungen. Transformationale Führung hingegen beinhaltet in der Regel

Führungsaspekte, erfolgt meist mitarbeiterorientiert und zielt auf die nachhaltige Veränderung der schulischen Lern- und Arbeitskultur ab. Dieser Führungsstil soll daher primär zu innerschulischen Innovationen und Veränderungen führen und ist maßgeblich dadurch geprägt, dass die Schulleitung eine sinnstiftende Zukunftsvision für die Schule entwickelt, Lehrkräfte inspiriert und motiviert, einzelne Lehrerinnen und Lehrer gezielt unterstützt und fördert sowie ihnen intellektuelle Herausforderungen bietet. Mit Blick auf Wirkungen zeigt sich, dass Schulleitungen in erster Linie indirekt, vermittelt über schulische, personale und unterrichtliche Faktoren, Wirkung auf Schülerleistungen entfalten und dass es insbesondere ein instruktionaler Führungsstil ist, der mit Schülerleistungen zusammen hängt.

Im Anschluss an diese konzeptionelle Vorarbeit werden die Annahmen für die Schulinspektion Hamburg mithilfe von Daten zum Schulleitungshandeln, die im Rahmen der Hamburger Schulinspektion mithilfe von international gängigen Instrumenten aus dem Bereich der Schuleffektivitätsforschung erhoben wurden, und anhand von Leistungsdaten der Hamburger KERMIT-Erhebungen validiert. In einem ersten Schritt wurde dafür ein mehrschrittiges, sequentielles Verfahren der Pfadmodellierung eingesetzt. Dieses Vorgehen erlaubt es, da das Pfadmodell im Gesamtmodell geschachtelt ist, den Gesamtfit des Modells in die Komponenten des Mess- sowie des Pfadmodells zu unterteilen und so zu prüfen, wie tragfähig das theoretisch postulierte logische Modell ist. Hierfür wurden erstens die spezifischen Messmodelle der einzelnen latenten Variablen separat geschätzt. Zweitens wurden dann die so ermittelten Item-Parameter fixiert und ein Gesamtmessmodell geschätzt. Drittens wurde anschließend das theoretisch beschriebene Pfadmodell spezifiziert.

Die Befunde zeigen, dass das Schulleitungshandeln einen sehr großen Einfluss auf die Arbeitsbedingungen von Lehrkräften ($R^2>.65$) hat. Auch die Arbeitszufriedenheit und das Commitment der Lehrkräfte lassen sich in weiten Teilen durch das Leitungshandeln in Verbindung mit den schulischen Rahmenbedingungen erklären ($R^2>.58$). Die Innovationskapazität der Lehrkräfte, also die durch sie berichtete Befähigung, Neuerungen in den eigenen Unterricht zu implementieren, hängt hingegen maßgeblich davon ab, wie stark die Kooperation im Kollegium ausgeprägt ist (0,60). Aber auch direkte Effekte des Schulleitungshandelns lassen sich nachweisen. So haben hier sowohl die Instruktionale als auch die Transaktionale Führung einen nachweisbaren Einfluss darauf, ob Lehrkräfte sich in der Lage sehen, Neuerungen und Innovationen in ihren Unterricht zu implementieren. Das Unterrichtshandeln der Lehrkräfte wiederum wird sowohl direkt durch das Schulleitungshandeln als auch indirekt über die Arbeitsbedingungen sowie die Innovationskapazität beeinflusst.

In einem zweiten Schritt wurde das spezifizierte Modell mit den oben genannten Leistungsdaten in Zusammenhang gebracht. Hierfür wurden zuerst für jeden Schüler bzw. jede Schülerin individuelle Leistungsentwicklungen in den Lese- sowie den Mathematiktests der KERMIT-Erhebungen in einem Zweischuljahres-Zeitraum (Klassenstufe 5 auf 7 bzw. 7 auf 9) ermittelt. Insgesamt lagen für alle 37 Schulen Leistungsentwicklungen von n = 23.943 Schülerinnen und Schülern für die Domänen Mathematik und Lesen und somit insgesamt n = 47.886 Lernentwicklungsinformationen vor. Anschließend wurden diese Informationen je Test auf Ebene der Schule aggregiert. Pro Schule lagen demnach folgende 12 Leistungsinformationen vor: 1) Leistungsentwicklung jeweils in Mathematik und Deutsch Leseverstehen von der Klassenstufe 5 auf 7 für die Schuljahre 2010/11 auf 2012/13, 2011/12 auf 2013/14 und 2012/13 auf 2014/15, 2) Leistungsentwicklung in Mathematik und Deutsch Leseverstehen von der Klassenstufe 7 auf 9 für die Schuljahre 2010/11 auf 2012/13, 2011/12 auf 2013/14 und 2012/13 auf 2014/15. Diese Daten wurden im Folgenden so zusammen gefasst, dass indiziert wurde, ob es sich bei einer Schule um eine hoch-performante handelt, um deine Schule, an der sich wiederholt auffallend hohe Leistungszuwächse finden. Anschließend wurden die Pfadmodelle bzw. deren Koeffizienten an den so erzeugten zwei Schultypen mithilfe eines Propensity-Score-Matchings (unter Kontrolle der sozialen Schülerzusammensetzung) miteinander verglichen.

Das Ergebnis zeigt, dass nur an leistungsstarken Schulen ein direkter Einfluss der Instruktionalen Führung der Schulleitungen auf die Unterrichtsgestaltung von Lehrkräften feststellbar ist, wobei dieser Einfluss durch indirekte Effekte noch deutlich verstärkt wird. Auffällig ist darüber hinaus, dass an den Schulen der Vergleichsgruppe vor allem Transformationale Führung eine Rolle spielt, die jedoch einen negativen Effekt auf die Unterrichtsgestaltung nach sich zieht. Während Schulleitungen an weniger leistungsstarken Schulen Instruktionale Führung mit Transformationaler und Transaktionaler Führung koppeln, um Einfluss auf die Innovationskapazität der Lehrkräfte zu nehmen, flankieren Schulleitungen an leistungsstarken Schulen ihre Instruktionalen Führungspraktiken durch ein grundsätzlich aktives Schulleitungshandeln in Verbindung mit einem Laissez-faire-Stil. Auch mit Blick auf die Arbeitsbedingungen der Lehrerinnen und Lehrer ist das Leitungshandeln an hoch performanten Schulen deutlicher durch einen instruktionalen Führungsstil geprägt als an den Schulen der Vergleichsgruppe, wobei jedoch auch deutlich wird, dass an ersteren Schulen der gesteigerte Einfluss dieses Führungsstils mit einem ebenfalls stärkeren Einfluss Transformationaler Führung einhergeht. Insgesamt zeigt sich damit theoriekonform, dass Schulleitungen an leistungsstarken Schulen häufiger instruktional führen als Schulleitungen an weniger leistungsstarken Schulen.

Grundsätzlich kommen wir daher zu dem Schluss: Schulleitungen nutzen unterschiedliche Führungsstile, sie wirken auf den Unterricht von Lehrkräften sowohl direkt als auch indirekt und insbesondere der Aspekt der Innovationskapazität von Lehrkräften spielt eine entscheidende Rolle als Mediator. Darüber hinaus führen Schulleitungen an hoch performanten Schulen auch häufiger instruktional und die Lehrkräfte an diesen Schulen legen besonderen Wert auf das Klassenmanagement – dies alles ist theoriekonform. Jedoch ist dies allein für das intendierte Ziel der inspektionsbasierten Schulentwicklung nicht hinreichend. Denn eine Verallgemeinerung bzw. Extrapolation ist wiederum nicht oder nur begrenzt möglich. So lassen sich zwar für die einzelnen Skalen die theoretisch postulierten und empirisch bekannten Zusammenhänge mit anderen Konstrukten nachweisen. Diese variieren jedoch in Stärke, Einfluss und Komplexität zwischen verschiedenen Kontexten, was darauf hindeutet, dass Schulleitungen an Schulen mit hoher Performanz grundsätzlich anders führen als ihre Kolleginnen und Kollegen an Schulen mit weniger hohen Lernzuwächsen. Mit anderen Worten: Schulleitungshandeln erfolgt weniger standardisiert als vielmehr situativ. Dies wiederum macht es in der Konsequenz dann auch schwierig aus einem standardisierten und mit Blick auf den Kontext invarianten Stärken-Schwächen-Profil, wie es im Rahmen von Schulinspektionen genutzt wird, zielgerichtete Entscheidungen zur Schulentwicklung zu treffen, die dann letztlich in verbesserte Schülerleistungen münden.

Die Befunde machen dabei zweierlei deutlich: Einerseits können quantitative, bewährte Instrumente zur Ermittlung von Schulleitungshandeln durch die Befragung von Lehrkräften belastbar eingesetzt werden, um erste, vergleichbare Diagnosen zur Qualität der Führung an Schulen zu ermöglichen. Andererseits sind diese Informationen als Grundlage für eine wissensbasierte Schulentwicklung jedoch nicht hinreichend, wenn sie, wie im Rahmen von Inspektionen üblich, primär im Rahmen eines Stärken-Schwächen-Profils verarbeitet und berichtet werden. Vielmehr bedarf es im Rahmen von Inspektionen einer umfassenderen Auseinandersetzung mit den Bedingungen vor Ort sowie der Analyse und Vermittlung der komplexen Zusammenhänge und Interaktionen zwischen einzelnen Merkmalen auf Ebene der Schule. Nur so können mit Hilfe von zielgerichteten und aufeinander abgestimmten Veränderungen die intendierten Entwicklungen auch wirklich mit hoher Wahrscheinlichkeit eintreten.

Beitrag 8: Pietsch, M. & Tulowitzki, P. (2017). Disentangling School Leadership and its Ties to Instructional Practices. An empirical Comparision of various Leadership Styles. School Effectiveness and School Improvement, 28(4), 629-649.

Beitrag 8 richtet sich an eine internationale Leserschaft und geht dabei noch einmal genauer bzw. differenzierter auf die Rolle von Schulleitungen im Rahmen einer evidenzbasierten Unterrichtsentwicklung ein und prüft den Zusammenhang von Schulleitungshandeln und der Unterrichtsgestaltung von Lehrkräften anhand einer Stichprobe von n=3.746 Lehrkräften an n=126 Schulen aller Schulformen in Hamburg. Hierbei kommt wiederum ein Strukturgleichungsmodell zum Einsatz, in dem sowohl direkte als auch indirekte Effekte des Schulleitungshandelns berücksichtigt werden. Dabei greift die Arbeit das Desiderat auf, dass bislang auch auf internationaler Ebene keine Forschungsarbeiten vorliegen, die (indirekte und direkte) Effekte unterschiedlicher Führungsstile von Schulleitungen auf die Unterrichtsgestaltung von Lehrkräften untersuchen.

Entsprechend wurde, wie auch in Beitrag 7, ein mehrschrittiges, sequentielles Verfahren der Pfadmodellierung eingesetzt. Ein zentraler Aspekt dieses Beitrages war dabei die Evaluation verschiedener Messmodelle zur Messung von Schulleitungshandeln mithilfe etablierter Skalen. Hierfür wurden verschiedene Modelle, die das Leitungshandeln beschreiben, empirisch miteinander verglichen. Dabei wies ein so genanntes Bi-Faktormodell den besten Modellfit auf. Daher wurde Schulleitungshandeln für die weitere Untersuchung als Bi-Faktormodell geschätzt. Bi-Faktormodelle sind vor allem dann geeignet, wenn es darum geht, herauszufinden, ob es a) einen Generalfaktor gibt, auf dessen Existenz sich Kommunalitäten zwischen Items zurückführen lassen und b) verschiedene, voneinander abgrenzbare Subfaktoren gibt, die einen eigenen, vom Generalfaktor unabhängigen Einfluss auf abhängige Variablen im Pfadmodell haben und c) Effekte sowohl des General- als auch der einzelnen Subfaktoren im Modell erwartet werden. In der Modellierung wird entsprechend angenommen, dass es einen allgemeinen Generalfaktor g gibt, der sich auf alle beobachteten Führungsindikatoren auswirkt und einzelne Führungsfacetten, die dann unabhängig voneinander Einfluss auf spezifische Indikatoren haben und sich zueinander ebenso wie zum Generalfaktor orthogonal verhalten.

Auch hier ging es, wie in Beitrag 7, im Weiteren darum zu prüfen, wie Schulleitungen auf die Unterrichtsgestaltung von Lehrkräften wirken. Die Befunde zeigen, entgegen der international gängigen Annahme, dass Schulleitungen in erster Linie direkt auf den Unterricht von Lehrkräften wirken. Dabei erklärt das Führungshandeln 10 bis 21 Prozent in der Variation der Unterrichtsgestaltung – dies deckt sich wiederum mit dem internationalen Forschungsstand. Dabei wirkt eine instruktionale Führung auf alle drei Basisdimensionen des Unterrichts, die in den Analysen

berücksichtigt wurden – Klassenmanagement, unterstützendes Lernklima und kognitive Aktivierung. Weiterhin zeigt sich aber auch, dass die Beeinflussung eines kognitiv aktivierenden Unterrichts durch die Schulleitung einen Führungsstilmix benötigt. Während Schulleitungen auf Klassenmanagement und unterstützendes Lernklima vor allem durch einen instruktionalen Führungsstil Einfluss ausüben, werden zur Veränderung eines kognitiv aktivierenden Unterrichts durch die Schulleitung auch Aspekte transformationaler, transaktionaler sowie passiver Führung benötigt.

Die Ergebnisse zeigen somit erstmals empirisch – dies auch auf internationaler Ebene –, dass einzelne Führungsstile, unter Kontrolle weiterer Führungsstile, differentielle sowie gemeinsame Effekte auf die Unterrichtsgestaltung von Lehrkräften ausüben können und dass diese Effekte, bei Kontrolle einer Vielzahl von Kovariaten, weniger indirekt als direkt sind.

Beitrag 9: Pietsch, M., Tulowitzki, P. & Koch, T. (angenommen). On the Differential and Shared Effects of Leadership for Learning on Teachers' Organizational Commitment and Job Satisfaction: A multilevel perspective. Educational Administration Quarterly.

In Beitrag 9, der sich ebenfalls an eine internationale Leserschaft richtet, wird anhand der bereits in Beitrag 8 genutzten Daten detailliert der Frage nachgegangen, welcher Zusammenhang zwischen Schulleitungshandeln und der Arbeitszufriedenheit sowie dem Commitment von Lehrkräften empirisch beobachtbar ist. Im Rahmen des Beitrages wird dabei vom Konzept des lernzentrierten Schulleitungshandelns – international auch Leadership for Learning genannt – ausgegangen und mithilfe von Mehrebenenstrukturgleichungsmodellen die gemeinsamen und differentiellen Effekte einzelner Führungsstile auf der Individual- sowie auf der Schulebene untersucht. Leadership for Learning zeichnet sich dabei dadurch aus, dass der Fokus der Schulverantwortlichen auf dem Lernen und dem Lehren an einer Schule liegt, Führungsverantwortung geteilt wird und darüber hinaus alle Schulbeteiligten zu einer Veränderung ihres Verhaltens und ihrer Lern- und Leistungsbereitschaft motiviert und inspiriert werden. Das Konzept des lernzentrierten Schulleitungshandelns überwindet dabei die Schwächen des Konzepts der instruktionalen Führung, insbesondere dessen Fokussierung auf die Person der Schulleiterin bzw. des Schulleiters sowie auf das Lehren, indem weitere relevante Führungsaspekte und Verhaltensweisen thematisiert und berücksichtigt werden, die relevant für die Lernerfolge von Schülerinnen und Schüler, aber auch für das Lernen auf allen Ebenen der Institution Schule sind. Entsprechend umfasst die Idee des Leadership for Learning Aspekte instruktionaler, transformationaler und geteilter Führung. Anders als in Beitrag 9 werden daher die Transaktionale sowie die Laissez-faire-Führung im Rahmen dieses Beitrages nicht als Leadership-Facetten in der empirischen Modellierung berücksichtigt.

Methodisch wird ein doppelt-latentes (doubly-latent) Strukturgleichungsmodell genutzt. Dies vor dem Hintergrund, dass es sich bei Führung in der Regel um ein Mehrebenenkonstrukt handelt und daher sowohl auf Ebene der Schule, gerichtet an alle Mitarbeiterinnen und Mitarbeiter, auf Ebene einzelner innerschulischer Gruppierungen aber auch zwischen einzelnen Lehrkräften und einer Schulleitung stattfinden kann, empirisch4 Untersuchungen, die Schulleitungshandeln mehrebenenanalytisch in den Blick nehmen dennoch äußert rar sind. Hinzu kommt, dass zwar für alle Leadership-Facetten des Leadership for Learning-Modells Befunde vorliegen, die zeigen, dass diese, jeweils für sich genommen, mit der Arbeitszufriedenheit und/ oder dem Commitment von Lehrkräften zusammen hängen. Eine Untersuchung jedoch, die die mutmaßlichen Effekte einzelner Facetten unter Kontrolle der anderen Führungsfacetten sowie deren gemeinsame Wirkung betrachtet, liegt bislang nicht vor. Da die empirischen Zusammenhänge zwischen den einzelnen Facetten jedoch teilweise sehr hoch sind (r zwischen .55 und .92), stellt sich insbesondere die Frage nach der Spezifität potentieller Effekte.

In einem ersten Schritt wurde daher im Rahmen des Beitrages geprüft, ob es sich bei den Führungskonstrukten wie angenommen um Mehrebenenkonstrukte handelt. Hierfür wurden Intraklassenkorrelationen sowie rwg-Indices (als Maß der Übereinstimmung) berechnet. Die diesbezüglichen Analysen zeigen, dass es sich bei allen drei Führungsfacetten – instruktional, transfromational, geteilt – sowie deren Subdimensionen um Mehrebenenkonstrukte handelt – so liegen die Werte der ICC(1) zwischen .24 und .37, die der ICC(2) zwischen .89 und .95 und die Werte der rwg-Indices zwischen .67 und .89. Somit liegt bis zu 37 Prozent der Varianz zwischen Schulen, und die Lehrkräfteantworten zum Schulleitungshandeln sind auf Schulebene sowohl hoch reliabel als auch übereinstimmend.

In einem zweiten Schritt wurden konfirmatorische Faktorenanalysen auf zwei Ebenen durchgeführt, um Messmodelle zum Führungshandeln auf Ebene der Schule und auf Ebene einzelner Lehrkräfte zu evaluieren. Diese Analyse zeigt, dass ein Modell, in dem auf beiden Ebenen ein möglichst ausdifferenziertes Modell mit sieben Subdimensionen des Führungshandelns genutzt wird, am besten auf die genutzten Daten passt.

In einem dritten Schritt wurden dann die gemeinsamen und differentiellen Effekte dieser Facetten auf die Arbeitszufriedenheit und das Commitment von Lehrkräften untersucht. Diesbezüglich zeigen die Befunde, dass beide Variablen in besonders starkem Maße mit dem Aspekt der geteilten Führung zusammenhängen – und dies sowohl auf der Individual- als auch der Schulebene. Darüber hinaus ließ

sich auf individueller Ebene noch ein Zusammenhang mit den abhängigen Variablen nachweisen, wenn Schulleitungen sich als Mentor oder Coach ihrer Mitarbeiterinnen und Mitarbeiter verstehen und auf deren individuelle Bedürfnisse eingehen (Individual Consideration). Auffällig ist, dass instruktionale Führung, anders als in der Forschung bisher berichtet, unter Kontrolle anderer Führungsfacetten, in keinem nachweisbaren Zusammenhang mit den beiden abhängigen Variablen steht, jedoch als einzige Führungsfacette systematisch mit dem sozialen Kontext einer Schule variiert. Dabei erklären die Führungsfacetten zusammen 54 bzw. 62 Prozent der Varianz in den abhängigen Konstrukten auf Individualebene und 67 bzw. 78 Prozent auf Schulebene.

Zusammen genommen zeigen die Befunde dieser Studie, dass instruktionale Führung, anders als häufig angenommen, fast keine nachweisbaren Effekte auf die Arbeitszufriedenheit und das Commitment von Lehrkräften hat und dass insbesondere eine Teilung von Führungsverantwortung positive Effekte nach sich zieht. Dies, dieser Befund ist ebenfalls neu, zeigt sich in etwa gleichem Maße auf Individual- wie auf Schulebene. Darüber hinaus wird deutlich, dass die gemeinsamen Effekte des lernzentrierten Schulleitungshandelns in etwa bei einer Effektstärke von d=2.16 bzw. d=2.52 und damit mehr als doppelt so hoch liegen wie für einzelne Facetten bislang festgestellt. Die Analysen verdeutlichen somit, dass es relevant ist, Schulleitungshandeln als Mehrebenenkonstrukt zu verstehen und zu analysieren und darüber hinaus in empirischen Arbeiten neue Modelle des Schulleitungshandelns zu berücksichtigen, die über das traditionelle Verständnis instruktionaler Führung hinaus gehen.

Einordnung der unter 3.3. berichteten Beiträge

Die drei Beiträge dieses Abschnitts nehmen die Schulleitung als zentralen Akteur im Rahmen einer evidenzbasierten Schul- und Unterrichtsentwicklung in den Blick. Dies vor dem Hintergrund, dass sich deren Rolle seit der Einführung der Neuen Steuerung im Bildungssystem zunehmend vom Primus inter Pares zum Manager und Leader gewandelt hat und Schulleitungen durch die Einführung neuer Steuerungsmechanismen in den letzten Jahren gegenüber Lehrkräften strukturell stark aufgewertet wurden.

Die vorgelegten Befunde machen dann auch deutlich, dass Schulleitungen an Schulen in Deutschland einen nachweisbaren Einfluss auf die Gestaltung von Unterricht und Schule und vermittelt hierüber auf Schülerleistungen haben. Diese Befunde decken sich mit der internationalen Befundlage und zeigen, dass insbesondere eine Führung, die den Unterricht in den Blick nimmt (Instruktionale Führung) nachweisbare Effekte auf die Gestaltung von Unterricht durch Lehrkräfte hat. Gleichwohl

machen die Befunde auch deutlich, dass dieses Konzept für sich genommen nicht ausreicht, um wirksames Schulleitungshandeln zu beschreiben. Zum einen, dies zeigen die vorgelegten Analysen, bedarf es zur Beeinflussung eines kognitiv aktivierenden Unterrichts eines komplexen Führungsstilmixes. Zum anderen hat eine solche Führung kaum Einfluss auf vermittelnde Variablen, wie z.B. das Commitment von Lehrkräften.

Gleichwohl zeigen die Analysen sehr deutlich, dass Führungspraktiken bzw. die Zusammenhänge von Führung mit weiteren Merkmalen in der Schule unter Umständen kontextsensitiv sind und somit der Kontext einer Schule einen eigenen Einfluss auf die evidenzbasierte Schul- und Unterrichtsentwicklung haben kann.

3.4. Kontexte evidenzbasierter Schul- und Unterrichtsentwicklung

Die letzten vier Beiträge untersuchen die Rolle schulischer Kontexte im Rahmen des aktuellen Steuerungsparadigmas. Abgestellt wird dabei einerseits auf den sozialen Kontext und andererseits auf den Wettbewerbskontext von Schulen. Dabei beschäftigen sich zwei der Beiträge eher methodisch-konzeptionell mit der Thematik, zwei weitere analytisch. Die Annahme lautet, dass Kontextbedingungen nachweisbare Effekte auf inner-schulische Variablen nach sich ziehen und insofern im Rahmen Neuer Steuerung eine entsprechende Rolle spielen. Dabei ist insbesondere der Wettbewerb zwischen Schulen eine zentrale Annahme im Rahmen des aktuellen Paradigmas: In einem Wettbewerbsmarkt können Schulen ihre Klientel nicht mehr länger als gegeben annehmen und müssen durch qualitativ hochwertige Angebote um Schülerinnen und Schüler werben, kurz: Wettbewerb zwischen Schulen erzeugt Bildungsqualität.

Beitrag 10: Schulte, K., Hartig, J. & Pietsch, M. (2016). Berechnung und Weiterentwicklung des Sozialindex für Hamburger Schulen. In B. Groot-Wilken, K. Isaac & J-P. Schräpler (Hrsg.), Sozialindices für Schulen – Hintergründe, Methoden und Anwendung (S. 157-172). Münster: Waxmann.

Dieser Beitrag wurde der Vollständigkeit halber beigefügt und bildet im Rahmen der vorliegenden Arbeit auch deswegen einen Sonderfall, da er Vorarbeiten aufgreift, die ich bereits vor längerer Zeit verfasst habe. Konkret wird in diesem Beitrag der Hamburger Sozialindex, auch KESS-Index genannt, revidiert und weiterentwickelt, der ursprünglich im Rahmen der Studie KESS 4 von mir entwickelt

wurde[1]. Durch den Einsatz von Sozialindizes sollen Hamburger Schulen in schwierigen Lagen mit zusätzlichen Ressourcen unterstützt werden, um Effekte der Schülerzusammensetzung kompensieren und chancenausgleichend wirken zu können: Gleiche Bildungschancen sollen mit ungleichem Mitteleinsatz erreicht werden. In Hamburg beschreibt der Sozialindex, seit 2006 basierend u.a. auf der Kapitaltheorie von Bourdieu, die sozialen Rahmenbedingungen der Schulen. Die auf dem Index basierende Zuordnung zu sechs abgestuften Belastungsgruppen hat Auswirkungen auf diversen Ebenen: Insbesondere determiniert der Sozialindex unterschiedliche Ressourcenallokationen (z.B. kleinere Klassenfrequenzen oder höhere Sprachfördermaßnahmen für Schulen mit niedrigeren Indizes). Für eine breitere Fundierung des Sozialindex über die hinsichtlich der verschiedenen theoretischen Facetten auf Schülerebene erfassten Indikatoren nach Bourdieu hinaus wurden zusätzlich Daten des Statistikamts Nord herangezogen. Dabei handelt es sich um Sozialraumdaten (z.B. die Arbeitslosenquote), die auf Ebene Statistischer Gebiete vorliegen. Diese Daten können gerade auch bei hohen Datenausfällen oder möglichen Verzerrungen durch selektives Beantworten wertvolle Informationen liefern.

Auf Basis einer Stichprobe von n = 24.452 aus allen staatlichen Hamburger Schulen, wurde der damals vorhandene Hamburger Sozialindex überarbeitet. Dafür wurden exploratorische und konfirmatorische Faktorenanalysen mit aggregierten Daten auf Schulebene durchgeführt. Um einen direkten Vergleich der Belastungswerte von Grund- und Sekundarschulen zu ermöglichen, war das Ziel, ein Modell für alle Schulen zu berechnen, und nicht wie bisher zwei separate Skalierungen mit unterschiedlichen Variablensätzen durchzuführen. Dem liegt die Forderung zugrunde, dass das Konstrukt sozialer Belastung über Schulformen hinweg konsistent definiert wird, d.h. Belastungsmerkmale für Kinder an Grundschulen grundsätzlich die gleichen sind wie solche für Schülerinnen und Schüler an weiterführenden Schulen. In methodischer Hinsicht wurde mit dem Einsatz von Faktorenanalysen der ursprüngliche, im Rahmen von KESS 4 etablierte Ansatz, wieder aufgegriffen und die Partial-Credit-Skalierung, die vom Institut für Schulentwicklungsforschung in den Zwischenjahren angewandt wurden wieder verworfen.

Mit insgesamt 53 Variablen wurde eine explorative Faktorenanalyse mit obliquer Bi-Geomin-Rotation berechnet. Dieses Rotationsverfahren wurde gewählt, um einen varianzstarken Generalfaktor zu erhalten, der möglichst viele gemeinsame Unterschiede zwischen den Schulen abbildet. Da nicht erwartet wurde, dass eine einzige Dimension zur Beschreibung der Daten ausreicht, wurden

[1] Pietsch, M., Bonsen, M. & Bos, W. (2006). Ein Index sozialer Belastung als Grundlage für Rückmeldungen von Leistungsergebnissen an Schulen und Klassen und für 'faire Vergleiche' von Grundschulen in Hamburg. In W. Bos & M. Pietsch (Hrsg.), KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen (S.225-245). Münster: Waxmann.

zusätzlich korrelierte Residualfaktoren zugelassen. Diese können durch den Generalfaktor nicht repräsentierte, variablenspezifische Abhängigkeiten abbilden, waren jedoch für das verfolgte Ziel einer eindimensionalen Beschreibung der Belastung inhaltlich nicht von Interesse. Auf Basis des aus der explorativen Faktorenanalyse gewonnenen Modells wurde dann ein konfirmatorisches Strukturgleichungsmodell mit einem Generalfaktor und sieben korrelierten Residualfaktoren spezifiziert. Hierbei wurden für die Residualfaktoren die jeweils höchsten Ladungen aus dem explorativen Modell freigesetzt, alle weiteren Ladungen wurden auf null fixiert.

Auf Grundlage dieses ersten Modells wurden nach den folgenden Kriterien 24 Variablen ausgewählt: Jeder der theoretisch angenommenen Bereiche sollte mit drei bis vier Indikatoren abgebildet sein, ausgewählt wurden zunächst die Variablen, welche innerhalb jedes Bereichs die jeweils höchsten Ladungen auf dem Generalfaktor aufwiesen. Darüber hinaus sollten drei bis vier Variablen der Daten des Statistikamts Nord in die Berechnung eingehen. Die Variablen aus der Fragebogenerhebung sollten dabei untereinander nicht redundant sein: war die Variable bei Eltern und Schülerinnen und Schülern vorhanden, wurde auf die Elternfrage zurückgegriffen. Wenn eine Variable aus den oben genannten Gründen die Eigenschaft eines Elternteils beschreibt, sollte immer auch der andere Elternteil herangezogen werden. Auch die sozialen Raumdaten wurden theoriegeleitet den Dimensionen zugeordnet, so zählten Arbeitslosigkeit und Hilfebedürftige Kinder zu dem ökonomischen Kapital, die Wahlbeteiligung wäre als soziales Kapital zu verorten.

Zur Überprüfung und Spezifikation des finalen Modells wurde mit den ausgewählten 24 Variablen eine weitere explorative Faktorenanalyse berechnet, die als Zwischenschritt dazu diente, die Faktorenspezifikationen der abschließenden konfirmatorischen Faktorenanalyse über die höchsten Faktorladungen der Variablen auf den jeweiligen Faktoren herzuleiten. Nachdem die Modellgütemaße bei der Analyse der 24 verbliebenen Variablen für ein Modell mit sechs Faktoren sprachen, wurde ein Generalfaktormodell mit fünf unkorrelierten Residualfaktoren spezifiziert, welches einen akzeptablen Gesamtfit zeigt. Auf Basis dieses Modells wurden Faktor-Scores für den Generalfaktor geschätzt, diese werden als geschätzter Belastungswert der Schulen verwendet. Die so ermittelten Sozialindices korrelierten mit den bislang vorliegenden hoch ($r>.90$), was auf eine hohe zeitliche Stabilität sozialer Belastungsfaktoren auf Ebene der Einzelschule hindeutet.

Dieser Beitrag ist insofern relevant, als im Rahmen der Arbeit ein Maß entwickelt bzw. überarbeitet wird, dass nicht nur für die Ressourcenallokation relevant ist, sondern das auch bei der Bildung repräsentativer Stichproben im Rahmen von wissenschaftlichen Untersuchungen und Evaluationen in Hamburg (z.B. bei der Auswahl einer repräsentativen Kernstichprobe von Schulen pro Schuljahr für

die Schulinspektion), bei der Berechnung und Rückmeldung von Vergleichswerten („fairer Vergleich“) für die schulbezogenen Ergebnisrückmeldungen im Rahmen von KERMIT oder bei der Bildung von Vergleichsgruppen im Kontext der Bildungsberichterstattung genutzt wird. Dabei hat der Index, anders als andere, vergleichbare Indices eine theoretische Anbindung und vereint darüber hinaus Individual- als auch sozialräumliche Daten in einem Modell. Nicht zuletzt weisen die Analysen auf die starke Persistenz sozialer Disparitäten an Schulen in Hamburg hin und zeigen, dass selbst über 10 Jahre hinweg kaum Veränderungen in der sozialen Zusammensetzung der Schülerschaft auf Schulebene nachweisbar sind.

Beitrag 11: Leist, S. & Pietsch, M. (2017). Bordering the Area of Spatial Relevance for Schools: A stochastic Network Approach using the Example of Hamburg, Germany. Belgeo - Revue Belge de Géographie, 2-3/2017 (Special Issue: Une géographie sociale de l’enseignement). OnlineFirst: https://doi.org/10.4000/belgeo.20332

Beitrag 11, der sich ebenfalls an eine internationale Leserschaft richtet, betrachtet das Thema Bildungsmärkte und damit einhergehend die Aspekte Wettbewerb zwischen Schulen und soziale Segregation anhand einer Vollerhebung von n = 14.032 Schülerverlaufsdaten an n = 400 Schulen im Stadtstaat Hamburg. Dabei befasst sich der Beitrag mit einem innovativen methodischen Ansatz zur Modellierung von Bildungsmärkten, der die Nachteile bisheriger Ansätze zur Marktabgrenzung aufwiegen kann sowie mit der Frage, wie sich der Wettbewerb zwischen Schulen darstellt. Nachdem regionale und lokale Bildungsmärkte mittels des empirischen Ansatzes eingegrenzt werden, werden der Wettbewerb zwischen weiterführenden Schulen um Schülerinnen und Schülern sowie soziale Segregationsprozesse beim Übergang von den Grundschulen in die weiterführenden Schulen analysiert.

Dabei betritt der Beitrag für den deutschsprachigen Raum Neuland. Denn während der Aspekt der Ouputsteuerung, die stardards-based reforms, im deutschsprachigen Raum in den vergangenen Jahren zunehmend Aufmerksamkeit aus dem Bereich der empirischen Bildungsforschung erfahren hat, wurden Wettbewerbsmechanismen, die choice policies, in der empirischen Forschung in Deutschland bislang weitestgehend ausgeblendet. Zwar beschäftigen sich einige ausgewählte Studien mit dem Thema der Schulwahl, also der Nachfrageseite. Empirische Studien zum Thema schulischer Wettbewerb, der Angebotsseite, lagen zum damaligen Zeitpunkt für den deutschsprachigen Raum nicht vor. Da jedoch davon auszugehen ist, dass die Neue Steuerung auf mittlere Sicht mit der Etablierung von Quasi-Märkten – Märkten, auf denen die Leistungserbringung unter Wettbewerbsbedingungen erfolgt, Finanzierung, Kontrolle und Steuerung jedoch in öffentlicher Hand liegen – einhergeht, ist es notwendig, auch diesen Aspekt zu betrachten.

Quasi-Märkte im Bildungs- bzw. Schulbereich zeichnen sich im Idealfall durch freie Schulwahlmöglichkeiten, leistungsbezogene Mittelzuweisungen an Schulen, Schulautonomie sowie schulinterne Personal- und Organisationsentwicklungsmöglichkeiten aus. Dabei benötigen die ökonomischen Regulierungssysteme Mechanismen, die die Markttransparenz erhöhen, indem sie Informationen bereitstellen und so Vergleiche möglich machen wie z.B. Evaluations- und Qualitätssicherungssysteme. Ein idealer Schulmarkt zeichnet sich dabei durch zwei Prämissen aus: Wettbewerb von Anbietern und Wahlfreiheit von Konsumenten. Die Idee hinter einem auf Marktstrukturen beruhenden Wettbewerb im Schulbereich ist relativ simpel: Wenn Schüler und ihre Familien aus einem Angebot an Schulen wählen können, dann können Schulen ihre Klientel nicht mehr länger als gegeben annehmen und müssen entsprechend dafür sorgen, ihre Erträge und Leistungen zu verbessern, um die Präferenzen, die Schüler und Eltern haben, zu befriedigen, um so auf dem Markt bestehen zu können. Verbesserte Schulwahloptionen für Eltern und Schüler tragen somit auch immer das Versprechen auf eine Qualitätssteigerung im Bildungsbereich in sich, indem sie einen positiv-kausalen Zusammenhang von Wahloptionen, Wettbewerb und Bildungsqualität unterstellen: Wahlmöglichkeiten für Schüler und ihre Familien erzeugen Wettbewerb zwischen Schulen und Wettbewerb zwischen Schulen erzeugt Bildungsqualität. Bekannt ist jedoch auch, dass eine Freigabe von Wahlmöglichkeiten im Bildungssystem zu sozioökonomisch segregierten Schulen und sozialen Benachteiligungen führen kann.

Inwieweit diese Annahmen auch auf Schulen bzw. ein Bildungssystem in Deutschland zutreffen, wird im Rahmen des Beitrages dann genauer untersucht. Hierfür wurden in einem ersten Schritt stochastische Netzwerkanalysen berechnet, also die beobachteten Wechselbeziehungen von Schülerinnen und Schülern am Übergang von der Primar- in die Sekundarstufe analysiert. Die stochastische Modellierung erfolgte in zwei Schritten. Zunächst sind für das gesamte Netzwerk latente Cluster modelliert worden. Der zweite Schritt umfasste dann die Unterteilung der in Schritt 1 ermittelten (regionalen) Cluster in (lokale) Subcluster oder anders ausgedrückt regionale Schulmärkte. Das Verfahren ähnelt dabei der latenten Klassenanalyse (LCA). Dieses Vorgehen wurde gewählt, da es gegenüber klassischen Ansätzen zur Modellierung von Schulmärkten, in denen diese durch die Forschenden anhand von regionalen oder administrativen Grenzen a priori definiert werden, auf tatsächlich beobachtete Informationen zurück greift. Während die Definition einzelner lokaler Märkte im klassischen Vorgehen somit komplett von den Vorannahmen der Forschenden zu den jeweiligen Grenzen von Märkten abhängt (und insofern Fehler in der Definition möglich sind), werden lokale Schulmärkte in Beitrag 11 anhand von Beobachtungen modelliert.

Mithilfe dieses Ansatzes ließen sich fünf regionale Schulmärkte in Hamburg nachweisen, die dann weiter in lokale Schulmärkte unterteilt wurden. Dabei wurde für die weiteren Analysen exemplarisch ein regionaler Schulmarkt (Hamburg-Süderelbe) herausgegriffen um anhand dieses regionalen Marktes und seiner lokalen (Sub-)Märkte erstmalig Wettbewerbsmaße und Segregationseffekte zu bestimmen. Die Erfassung der sozialen Segregation erfolgte mithilfe von Daten des räumlichen Sozialmonitorings des Rahmenprogramms Integrierte Stadtteilentwicklung, kurz „RISE“, der Stadt Hamburg. Als Maß für den Wettbewerb der Schulen um die Schülerschaft wurde der sogenannte Herfindahl-Index verwendet. Der Herfindahl-Index bemisst Wettbewerb anhand der Anbieterkonzentration. Der Index kann Werte zwischen 1/N und 1 annehmen und es gilt: Je größer der Wert, umso geringer der Wettbewerb.

Die Befunde zum Wettbewerb machen deutlich, dass dieser Index auf dem betrachteten regionalen Schulmarkt stark ausgeprägt ist – der Herfindahl-Index beträgt im Mittel 0,05 und liegt damit deutlich unter den Werten, die aus anderen Studien (z.B. den USA, dort wird ein mittlerer Index von 0,37 berichtet) bekannt sind. Auch auf den lokalen Schulmärkten ist der Wettbewerb zwischen Schulen um Schülerinnen und Schülern vergleichsweise hoch (Herfindahl-Index zwischen 0,10 und 0,17). Es sind demnach nahezu perfekte Märkte mit dem entsprechenden Wettbewerb zu beobachten. Betrachtet man jedoch zusätzlich die sozialen Hintergründe der Schülerinnen und Schüler, zeigt sich, dass es nahezu keinen Wettbewerb um Schülerinnen und Schülern gibt, die über einen hohen Sozialstatus verfügen – diese Schülerschaft konzentriert sich auf wenige ausgewählte Schulen. Besonders auffällig ist dabei der lokale Schulmarkt Wilhelmsburg. Dieser Markt ist von den anderen Märkten der Region abgeschottet und die dort ansässigen Schulen rekrutieren fast ausschließlich (17 von 20) Schülerinnen und Schüler aus benachteiligten sozialen Lagen.

Damit zeigt dieser Beitrag erstmals für eine Region in Deutschland, welche Folgen die Neue Steuerung im Bildungssystem auf den Wettbewerb zwischen Schulen sowie auf damit potenziell einhergehende Segregationseffekte im Schulsystem haben kann. Dabei liegt der besondere Wert zum einen im neu entwickelten Ansatz zur Modellierung schulischer Märkte, der es auch ermöglicht, Veränderungen von Marktstrukturen im Laufe der Zeit zu berücksichtigen. Zum anderen macht der Beitrag aber auch deutlich, dass der Wettbewerb zwischen Schulen in Deutschland stark, und ggf. sogar stärker als aus anderen Nationen bekannt, sein kann und dass eine Wettbewerbssteuerung auch in Deutschland mit negativen sozialen Nebeneffekten einhergeht.

Beitrag 12: Pietsch, M. , Graw, S. & Schulte, K. (angenommen). Inspektionsbasierte Unterrichtsentwicklung an Schulen in schwieriger Lage. In T. Stricker (Hrsg.), Zehn Jahre Fremdevaluation in Baden-Württemberg - Stand der Forschung – Zwischenbilanz – Perspektiven. Wiesbaden: VS.

Beitrag 12 wendet den in Beitrag 10 ermittelten Sozialindex sowie die inhaltsanalytischen Vorarbeiten aus Beitrag 5 an, um zu prüfen, in wie weit eine inspektionsbasierte Unterrichtsentwicklung durch die sozialen Kontextbedingungen einer Schule beeinflusst wird. Dabei wird davon ausgegangen, dass die Grundvoraussetzungen für eine erfolgreiche Schulentwicklung an Schulen in sozial schwierigen Lagen primär folgende sind: a) Die Schaffung eines geregelten, geordneten und lernförderlichen Schulklimas, b) die Befähigung von Schulleitungen, den Veränderungsprozess aktiv zu gestalten, c) die Etablierung angemessen hoher Erwartungen an alle Schulbeteiligten, d) die Kultivierung einer innerschulischen Feedbackkultur, in der die Beteiligten Rückmeldungen als Entwicklungsimpuls verstehen sowie e) die Befähigung aller Schulbeteiligten, sich kontinuierlich weiter zu entwickeln. Im Rahmen des Beitrages werden für die Analyse von Daten, die an n = 49 Schulen erhoben wurden, Mehrebenenstrukturgleichungsmodelle genutzt und der direkte und indirekte (über Schulleitungen vermittelte) Einfluss schulischer Rahmenbedingungen auf Unterrichtentwicklungsmaßnahmen analysiert. Konkret wurde ein wurde ein so genanntes doppelt-latentes (doubly-latent) Modell genutzt, dass Stichprobenfehler auf der Individual- sowie Messfehler auf der Individual- und der Schulebene korrigiert.

Die inspektionsbasierte Unterrichtsentwicklung wurde dabei anhand von Entwicklungsberichten dargestellt, die durch Schulleitungen zum Zeitpunkt der zweiten Inspektion abgegeben wurden. Hierin werden die Schulleitungen darum gebeten, anzugeben, welche für die Schulen bedeutenden Entwicklungsmaßnahmen infolge der ersten Inspektion ergriffen sowie ob diese Maßnahmen umgesetzt wurden. Die Berichte sollen auf den zwischen Schule und Schulaufsichten getroffenen Ziel- und Leistungs-Vereinbarungen basieren. Der Schulaufsicht obliegt es auch, die Umsetzung der Maßnahmen zu kontrollieren. Insgesamt wurden nach Aussage der Schulleitungen der 49 Schulen insgesamt 119 Unterrichtentwicklungsmaßnahmen durchgeführt, im Schnitt rund 2,4 Maßnahmen je Schule, wobei die Spannbreite von null bis sieben Maßnahmen reichte.

Das Schulleitungshandeln wird anhand von Daten modelliert, die zum Zeitpunkt der zweiten Inspektion mithilfe von Fragebögen an Lehrkräfte erhoben wurden. Dabei knüpft die Inspektion an die Idee des lernzentrierten Führungshandelns an und erfasst daher u.a. die Dimensionen Instruktionale, Transformationale und Geteilte Führung mithilfe etablierte Instrumente: a) Transformationale Führung: Skalen aus dem Multifactor Leadership Questionnaire (MLQ), b)

Instruktionale Führung: Skalen aus dem Teaching and Learning International Survey (TALIS) und c) Geteilte Führung. Die Stichprobe der Lehrkräfte an den 49 Schulen umfasste n = 1.140 Personen.

Der soziale Kontext der Schulen wiederum wird anhand des in Beitrag 10 ermittelten Hamburger Sozialindex für Schulen aus dem Jahr 2013 beschrieben. Ein niedriger Wert zeigt dabei eine geringe soziale Belastung der Schule, ein hoher Wert eine hohe soziale Belastung an. Für die genutzte Schulstichprobe reichen die Werte von -1,53 bis zu 1,90. Der Mittelwert liegt bei -0,07, die Standardabweichung liegt bei 0,94. Als Kontrollvariablen werden darüber hinaus berücksichtigt: a) die Schulform, b) ob zwischen der ersten und der zweiten Inspektion an der Schule ein Schulleiterwechsel stattgefunden hat (Dummy-kodiert) und c) ob sich die Struktur der Schule in diesem Zeitraum (Überführung von Grund-, Haupt- und Realschule in reine Grundschule) verändert hat (Dummy-kodiert).

Die Ergebnisse der Studie zeigen, dass Inspektionseffekte in der vorliegenden Stichprobe nicht über Schulleitungsvariablen vermittelt werden – weder der Führungsstil der Schulleitung, noch ob ein Schulleitungswechsel stattfand, hat einen Einfluss darauf, ob und wie viele Unterrichtsentwicklungsmaßnahmen infolge einer Schulinspektion ergriffen werden. Ein interessanter Nebenbefund ist, dass – anders als international häufig berichtet – der soziale Kontext einer Schule keinen nachweisbaren Einfluss darauf hat, ob ein Schulleiterin bzw. ein Schulleiter eine Schule verlässt. Die nachweisbaren Effekte sind insofern direkt, wobei sich für sämtliche Kontroll- und Mediatorvariablen kein Zusammenhang mit inspektionsbasierten Unterrichtentwicklungsmaßnahmen nachweisen lässt. Einzig für den sozialen Kontext finden sich nachweisbare Effekte auf Ebene der Schule: Einerseits haben die sozialen Rahmenbedingungen tendenziell einen Effekt darauf, ob an einer Schule auf Basis einer Schulinspektion überhaupt Maßnahmen für die Unterrichtsentwicklung ergriffen werden. Andererseits haben diese Bedingungen dann wiederum einen noch größeren Effekt darauf, wie viele Maßnahmen in der Folge durchgeführt werden. Schulen in sozial schwieriger Lage gehen infolge einer Inspektion demnach seltener die Unterrichtsentwicklung an und setzen selbst dann eher weniger Maßnahmen um als Schulen in weniger belasteten sozialen Lagen. Dabei lassen sich mit den Modellvariablen insgesamt rund 70 Prozent der Unterschiede in der inspektionsbasierten Unterrichtsentwicklung auf Schulebene erklären.

Beitrag 12 ist dabei vor allem deswegen bedeutsam, da er die Wirkungsweise von Schulinspektionen auf Schulebene untersucht und dabei verdeutlicht, dass der soziale Kontext eine entscheidende Rolle bei der Verarbeitung der Ergebnisse spielt. Für die Praxis wirft dies die Frage auf, wie Rückmeldungen

bzw. Inspektionsverfahren an sich gestaltet werden können, damit Schulen in schwierigen sozialen Lagen als Folge auch tatsächlich die intendierten Maßnahmen für die Unterrichtsentwicklung ergreifen. Ein wichtiger und für Deutschland neuer Befund ist darüber hinaus, dass sich kein Zusammenhang von Schulleiterwechseln und sozialem Kontext finden lässt. Diesbezüglich stellt sich daher die Frage, inwiefern internationale Befunde zum Thema auf den hiesigen Kontext übertragbar sind. Auch die Tatsache, dass sich bei der Zusammenführung verschiedener Erhebungsinstrumente kein Effekt des Schulleitungshandelns auf die Unterrichtsentwicklung nachweisen lässt, ist bemerkenswert, da dies ggf. Probleme bzw. Artefakte in rein korrelativen Studien zum Thema, für die Daten aus nur einer Perspektive bzw. mit nur einem Instrument erhoben werden, indizieren kann.

Beitrag 13: Pietsch, M. & Leist, S. (angenommen). The Effects of Competition in Schooling Markets on Leadership for Learning. Zeitschrift für Bildungsforschung.

Beitrag 13 nutzt verschiedene Vorarbeiten, die hier bereits vorgestellt wurden, um der Frage nachzugehen, welchen Einfluss der Wettbewerb zwischen Schulen auf das Schulleitungshandeln hat. Als Datengrundlage dienen Befragungsdaten von n = 3.950 Lehrkräften aus 74 Hamburger Sekundarschulen sowie die Daten einer kompletten Jahrgangskohorte von Schülerinnen und Schülern (N=14.481), die im Schuljahr 2014/2015 von der Grundschule in die Sekundarstufe gewechselt haben. Dabei knüpft die Arbeit inhaltlich an Untersuchungen aus dem anglo-amerikanischen Raum an, die zeigen konnten, dass Wettbewerb zwischen Schulen in der Regel mit einer verbesserten Qualität von Unterricht und Schule einhergeht. Theoretisch wird dabei angenommen, dass Wettbewerb das Verhalten aller Schulbeteiligten verändert, auch das von Schulleitungen. Insbesondere sollte sich, so die Annahme in der internationalen Diskussion, ein positiver Zusammenhang mit instruktionalen sowie geteilten Führungspraktiken finden. Ob dies tatsächlich so ist, ist bislang jedoch kaum empirisch belegt. Diesbezüglich betritt Beitrag 13 insofern Neuland. Daher wendet sich der Beitrag, obwohl in der Zeitschrift für Bildungsforschung publiziert, insbesondere auch an eine internationale Leserschaft und wurde entsprechend auf Englisch verfasst.

Um die eingangs beschriebene Frage zu beantworten, wurde in einem ersten Schritt die Studie zu lokalen Schulmärkten, die in Beitrag 11 vorgestellt wurde, repliziert. Es wurden also erneut anhand der Schülerverlaufsdaten zuerst regionale Bildungsmärkte und im Anschluss daran lokale Bildungsmärkte mithilfe eines stochastischen Netzwerkansatzes modelliert. Letztere Märkte wurden diesmal jedoch, anders als in Beitrag 11, nicht nur exemplarisch, sondern für das gesamte Bundesland Hamburg ermittelt. Die Befunde zeigen eine hohe Stabilität der regionalen Bildungsmärkte: Wie auch in Beitrag 11 ließen sich mithilfe der Daten fünf regionale Märkte in

Hamburg nachweisen. Diese fünf regionalen Märkte ließen sich wiederum in 14 lokale Schulmärkte unterteilen.

Auf Basis dieser Informationen wurden in einem zweiten Schritt Wettbewerbsmaße berechnet. Als Maß für den Wettbewerb der Schulen um die Schülerschaft wurde auch hier der sogenannte Herfindahl-Index verwendet, der international häufig genutzt wird, um den Wettbewerb zwischen Schulen zu beschreiben. Der Herfindahl-Index bemisst Wettbewerb dabei anhand der Anbieterkonzentration. Der Index kann Werte zwischen 1/N und 1 annehmen und es gilt: Je größer der Wert, umso geringer der Wettbewerb. Darüber hinaus wurden Marktanteile sowie Verbleibsquoten berechnet. Wie auch in Beitrag 11 zeigt sich ein starker Wettbewerb zwischen Schulen in Hamburg, die Herfindahl-Indices der lokalen Schulmärkte liegen zwischen 0.08 und 0.22. Nur drei Märkte weisen eine etwas höhere Konzentration auf, was auf ein Oligopol hindeutet. Dies wird durch die Befunde zu den Marktanteilen von Schulen gestützt: Auf den meisten Schulmärkten verfügen die drei mächtigsten Anbieter in der Regel zusammen über einen Marktanteil von 30 bis 50 Prozent. Bei den drei auffälligen Märkten vereinen die drei marktmächtigsten Schulen jedoch rund 60 bis 75 Prozent der Schülerschaft auf sich.

In einem dritten Schritt wurden diese Informationen dann wiederum genutzt, um zu prüfen, ob sich ein Zusammenhang zwischen schulischem Wettbewerb und dem Leitungshandeln von Sekundarschulleitungen findet. Auch hier wird das Konzept des lernzentrierten Schulleitungshandelns (Leadership for Learning) genutzt und konkret der Zusammenhang mit den Führungsfacetten Instruktionale, Transformationale und Geteilte Führung bzw. deren Subdimensionen untersucht. Auch hier wurde wieder ein so genanntes doppelt-latentes (doubly-latent) Mehrebenenstrukturgleichungsmodell genutzt, dass Stichprobenfehler sowie Messfehler korrigiert. Aufgrund der Komplexität der Daten sowie der relativ kleinen Schulstichprobe wurden die Zusammenhänge für die einzelnen Führungsfacetten jeweils separat ermittelt.

Die Befunde zeigen, dass der Wettbewerbskontext einen Effekt darauf hat, wie eine Schulleitung führt. Dabei lassen sich, theoriekonform, insbesondere Zusammenhänge zwischen Wettbewerb und instruktionalen sowie geteilten Führungspraktiken finden. Es zeigt sich: Je größer der Wettbewerb zwischen Schulen, umso häufiger teilen Schulleitungen Führungsverantwortung und arbeiten an der Qualität des Unterrichts, indem sie sich bemühen, das Unterrichten von Lehrkräften durch zielgerichtete Interventionen konkret zu verbessern. Auffällig ist dabei auch, dass der soziale Kontext einer Schule sowie weitere strukturelle Kontextvariablen keinen Einfluss auf das Leitungshandeln haben, wenn der Wettbewerb um Schülerinnen und Schüler berücksichtigt wird. Damit

widersprechen die Befunde denjenigen aus dem anglo-amerikanischen Raum, die darauf hinweisen, dass Leitungshandeln stark durch den sozialen Kontext einer Schule beeinflusst wird und Wettbewerbseffekte auf schulische Variablen weniger stark ausfallen, wenn in empirischen Analysen der soziale Kontext und/oder strukturelle Merkmale von Schule und Schulsystem berücksichtigt werden.

Einordnung der unter 3.4. berichteten Beiträge

Diese letzten vier Beiträge beschäftigen sich mit der Frage, welche Effekte schulische Kontextbedingungen auf die evidenzbasierte Schul- und Unterrichtsentwicklung sowie dafür relevante innerschulische Voraussetzungen haben. Dabei beschäftigen sich zwei der vier Beiträge mit inhaltlichen und methodischen Vorarbeiten, und die anderen beiden Beiträge wenden diese Vorarbeiten dann auf konkrete Fragestellungen an.

Die berichteten Arbeiten zeigen dabei insbesondere auf, welche Möglichkeiten es gibt, den sozialen Kontext sowie den Wettbewerbskontext von Schulen empirisch zu modellieren und wenden diesbezüglich innovative Ansätze an. Erstmalig im deutschsprachigen Raum wird darüber hinaus an die anglo-amerikanische Forschung zum Wettbewerb zwischen Schulen angeknüpft, und es werden Effekte dieses Wettbewerbs auf Schulleitungshandeln und soziale Segregation untersucht. Diesbezüglich machen die Befunde deutlich, dass Wettbewerb auch in Deutschland (aus dem anglo-amerikanischen Raum ist dies bekannt) sowohl mit positiven als auch negativen Konsequenzen einher geht. Darüber hinaus konnte gezeigt werden, dass die inspektionsbasierte Unterrichtsentwicklung nicht nur, wie in Abschnitt 3.2 gezeigt, von innerschulischen sondern auch von Kontextbedingungen abhängt.

Zusammengenommen machen die Studien damit einerseits die Relevanz deutlich, den Kontext, unter dem Schul- und Unterrichtsentwicklung stattfindet, methodisch angemessen zu modellieren. Zum anderen verdeutlichen sie die Komplexität und Kompliziertheit solcher Prozesse die eben auch abhängig von den jeweiligen Rahmenbedingungen sind.

4. Zusammenfassung und Diskussion

Die Einführung der Neuen Steuerung im Deutschen Bildungssystem ist mit der Annahme verbunden, dass Schule und Unterricht infolge einer Kombination verschiedener Mechanismen evidenzbasiert entwickelt werden können und sollen. Geschehen soll dies durch eine outputorientierte Steuerung in

Verbindung mit einer zunehmenden Verlagerung von Entscheidungs- und Steuerungskompetenzen auf die Ebene der Einzelschule, Datenerhebungen inkl. Rückmeldeverfahren durch interne und externe Evaluationsmaßnahmen sowie die Etablierung von Marktmechanismen. Ob und wie infolge der Einführung dieser Mechanismen jedoch tatsächlich eine evidenzbasierte Entwicklung von Schule und Unterricht stattfindet, ist großteils unklar. Weder liegen belastbare Modelle vor, die theoretisch begründen, wie und warum diese Maßnahmen überhaupt wirken sollen, noch sind Modelle vorhanden, die es ermöglichen, den Prozess der Schul- und Unterrichtsentwicklung theoretisch zu beleuchten. Die empirischen Befunde wiederum sind heterogen, und es ist unklar, ob die erwünschten Wirkungen tatsächlich eintreten oder nicht und woran dies liegt.

Der erste hier vorgelegte Beitrag setzt daher den Rahmen für die weiteren Beiträge, indem er einerseits die Problematik aufzeigt, die mit der Evaluation von Wirkungen komplex-komplizierter Interventionen im Bildungssystem verbunden ist und andererseits die bisherige Forschungslage für einen derartigen Mechanismus – die Schulinspektion – kritisch-systematisch aufarbeitet. Leitend für die weiteren hier vorgelegten Arbeiten wird abschließend empfohlen, sich an der aktuellen Diskussion zur Evaluation der Effekte von Programmen und Maßnahmen (z.B. Cook et al. 2010, Reichhardt 2011) zu orientieren und somit sowohl methodisch belastbare Black-Box-Evaluationen als auch programmtheoretische Ansätze zu nutzen, um die Wirkungen von Maßnahmen der Neuen Steuerung im Bildungssystem systematisch zu evaluieren. Die weiteren im ersten Abschnitt zusammengefassten Arbeiten schlagen dann die Brücke von der Evaluations- zur empirischen Bildungsforschung und machen erneut die Komplexität des Untersuchungsgegenstandes deutlich.

Diejenigen Beiträge wiederum, die im darauf folgenden Abschnitt zusammen gefasst sind, greifen die eingangs vorgestellte Idee auf und evaluieren entsprechend einerseits die Wirksamkeit von Schulinspektion auf Schülerleistungen mithilfe eines Black-Box-Ansatzes und versuchen andererseits heraus zufinden, ob und wie dieses Instrument auf die Entwicklung von Schule und Unterricht wirkt. Die Befunde zeigen auf der einen Seite, dass sich leicht positive Effekte auf Schülerleistungen finden lassen. Auf der anderen Seite wird aber auch deutlich, dass innerschulische Voraussetzungen eine wichtige Determinante sind, wenn es um eine evidenzbasierte Entwicklung von Schule und Unterricht im Kontext des aktuellen Steuerungsparadigmas geht. Deutlich wird in den Studien auch, dass im Rahmen von Schulinspektionsverfahren insbesondere die Schulleitung eine wichtige Rolle dabei spielt, ob und wie Effekte entstehen.

Dieser Befund wiederum bildet die Ausgangslage für die in Abschnitt drei zusammen gefassten Studien. Untersucht wird hier, ob und wie Schulleitungen Einfluss auf den Unterricht von Lehrkräften

sowie diesbezügliche innerschulische Voraussetzungen haben. Auch hier ist Beitrag eins leitend, indem im Anschluss an diesen versucht wird, alternative Erklärungen für potenzielle Effekte durch den simultanen Einbezug verschiedener Führungsstile in Kontrolldesigns auszuschließen (Reichhardt 2011, S. 255):

> *"The way to warrant any claim to knowledge is by identifying alternatives and assessing their implications. If you have been diligent in coming up with alternatives and in assessing their implications, and if the implications of one explanation hold true much more than the implications of any of the alternatives, you are then justified in tentatively accepting that explanation as true. Such a path to knowledge is the path of eliminating alternative explanations."*

Die vorgelegten Befunde zeigen, wie komplex bereits die innerschulischen Zusammenhänge sind, dass einfach-lineare und monokausale Assoziationen von Schulleitungshandeln und der Unterrichtsgestaltung und diesbezüglichen innerschulischen Voraussetzungen nicht zu erwarten sind, und machen darüber hinaus deutlich, dass der Kontext einer Schule wiederum mit den innerschulischen Variablen und den diesbezüglichen Interaktionen zusammenhängt und diese beeinflussen kann.

Hier schließen dann wiederum die Arbeiten des letzten Abschnitts an und untersuchen den Einfluss von Wettbewerb – einem zentralen Aspekt im Paradigma der Neuen Steuerung – sowie des sozialen Kontexts auf die evidenzbasierte Unterrichtsentwicklung an Schulen sowie auf das Leitungshandeln von Schulleitungen. Im Rückgriff auf eigene konzeptionelle und methodische Vorarbeiten zeigen die empirischen Befunde, dass insbesondere der Wettbewerbskontext einer Schule eine Rolle dabei spielt, wie Schulleitungen agieren und dass Maßnahmen zur Unterrichtsentwicklung, die infolge einer Schulinspektion ergriffen werden, wiederum maßgeblich von den sozialen Rahmenbedingungen einer Schule abhängen.

Zusammengefasst machen die Befunde der hier vorgelegten empirischen Studien deutlich, in welch einem komplex-komplizierten Setting erwartet wird, dass Instrumente der Neuen Steuerung im Bildungssystem Wirkungen entfalten und dass die Annahme, dass Rückmeldungen kausal-linear zur einer evidenzbasierten Entwicklung von Schule und Unterricht und in der Folge zu verbesserten Schülerleistungen führen sollten, vergleichsweise naiv ist. Denn eine Vielzahl von interagierenden Akteuren ist auf verschiedenen Ebenen des Schulsystems verantwortlich dafür, dass Effekte eintreten

- dies unter Kontextbedingungen, die wiederum Einfluss auf das Handeln der jeweiligen Akteure sowie auf die Wirksamkeit der Steuerungsinstrumente an sich haben.

Dabei ist im Rahmen entsprechender Analysen der Ausschluss alternativer Erklärungen gar nicht so einfach. Denn was wie worauf wie stark wirkt, hängt mit den jeweiligen Zielkriterien zusammen. So zeigen die Analysen z.B., dass Schulleitungen das Klassenmanagement von Lehrkräften vor allem durch instruktionales Handeln beeinflussen können, die Einflussnahme auf einen kognitiv aktivierenden Unterricht jedoch einen Führungsstile-Mix voraussetzt. Wollen Schulleitungen hingegen die Voraussetzungen für gelingenden Unterricht sowie gelingende Schul- und Unterrichtsentwicklung schaffen, ist es wiederum wichtig, dass die Führungsverantwortung auf das Kollegium verteilt wird. Dabei werden diese Zusammenhänge dann wiederum durch den jeweils spezifischen Kontext einer Schule beeinflusst.

Gleichwohl zeigt sich in den empirischen Studien, dass durchaus positive Effekte infolge Neuer Steuerungsmechanismen auftreten können: So werden infolge von Schulinspektionen innerschulische Stärken ausgebaut und Schwächen abgebaut – es findet also eine Entwicklung von Schule statt. Auch der Unterricht wird verändert, es findet Unterrichtsentwicklung statt. Und sogar Effekte auf Schülerleistungen konnten in den empirischen Arbeiten nachgewiesen werden. All dies ist jedoch fragil, und Erklärungen, warum diese Effekte zu beobachten sind, können nur begrenzt getroffen werden. Dennoch lassen sich diesbezüglich Muster feststellen. Deutlich wird: Sowohl innerschulische Voraussetzungen als auch der Kontext spielen eine wichtige Rolle dabei, ob und wie Schule und Unterricht weiter entwickelt werden. Insbesondere das Führungsverhalten von Schulleitungen scheint hierfür relevant zu sein. Aber auch die Innovationskapazität von Lehrkräften – sehen sie sich in der Lage, Innovationen und Veränderungen im eigenen Unterricht umzusetzen? – sind den empirischen Studien zufolge eine wichtige Voraussetzung für eine evidenzbasierte (Weiter-)Entwicklung von Schule und Unterricht. Wettbewerb zwischen Schulen und – in begrenztem Maße – der soziale Kontext haben wiederum Einfluss auf diese Faktoren sowie darauf, ob infolge der Rückmeldung empirisch gewonnener Informationen überhaupt eine Entwicklung stattfindet.

Interessanterweise wurden die meisten dieser Aspekte – z.B. Wettbewerb, Schulleitungshandeln, Innovationskapazität – im Rahmen der empirischen Forschung zur Qualität von Schule und Unterricht im deutschsprachigen Raum bislang kaum erforscht. Die vorgelegten Arbeiten stellen teilweise sogar die einzigen verfügbaren Quellen dar, in deren Rahmen empirische Evidenz zu diesen – international empirisch teilweise durchaus umfangreich behandelten – Themen für den deutschsprachigen Raum generiert wurde. Auffällig ist dabei auch, dass eben diese Merkmale, die sich im Rahmen der

empirischen Studien als relevant für eine evidenzbasierte Schul- und Unterrichtsentwicklung heraus kristallisiert haben, im internationalen politischen ebenso wie im wissenschaftlichen Diskurs als zentrale Aspekte wirksamer Neuer Steuerung im Bildungssystem gelten, im deutschsprachigen Diskurs jedoch bestenfalls bzw. wenn überhaupt nur am Rande und in der Regel selbst dann eher implizit als explizit thematisiert werden.

Insofern untermauern die hier vorgelegten Beiträge auch die eingangs berichtete von Berkemeyer (2010) und Maag Merki (2018) aufgestellte These, dass die aktuellen Bildungsreformen im deutschsprachigen Bildungssystem vor allem normativ und weniger evidenzbasiert bzw. wissenschaftlich fundiert erfolgten. So werden Konzepte wie dasjenige der Neuen Steuerung unvollständig übernommen, und wichtige (Kern-)Aspekte werden nicht berücksichtigt. Wirkungs- und Wirksamkeitsannahmen werden stark verkürzt aufgestellt und tragen der Komplexität des Gegenstandes kaum Rechnung. Und Wirkungserwartungen werden wiederum normativ aufgeladen und überhöht. Dass dies so ist, daran dürfte die deutschsprachige Erziehungswissenschaft nicht unschuldig sein. Denn sie bleibt einerseits kommunizier- und empirisch prüfbare theoretische Modelle schuldig, die zu einer wissensbasierten Steuerung des Bildungssystems beitragen könnten. Andererseits sucht sie kaum den Anschluss an die internationale empirische Forschung zum Themengebiet der School-Effectiveness- und der School-Improvement-Forschung.

Die hier vorgelegten Studien machen deutlich, dass dies jedoch ein lohnenswerter und zielführender Ansatz sein kann und in Kombination mit einem programmtheoretischen Forschungsverständnis dazu beitragen könnte, die eklatanten Wissenslücken zu füllen. Dabei handelt es sich um einen iterativen, kleinschrittigen Prozess oder wie es Astbury und Leeuw (2010, S. 374) formulieren: „Program theory building with mechanisms involves constant shuttling between theory and empirical data, using both inductive and deductive reasoning.“ Mit dieser Arbeit wurde ein erster Schritt unternommen, dies systematisch anzugehen. Damit eine umfassend evidenzbasierte Steuerung des Bildungssystems sowie wissensbasierte Entscheidungen im Hinblick auf Reformmaßnahmen jedoch tatsächlich möglich werden, bedarf es einer Vielzahl weiterer, konzeptionell und empirisch ähnlich gelagerter Arbeiten sowie einer ständigen Überprüfung und Revision von Annahmen zu Wirkungen und zu Wirkmechanismen implementierter Steuerungsmechanismen im Deutschen Bildungssystem.

5. Literatur

Altrichter, H., & Maag Merki, K. (2016). Steuerung der Entwicklung des Schulwesens. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 1-27). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H., Moosbrugger, R., & Zuber, J. (2016). Schul-und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 235-277). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H., Rürup, M., & Schuchart, C. (2016). Schulautonomie und die Folgen. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 107-149). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bellmann, J. (2006). Bildungsforschung und Bildungspolitik im Zeitalter 'Neuer Steuerung'. *Zeitschrift für Pädagogik*, *52*(4), 487-504.

Berkemeyer, N. (2010). *Die Steuerung des Schulystems*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Berkemeyer, N., & Hermstein, B. (2018). Schulentwicklung(-sforschung) – Quo vadis?. In K. Drossel, & B. Eickelmann (Hrsg.), *Does' What works' work? Bildungspolitik, Bildungsadministration und Bildungsforschung im Dialog*. Münster: Waxmann.

Böttcher, W., & Keune, M. (2012). Externe Evaluation und die Steuerung der Einzelschule: Kontrolle oder Entwicklung?. In M. Raterman, & S. Stöbe-Blossey (Hrsg.), *Governance von Schul- und Elementarbildung* (S. 63-80). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bonsen, M. (2016). Schulleitung und Führung in der Schule. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 301-323). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

Chen, H.-T. (1990). *Theory-driven evaluations*. Newbury Park: Sage.

Chrispeels, J. H., Brown, J. H. & Castillo, S. (2000). School leadership teams: Factors that influence their development and effectiveness. *Advances in Research and Theories of School Management and Educational Policy*, 4, 39-73.

Cook, T. D., Scriven, M., Coryn, C. L., & Evergreen, S. D. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, 31(1), 105-117.

Cortez, D., Gayle, B. M., & Preiss, R. W. (2006). An overview of teacher effectiveness research: Components and processes. In B. M. Gayle, R. W. Preiss, N. Burrell, & M. Allen (Hrsg.), *Classroom communication and instructional processes: Advances through meta-analysis* (S. 263-277). London: Lawrence Erlbaum.

Dedering, K. (2012). *Steuerung und Schulentwicklung: Bestandsaufnahme und Theorieperspektive*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Demski, D. (2017). *Evidenzbasierte Schulentwicklung: Empirische Analyse eines Steuerungsparadigmas*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fend, H. (1986). „Gute Schulen–schlechte Schulen ". Die einzelne Schule als pädagogische Handlungseinheit. *Die Deutsche Schule*, 78(3), 275-293.

Fend, H. (2008). Schule gestalten. *Systemsteuerung, Schulentwicklung und Unterrichtsqualität. Wiesbaden:* VS Verlag für Sozialwissenschaften.

Fend, H. (2011). Die Wirksamkeit der neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung*, 1(1), 5-24.

Fuchs, H. W. (2009). Neue Steuerung – Neue Schulkultur. *Zeitschrift für Pädagogik*, 55(3), 369-380.

Hood, C. (2002). Control, bargains, and cheating: The politics of public-service reform. *Journal of Public Administration Research and Theory*, *12*(3), 309-332.

Klieme, E., & Steinert, B. (2009). Schulentwicklung im Längsschnitt. Ein Forschungsprogramm und erste explorative Analysen. In M. Prenzel, & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006* (S. 221-238). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maag Merki, K. (2008). Die Architektur einer Theorie der Schulentwicklung. Strukturanalyse und Interdependenzen. *Journal für Schulentwicklung*, 2, 22-30.

Maag Merki, K. (2018). Reformen im Bildungswesen. In F. Imlig, L. Lehmann, & K. Manz (Hrsg.), *Schule und Reform* (S. 243-254). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rahm, S. (2005). Einführung in die Theorie der *Schulentwicklung*. Weinheim: Beltz.

Scheerens, J., Glas, C. A., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring: a systemic approach* (Vol. 13). London: Taylor & Francis.

Jann, W. (2005). Neues Steuerungsmodell. In B. Blanke, S. v. Bandemer, F. Nullmeier, & G. Wewer (Hrsg.), *Handbuch zur Verwaltungsreform* (S. 74-84). Wiesbaden: VS Verlag für Sozialwissenschaften.

Johnson, K., Greenseid, L. O, King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, 30(3), 377-410.

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.

Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration & Leadership*, *43*(2), 269-289.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Punton, M., Vogel, I., & Lloyd, R. (2016). *Reflections from a realist evaluation in progress: scaling ladders and stitching theory*. Brighton: Institute of Development Studies.

Ramaprasad, A. (1983). On the Definition of Feedback. *Behavioral Science*, 28(1), 4-13.

Reichardt, C. S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation*, 32(2), 246-272.

Reinbacher, P. (2016). Ein theoretischer Bezugsrahmen für «Schulentwicklung». *Schweizerische Zeitschrift für Bildungswissenschaften*, *38*(2), 295-318.

Reynolds, D., Teddlie, C., Creemers, B. P. M., Scheerens, J., & Townsend, T. (2000). An introduction to school effectiveness research. In C. Teddlie & D. Reynolds (Hrsg.), *The international handbook of school effectiveness research* (S. 3-25). London: Falmer Press.

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities* (Vol. 17). Dordrecht: Springer Science & Business Media.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School effectiveness and school improvement*, 28(2), 242-258.

Schröter, E., & Wollmann, H. (2005). New Public Management. In B. Blanke, S. v. Bandemer, F. Nullmeier, & G. Wewer (Hrsg.), *Handbuch zur Verwaltungsreform* (S. 63-74). Wiesbaden: VS Verlag für Sozialwissenschaften.

Steffens, U., & Bargel, T. (2016). *Schulqualität – Bilanzen und Perspektiven*. Münster: Waxmann.

Stufflebeam, D. L. (1972). Evaluation als Entscheidungshilfe. In Wulf, C. (Hrsg.), *Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen* (S. 113-145). München: Piper.

Tarter, C. J. & Hoy, W. K. (1998) Toward a contingency theory of decision making. *Journal of Educational Administration*, 36(3), S. 212-228.

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research.* London: Falmer Press.

Thiel, F., Cortina, K. S., & Pant, H. A. (2014). Steuerung im Bildungssystem im internationalen Vergleich. *Das Selbstverständnis der Erziehungswissenschaft: Geschichte und Gegenwart Zeitschrift für Pädagogik: Beiheft*, *60*, 123-138.

Thiel, F., & Thillmann, K. (2012). Interne Evaluation als Instrument der Selbststeuerung von Schulen. In A. Wacker, U. Maier, & J. Wissinger (Hrsg.), *Schul-und Unterrichtsreform durch ergebnisorientierte Steuerung* (S. 35-55). Wiesbaden: VS Verlag für Sozialwissenschaften.

Weber, M. (1980). *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie*. Tübingen: Mohr.

Weber, K. (2006). From nuts and bolts to toolkits: Theorizing with mechanisms. *Journal of Management Inquiry*, 15, 119-123.

6. Gesamtverzeichnis der eingereichten Beiträge

Beitrag 1: 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, N. (Hrsg.), *Schul- und Bildungsforschung. Diskussionen, Befunde und Perspektiven* (S. 167-185). Münster: Waxmann.

Beitrag 2: Pietsch, M., van den Ham, A.-K. & Köller, O. (2015). Wirkung von Schulinspektion: Ein Rahmen zur theoriegeleiteten Analyse von Schulinspektionseffekten. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), *Schulinspektion in Hamburg. Der erste Zyklus der 2007 – 2013: Grundlagen, Befunde, Perspektiven* (S. 117-136). Münster: Waxmann.

Beitrag 3: Ehren, M. & Pietsch, M. (2016). Validation of Inspection Frameworks and Methods. In M. Ehren (Hrsg.), *Methods and Modalities of effective School Inspections* (S. 47-68). London: Springer.

Beitrag 4: Pietsch, M., Janke, N. & Mohr, I. (2014). Führt Schulinspektion zu besseren Schülerleistungen? Difference-in-Differences-Studien zu Effekten der Schulinspektion Hamburg auf Lernzuwächse und Leistungstrends. *Zeitschrift für Pädagogik*, 60(3), 446-470.

Beitrag 5: Pietsch, M., Feldhoff, T. & Petersen, L. S. (2016). Schulentwicklung durch Inspektion? Welche Rolle spielen innerschulische Verarbeitungskapazitäten? In AG Schulinspektion (Hrsg.), *Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten* (S. 227-262). Wiesbaden: VS.

Beitrag 6: Pietsch, M. & Hosenfeld, I. (2017). Von der Schulinspektion zur Unterrichtsentwicklung: Welche Rolle spielt die Schulleitung. *Empirische Pädagogik*, 31(2), 202-220.

Beitrag 7: Pietsch, M., Lücken, M., Thonke, F., Klitsche, S. & Musekamp, F. (2016). Der Zusammenhang von Schulleitungshandeln, Unterrichtsgestaltung und Lernerfolg: Eine argumentbasierte Validierung zur Interpretier- und Nutzbarkeit von Schulinspektionsergebnissen im Bereich Führung von Schulen. *Zeitschrift für Erziehungswissenschaft*, 19(3), 527-555.

Beitrag 8: Pietsch, M. & Tulowitzki, P. (2017). Disentangling School Leadership and its Ties to Instructional Practices. An empirical Comparision of various Leadership Styles. *School Effectiveness and School Improvement*, 28(4), 629-649.

Beitrag 9: Pietsch, M., Tulowitzki, P. & Koch, T. (angenommen). On the Differential and Shared Effects of Leadership for Learning on Teachers' Organizational Commitment and Job Satisfaction: A multilevel perspective. *Educational Administration Quarterly*.

Beitrag 10: Schulte, K., Hartig, J. & Pietsch, M. (2016). Berechnung und Weiterentwicklung des Sozialindex für Hamburger Schulen. In B. Groot-Wilken, K. Isaac & J-P. Schräpler (Hrsg.), *Sozialindices für Schulen – Hintergründe, Methoden und Anwendung* (S. 157-172). Münster: Waxmann.

Beitrag 11: Leist, S. & Pietsch, M. (2017). Bordering the Area of Spatial Relevance for Schools: A stochastic Network Approach using the Example of Hamburg, Germany. *Belgeo - Revue Belge de Géographie*, 2-3/2017 (Special Issue: Une géographie sociale de l'enseignement). OnlineFirst: https://doi.org/10.4000/belgeo.20332

Beitrag 12: Pietsch, M. , Graw, S. & Schulte, K. (angenommen). Inspektionsbasierte Unterrichtsentwicklung an Schulen in schwieriger Lage. In T. Stricker (Hrsg.), *Zehn Jahre Fremdevaluation in Baden-Württemberg - Stand der Forschung – Zwischenbilanz – Perspektiven*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Beitrag 13: Pietsch, M. & Leist, S. (angenommen). The Effects of Competition in Schooling Markets on Leadership for Learning. *Zeitschrift für Bildungsforschung*.

Die Forschungsarbeiten 1 und 2, 4-9 sowie 12 und 13 wurden maßgeblich von mir konzipiert, durchgeführt und verfasst. Die Koautorinnen und Koautoren wirkten großteils beratend und unterstützend, teils als Fachexpertinnen und Fachexperten, mit und steuerten vor allem inhaltliche und/ oder methodische Anregungen, Revisionen und Ergänzungen bei (so überprüfte z.B. Herr Prof. Koch die Doubly-Latent-Messmodelle in Beitrag 9 und revidierte diese leicht). In Beitrag 5 wurde Abschnitt drei fast allein von Herrn Prof. Dr. Feldhoff verantwortet. In Beitrag 7 wurden die Daten der KERMIT-Erhebung von Herrn Dr. Lücken aufbereitet und in Beitrag 13 wurden die geografischen Analysen (Modellierung der Märkte) von Herrn Leist durchgeführt. Arbeit 3 wurde in einem iterativen Vorgehen komplett gemeinsam vorbereitet und verfasst und basiert auf Vorarbeiten beider Autorinnen bzw. Autoren. Arbeit 10 wurde in erster Linie von Frau Dr. Schulte verantwortet, mein Anteil daran war vor allem konzeptioneller und methodischer Art. Arbeit 11 wurde gemeinsam verantwortet, wobei Idee und Konzeption sowie die Befunde zum Wettbewerb von mir stammen, Herr Leist jedoch alle Berechnungen durchführte und den überwiegenden Teil des Textes verfasste (da prospektiv bereits Beitrag 13 geplant war und Beitrag 11 hierzu eine eher methodisch orientierte, geografische Vorarbeit darstellt).

7. Kopien der eingereichten Beiträge

Knut Schwippert
Martin Bonsen
Nils Berkemeyer (Hrsg.)

Schul- und Bildungsforschung

Diskussionen, Befunde und Perspektiven

Festschrift für Wilfried Bos

Waxmann 2013
Münster / New York / München / Berlin

Marcus Pietsch, Nike Janke und Ingola Mohr

Führt Schulinspektion wirklich nicht zu besseren Schülerleistungen?

Eine Einschätzung zur Belastbarkeit vorliegender Wirksamkeitsstudien aus programmtheoretischer Perspektive

Die externe Evaluation von Schulen mittels Schulinspektionsverfahren gilt in den Ländern der Bundesrepublik Deutschland neben der outputorientierten Diagnose der Leistungsfähigkeit des Bildungssystems seit einigen Jahren als dasjenige Steuerungselement, von dem eine besonders hohe Innovationskraft für den Bildungsbereich erwartet wird. Erwartet wird insbesondere, dass Schulinspektion zu einer Verbesserung von Prozessqualitäten in Schulen führt und über diese in gesteigerte Schülerleistungen mündet (vgl. Böttcher & Kotthoff, 2010; Ehren & Visscher, 2006). Diejenigen Studien jedoch, die zu klären versuchen, ob Schulinspektionsinterventionen einen empirisch nachweisbaren Einfluss auf die Verbesserung von Schülerleistungen haben, weisen in der Regel nach, dass Schulinspektionen von durchschlagender Wirkungslosigkeit gekennzeichnet sind (vgl. Cullingford & Daniels, 1999; Luginbuhl, Webbink & Wolf, 2009; Matthews & Sammons, 2004; Rosenthal, 2004; Shaw, Newton, Aitkin & Darnell, 2003; Wilcox & Gray, 1996). So zeigen all jene Studien, die sich auf Effekte der englischen Schulinspektion OFSTED (The Office for Standards in Education) beziehen, dass keine oder sogar leicht negative Wirkungen auf Schülerleistungen als Folge von Schulinspektionen in England beobachtet werden können (vgl. Wolf & Janssens, 2007). Rosenthal (2004) weist beispielsweise nach, dass im Jahr nach einer Schulinspektion die Schülerleistungen an evaluierten Schulen um ca. einen halben Prozentpunkt niedriger ausfallen als an vergleichbaren Schulen, an denen keine Schulinspektion stattgefunden hat. Einzig eine rezente Studie aus den Niederlanden (vgl. Luginbuhl et al., 2009) zeigt entgegen der allgemeinen Befundlage, dass Schulinspektionen gegebenenfalls zu leichten Steigerungen der Schülerleistungen in der Größenordnung von zwei bis drei Prozent einer Standardabweichung in den ersten zwei Jahren nach einer Schulinspektion führen könnten. Nachdem die Autoren dieser Studie jedoch ein methodisch robusteres Alternativdesign anlegten, um eventuelle Stichprobenverzerrungen respektive Selektionseffekte zu kontrollieren, ließen sich in dieser Studie keine Effekte von Schulinspektionsbesuchen auf Schülerleistungen mehr nachweisen.

Schulinspektion wirkt nicht! Aber warum?

Studien zur Wirksamkeit von Schulinspektionen versuchen in der Regel aus den Ergebnissen von Schülerleistungstests oder zentralen Abschlussarbeiten abzuleiten, ob die Nichterreichung intendierter Ziele darauf zurückzuführen ist, dass Schulinspektionen gegebenenfalls nicht gut oder nicht korrekt durchgeführt werden. Die Überprüfung der Inspektionswirksamkeit anhand von Schülerleistungsdaten gilt den meisten Autoren dabei als „der strengste Maßstab für die Prüfung" (vgl. Böttcher & Kotthoff, 2010, S. 309).

Die derzeit vorliegenden Befunde weisen somit anscheinend empirisch verlässlich darauf hin, dass Schulinspektionen die avisierten Ziele in der Regel verfehlen oder sogar kontraproduktiv sind. Entsprechend resümieren viele Autoren (vgl. Gärtner & Pant, 2011; Husfeldt, 2011; Wolf & Janssens, 2007), dass Schulinspektion sich nachteilig auf die Entwicklung von Schülerleistungen auswirke.

Gleichwohl können auch andere Gründe hinter der beobachteten Nicht-Wirksamkeit von Inspektionen stecken. So könnten die divergierenden Befunde aus England und den Niederlanden beispielsweise auch darauf hindeuten, dass die unterschiedlichen Effekte auf verschiedenartige Entwicklungsparadigmen zurückzuführen sind – in England wurden die Befunde der Schulinspektion veröffentlicht, und es wird entsprechend eine Entwicklung durch Wettbewerb (und damit potenziell einhergehender Performanz-Paradoxa, vgl. hierzu Leeuw & van Thiel, 2002) forciert, während Inspektionsbefunde in den Niederlanden nur den Schulbeteiligten und der Bildungsadministration zur Verfügung gestellt wurden und eine Entwicklung durch die Bereitstellung von Informationen das Ziel der Inspektion ist (vgl. hierzu Böttger-Beer & Koch, 2008). Entsprechend stellt sich bei Betrachtung der bisher durchgeführten Studien u. a. die Frage, ob diese theoretischen Annahmen in den Analysen ausreichend beachtet wurden.

Und nicht zuletzt deutet die studieninterne Inkonsistenz der rezenten niederländischen Befunde auf mögliche gravierende Probleme in der bisherigen Anwendung kausalanalytischer Forschungsmethoden bei der Evaluation von Inspektionseffekten hin. Wenn die Anwendung unterschiedlicher Verfahren zum Umgang mit Selektionseffekten innerhalb einer Studie bereits solch markante Unterschiede in den Studienergebnissen nach sich zieht, wie ist es dann um die Belastbarkeit der Studien zur Inspektionswirksamkeit insgesamt bestellt? Gemutmaßt werden kann so z. B., dass viele der bislang durchgeführten empirischen Analysen eventuell mithilfe inadäquater methodischer Herangehensweisen generiert wurden. Eine Vermutung, die ein Review von Wolf und Janssens (2007) zur Wirksamkeit von Schulinspektionen *grosso modo* nährt.

Forschungsfrage und Überblick

Deutlich sichtbar wird bereits anhand dieser Beispiele, dass, je nach Perspektive, ein anderer Grund für das Verfehlen der intendierten Ziele angenommen werden kann und eine mutmaßliche Wirkungslosigkeit entsprechend nicht zwangsläufig auf eine mangelhafte Durchführungspraxis schließen lässt. Denn, warum eine Intervention empirisch für nicht-wirksam befunden wird, kann seine Ursache sowohl in fehlerhaften bzw. unzureichend beschriebenen Modellannahmen, einer tatsächlich ungenügenden Implementations- und Umsetzungspraxis aber auch in studienimmanenten methodischen Unzulänglichkeiten haben. Inwieweit die Frage, ob Schulinspektion auf Schülerleistungen wirkt, daher so eindeutig mit *Nein* beantwortet werden kann, wie es viele rezente Übersichten zum Forschungsstand derzeit implizieren, ist daher eher ungewiss.

Ziel des vorliegenden Beitrags ist es, die Probleme in der aktuellen Forschung zur Wirksamkeit von Schulinspektion zu systematisieren und Lösungswege aufzuzeigen. Zuerst wird daher ein Analyserahmen erarbeitet, der es ermöglicht, Studien zur Wirksamkeit von Schulinspektionsverfahren auf ihre Belastbarkeit hin zu untersuchen. Anschließend werden die bislang vorliegenden Studien anhand dieses Rasters beurteilt und eine Einschätzung zum Stand der empirischen Forschung zur Schulinspektionswirksamkeit

gegeben. Abschließend werden Möglichkeiten, die berichteten Forschungsdesiderata anzugehen, diskutiert.

1 Analyserahmen: Warum wirkt Schulinspektion nicht?

In diesem Abschnitt wird der genutzte Analyserahmen für das nachfolgende Review vorgestellt. Zentral in diesem Abschnitt sind die Vorstellung eines Modells zur theoriebasierten Evaluation von Interventionen und die Klassifikation möglicher Fehler, die bei einer solchen Evaluationspraxis auftreten können.

Schulinspektionen als Intervention auf Schulebene

Die Funktionen, die Schulinspektion erfüllen soll, unterscheiden sich nicht grundlegend von denen anderer Werkzeuge des Bildungsmonitorings (Maritzen, 2008; Scheerens, Glas & Thomas, 2003). So hat Schulinspektion die Aufgabe, Informationen aus dem Bereich Bildung und Erziehung systematisch zu beschaffen und aufzubereiten, um auf diesem Wege Akkreditierungen, Rechenschaftslegungen und Diagnosen für systemisches Lernen im Bildungssystem zu ermöglichen. Schulinspektionen sollen entsprechend dazu beitragen, elementare Standards von Bildungsqualität zu gewährleisten, einen verbesserten Service für das einzelschulische Qualitätsmanagement zu bieten und unterschiedlichen Akteuren im Bildungssystem verwertbare Informationen zu schulischen Prozessqualitäten bereitzustellen (vgl. Döbert, Rürup & Dedering, 2008; Maritzen, 2007, 2008, 2009; Pietsch, Schnack & Schulze, 2009). Dabei liegt der Schwerpunkt der Inspektionsarbeit auf der Evaluation einzelschulischer Prozesse (vgl. Böttcher & Kotthoff, 2010; Döbert et al., 2008; Maritzen, 2009; van Ackern & Klemm, 2009), wobei die Inspektionen in den Ländern trotz augenscheinlicher Unterschiede in Ausstattung, Ausgestaltung und ministerieller Anbindung „meist auf Grundlage von Verfahren mit wiederkehrenden Standardelementen" erfolgen (Maritzen, 2008, S. 87).

Grundlage für die externe Einzelschulevaluation bilden dabei in allen Ländern die landesspezifischen Qualitätsrahmen oder Qualitätstableaus, die im Sinne eines Input-Prozess-Output-Modells (teilweise unter Berücksichtigung von Kontextfaktoren) Anforderungen an Schulqualität definieren (vgl. Bos, Holtappels & Rösner, 2006; Müller, 2010). In allen Inspektionen wurden, basierend auf diesen Qualitätskatalogen, Leistungsindikatoren definiert, die es ermöglichen sollen, die Qualität einzelner Prozesse möglichst differenziert zu erfassen (vgl. Böttcher & Kotthoff, 2007). Geschulte Experten – meist Lehrkräfte, Schulleitungen oder Schulaufsichten – evaluieren anhand dieser Indikatoren die Schulen in Teams von zwei bis vier Personen (vgl. Döbert et al., 2008; Müller, 2010). Die Datenerhebung selbst erfolgt über ein umfangreiches Repertoire an Methoden. Fragebögen und Interviews, gerichtet an verschiedene Schulbeteiligte, sowie Schulbegehungen und die Begutachtung schulinterner Dokumente und Statistiken gehören derzeit ebenso wie die Beobachtung von Unterrichtseinheiten zum länderübergreifenden Methodenstandard (vgl. Bos et al., 2006; Döbert et al., 2008). Die derart ermittelten Befunde werden in einem Bericht zusammengefasst und den Schulen sowie gegebenenfalls weiteren Beteiligten, wie z. B. Schulaufsichten, übergeben. In regelmäßi-

gem Turnus werden darüber hinaus Berichte verfasst, die Befunde der Schulinspektionen auf Systemebene präsentieren.

Durch diese Rückmeldungen aus Schulinspektionen erhalten die Schulbeteiligten, ebenso wie Bildungsadministration und Bildungspolitik, Informationen zur extern wahrgenommenen Qualität innerschulischer Prozesse auf verschiedenen Ebenen. Obwohl dies häufig nicht explizit benannt wird, steht hinter dem Vorgehen, aufbereitete Informationen an verschiedene Gruppen auf Schul- und Schulsystemebene zurückzumelden, die Idee, dass die Rückmeldungen von Evaluationsergebnissen als Grundlage für Entwicklungen an den evaluierten Schulen genutzt werden, die in ihrer Konsequenz in verbesserte fachliche Schülerleistungen münden (vgl. Wolf & Janssens, 2007; Ehren & Visscher, 2006; Ehren, Leeuw & Scheerens, 2005). Insbesondere in den Deutschen Ländern zeichnet sich diesbezüglich aktuell ein Trend ab (vgl. Böttcher & Kotthoff, 2010, S. 307) „demzufolge Schulinspektionen ... als ein Unterstützungs- und Dienstleistungsangebot für die Schule verstanden werden, deren Erkenntnisse wirksam für die interne Schulentwicklung genutzt werden können bzw. sollen". Folgerichtig betont Köller (2009), dass der Wert von Schulinspektionen insbesondere darin läge, dass sie Hinweise auf die Prozesse vor Ort gäben, mithilfe derer Hypothesen darüber generiert werden könnten, welche Veränderungen von Schule und Unterricht zu höheren Lernerträgen der Schülerinnen und Schüler führen können.

Modelle zur Wirksamkeit von Interventionen

Ob Schulinspektionen die intendierten Wirkungen tatsächlich erzielen, muss sich wiederum mithilfe von Wirksamkeitsanalysen überprüfen lassen. Die Wirksamkeit von Interventionen kann dabei entweder mithilfe von *Blackbox*-, *Greybox*- oder *Whitebox*-Verfahren bestimmt werden (vgl. Scriven, 1994). Im Rahmen von *Blackbox*-Verfahren werden die intendierten Effekte einer Interventionsmaßnahme, in diesem Falle der Schulinspektion, evaluiert, ohne die Wirkungsmechanismen und Wirkungsbedingungen in den Blick zu nehmen. In *Greybox*-Verfahren werden Interventions- und Wirkungszusammenhänge betrachtet, ohne darauf einzugehen, wie diese im Detail funktionieren. In *Whitebox*-Verfahren wiederum werden auch die inneren Prozesse und Wirkungsweisen der Intervention sowie Wirkungszusammenhänge detailliert und in Gänze dargestellt und betrachtet. Während es *Grey*- und *Whitebox*-Ansätze somit ermöglichen, explizite und implizite Annahmen einer Intervention und ihrer Wirkungsweise modellhaft sichtbar zu machen und auf diesem Wege im besten Falle herauszufinden, *wie* und *warum* sie Wirksamkeit entfaltet oder nicht, können *Blackbox*-Ansätze ausschließlich klären, *ob* eine intendierte Wirkung infolge einer Intervention eintritt oder nicht (vgl. Chen, 2005, 2006).

Diese Unterscheidung findet ihre Entsprechung in der Anlage von Programmtheorien. Programmtheorie bedeutet im Kontext der empirischen Bildungsforschung, dass für ein Bildungsprogramm der Zusammenhang zwischen Ursachen und Wirkungen anhand einer gegenstandsbezogenen Theorie des Programmablaufs beschrieben wird, eines logischen Modells, das theoretisch plausible Annahmen darüber trifft, welche Prozesse auf dem Weg vom Input zum Output Veränderungen bewirken (vgl. Chen, 2005; Rogers, 2002; Weiss, 1997). Während ökonomische Input-Output- oder *Blackbox*-Modelle nur Aussagen zum Zusammenhang von Intervention und Effekten ermöglichen, gestatten es einfach-lineare Programmtheorien, darüber hinaus die vermittelnden Prozesse, Aktivitä-

ten und Mechanismen zu beleuchten, die bei der Generierung von Effekten wirken (vgl. Chen, 2006). Nicht-lineare Programmtheorien ermöglichen es, ergänzend weitere Sachverhalte in Wirksamkeitsanalysen – z. B. Kontexteffekte oder auch Voraussetzungen der Interventionsmaßnahme – zu berücksichtigen (vgl. Coryn, Noakes, Westine & Schröter, 2011). Komplexere Programmtheorien bestehen daher in der Regel (vgl. z. B. Chen 2006, Coryn et al., 2011) aus mindestens zwei Teilen (vgl. Abb. 1), die eng miteinander verwoben sind; einerseits aus einem Veränderungsmodell und andererseits aus einem Handlungsmodell (vgl. Chen, 2006, S. 277):

> The action model and change model are closely related to each other and are essential for the success of a program. On the one hand, a change model is needed to justify the selection of an intervention for achieving the goals and/or outcomes and it provides a basis for developing the action model. On the other hand, the action model provides a blueprint to organize program activities and to activate and energize the change model for achieving program goals.

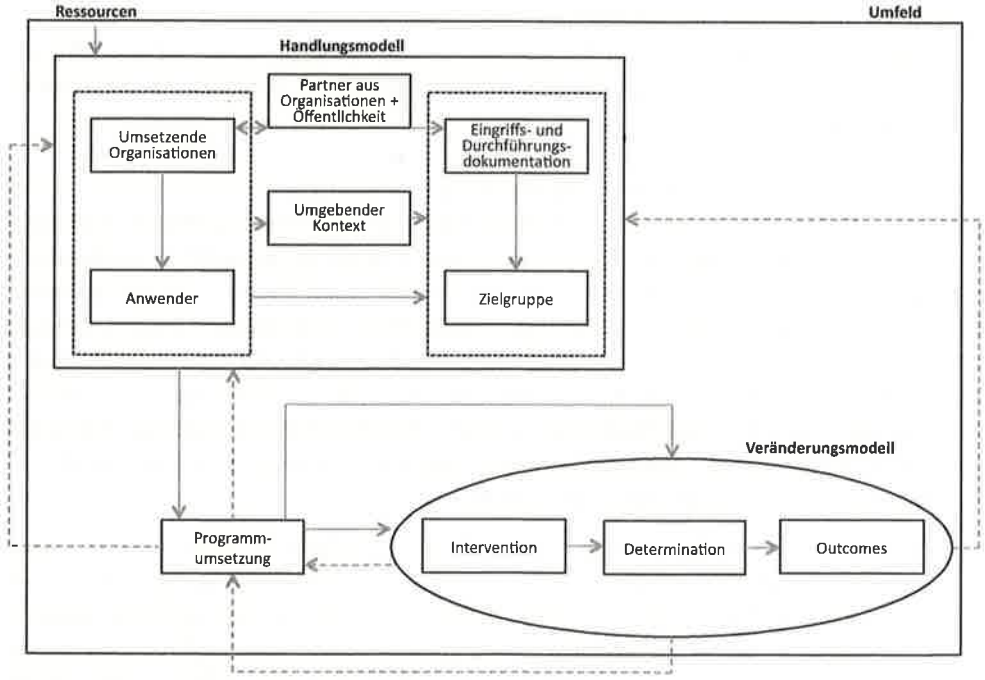


Abbildung 1: Basismodell zur Evaluation von Interventionswirkungen (Chen, 2006)

Dabei umfasst das Veränderungsmodell, entsprechend der einfach-linearen Modelle, drei Komponenten: a) die Intervention, die sich auf das Programm bezieht, mit dessen Hilfe Veränderungen bewirkt werden sollen, b) die Determination, die diejenigen Prozesse und Mechanismen umfasst, die den Input durch die Intervention, mit Blick auf mögliche Effekte, vermitteln und c) den Outcomes, womit die antizipierten Effekte des Programms gemeint sind. Das Veränderungsmodell unterstellt entsprechend, dass

die Implementation einer Intervention diejenigen Determinanten kausal beeinflusst, die ihrerseits zu Veränderungen in den Outputs führen.

Das Handlungsmodell hingegen beschreibt die Voraussetzungen und Annahmen der Intervention selber, stellt somit also dar, mit welchen Mitteln, unter welchen Voraussetzungen und gerichtet an wen die Intervention Wirkung erzielen soll. Das Handlungsmodell umfasst dabei sechs Bereiche:

1) *Umsetzende Organisationen:* Die Organisationen, die für die Organisation von Personal und die Ressourcenallokation bezüglich der Intervention verantwortlich sind.
2) *Anwender:* Die Personen, die für die konkrete Umsetzung respektive Implementierung des Programms verantwortlich sind.
3) *Partner aus Organisationen und Öffentlichkeit:* Die Organisationen oder Partner, die zusätzlich zur Primärorganisation notwendig sind, um die Intervention zielgerichtet durchzuführen.
4) *Umgebender Kontext:* Der ökologische Kontext, der direkt mit dem zu implementierenden Programm interagiert, z. B. in Form von Zielvorgaben, Normen, Gesetzen.
5) *Eingriffs- und Durchführungsdokumentation:* Der Umsetzungsplan, der die Intervention inklusive ihrer spezifischen Inhalte und Aktivitäten beschreibt und dabei die Schrittfolge definiert, mit der die Intervention im Feld umgesetzt werden soll.
6) *Zielgruppe:* Die Zielgruppe, für die das intervenierende Programm gedacht ist, dabei insbesondere das Vorhandensein von klar definierten Ansprechpartnern und die Einbindung in das Programm.

Fehlerquellen bei der Evaluation von Interventionen

In einem solchen kausalanalytischen Setting kann ein nicht nachweisbarer Wirkungszusammenhang dann wiederum auf maximal drei Gründe zurückgeführt werden: Erstens kann ein Programm- respektive Implementations-, zweitens ein Theorie- und/oder drittens ein Methodenfehler vorliegen (vgl. Stame, 2010). Während ein Programmfehler vorhanden ist, wenn es mittels einer Intervention nicht gelingt, eine intendierte Wirkung nachweisbar zu erzeugen, liegt ein Theoriefehler vor, sofern ein theoretisch postulierter Wirkungszusammenhang nicht ausreichend valide begründet wurde (vgl. Suchman, 1969). Ein Beispiel aus der Bildungsforschung, das diese Unterscheidung anschaulich illustriert, geben Raudenbush und Sadoff (2008, S. 139):

> Consider a study in which the assignment of schools or classrooms to a novel instructional innovation is found to have no significant impact on student learning. Assume that the study design was unbiased and provided adequate statistical power to detect a nonnegligible effect. Two explanations immediately arise. Program evaluators refer to these as 'theory failure' versus 'implementation failure' … First, it may be that the innovation changed classroom instruction in the ways intended but that those classroom changes made no difference in student learning. The term theory failure describes this scenario because the theory that links intended changes in instruction to intended student outcomes will have proven incorrect. Second, the innovation may never have been effectively implemented in classrooms. Perhaps the innovators lacked skill in working with teachers or perhaps the teachers lacked the

> skill, knowledge, or motivation to put the innovative ideas to work in their teaching. In any case, program theory about the relationship between the intended instruction and student outcomes was never tested, leading to 'implementation failure'.

Gleichwohl wird in diesem Beispiel unterstellt, dass die eingesetzten Forschungsmethoden zur Evaluation der Interventionsmaßnahme adäquat eingesetzt wurden; eine Annahme, die häufig nicht zutrifft (vgl. Lipsey, Crosse, Dunkle, Pollard & Stobart, 1985; Stame, 2010; White, 2010). Entsprechend kann eine nicht nachweisbare Wirkung auch darauf zurückzuführen sein, dass Methodenfehler vorliegen. Die Nicht-Wirksamkeit einer Intervention rührt in diesem Falle daher, dass es im Rahmen der empirischen Kausalstudie aufgrund methodologischer Unzulänglichkeiten nicht gelingt, den angenommenen Wirkungszusammenhang empirisch verlässlich zu überprüfen (vgl. Lipsey et al., 1985).

2 Studien zur Wirksamkeit von Schulinspektionen

Die Evaluation von Schulinspektionswirksamkeit findet im Rahmen des oben genannten Paradigmas statt. In diesem Abschnitt werden einerseits die möglichen Fehlerquellen bei der Evaluation von Schulinspektionswirksamkeit ausführlicher beschrieben und andererseits bislang vorliegende Studien zur Wirkung und Wirksamkeit von Schulinspektionsverfahren diesbezüglich untersucht.

Theoriefehler als Gegenstand der Evaluation von Inspektionseffekten

(Bildungs-)Programme haben immer inhärente Theorien und Annahmen darüber, wie und warum eine Intervention zu bestimmten Ergebnissen führen soll. In diesem Zusammenhang meint Bildungsprogramm eine Sammlung speziell zugeschnittener Maßnahmen, mit deren Hilfe spezifische Bildungsziele erreicht werden sollen (vgl. Rogers, 2002). Bei einem Programm, in Abgrenzung zu einem Projekt, handelt es sich um pädagogische Maßnahmen, die nicht von vornherein auf eine bestimmte Zeit begrenzt sind (vgl. Ditton, 2010). Mit Blick auf Schulinspektionen sind diese Theorien häufig nicht expliziert und müssen entsprechend durch die Rekonstruktion von Programmtheorien sichtbar gemacht werden (vgl. Ehren et al., 2005; Ehren & Honingh, 2011). Bei Studien zur Evaluation von Theoriefehlern geht es daher einerseits darum, den implizit angenommenen Wirkungszusammenhang von Eigenschaften der Intervention mit erwarteten Programmzielen zu verdeutlichen (vgl. Stame, 2010) und andererseits darum, die Komplexität und Kompliziertheit einer Interventionsmaßnahme darzustellen (vgl. Rogers, 2008).

Programmtheorien müssen daher in erster Linie darauf eingehen, warum und auf welchem Wege eine Intervention Wirkung entfalten soll. Die Wirkungsweise muss dabei im Rahmen von Handlungs- und Veränderungsmodellen verdeutlicht werden (vgl. Chen, 2006), die theoretische Annahmen zur Wirkungsweise der Intervention postulieren oder rekonstruieren (vgl. Astbury & Leeuw, 2010). Dabei gilt: Je elaborierter eine Programmtheorie ausgearbeitet ist und je besser die theoretischen Annahmen zur Wirk-

samkeit – also die aufgestellten Wirksamkeitsregeln – durch empirische Evidenz belegt sind, desto wahrscheinlicher ist es, dass die Theorie auch in der Alltagspraxis funktioniert (vgl. Janssens & Wolf, 2009). Im Rahmen von Wirksamkeitsstudien wird der Begriff *Theorie* jedoch häufig als Synonym für (Transformations-)*Mechanismen* verstanden (vgl. Astbury & Leeuw, 2010), was wiederum bedeutet, „that one is doing one species of analytic evaluation … which involves no theory in anything like a proper use of that term" (Scriven, 1998, S. 59).

Wie Weber (2006) betont, birgt ein solches Theorieverständnis die Gefahr, dass die inhärent-theoretischen Annahmen einem einfach darstellbaren Kausalmodell untergeordnet werden, das dann jedoch die Wirkungsbedingungen unterschätzt. Programmtheorien müssen daher, wie Rogers (2008) in Anlehnung an Glouberman und Zimmerman (2002) zeigt, um angemessene kausalanalytische Schlüsse zu ermöglichen, mindestens drei Formen von Interventionen anhand ihrer Komplexität und Kompliziertheit unterscheiden: a) einfache Interventionen, die voll standardisiert ablaufen und durch einen einzelnen Akteur oder eine einzelne Organisation implementiert werden, b) komplizierte Interventionen, die aus mehreren Komponenten bestehen und z. B. multidimensional angelegt sind und/oder nur im Zusammenspiel mit anderen Interventionen Wirkung zeigen und c) komplexe Interventionen, die durch verschiedene Akteure oder Organisationen implementiert werden, die in Abhängigkeit von den gegebenen Voraussetzungen unterschiedliche Wirkungen entfalten können und daher sowohl abhängig von den Ausgangs- als auch den Kontextbedingungen sind, unter denen die Intervention stattfindet, und sich darüber hinaus adaptiv verhalten.

Die Möglichkeit des kausalen Schließens variiert dann wiederum in Abhängigkeit von der Interventionsform (vgl. Stame, 2010): So reicht es bei einfachen Interventionen aus anzunehmen, dass diese für beobachtete Effekte verantwortlich sind. Bei komplizierten Interventionen kann es hingegen sein, dass eine Komponente der Intervention nicht ausreicht, um Effekte nach sich zu ziehen oder diese nicht die einzige Ursache für Wirkungen darstellt. Bei komplexen Interventionen kann es wiederum dazu kommen, dass zirkuläre bzw. rekursive Kausalitäten Effekte nach sich ziehen. Das heißt, die Bedingungen, die der Intervention vorausgehen, stellen Voraussetzungen für weitere Erfolge dar und bieten die Chance, dass Interventionen in eine Entwicklungsspirale münden, sie stellen aber gegebenenfalls auch einen Hinderungsgrund dafür dar, dass eine Intervention einen Erfolg zeigt (vgl. Rogers, 2008).

Ausgearbeitete Programmtheorien zur Wirksamkeit von Schulinspektionen liegen bislang nur vereinzelt vor. So hat Ehren, begleitet durch verschiedene Kolleginnen und Kollegen (vgl. Ehren et al. 2005; Ehren & Honingh, 2011), die für die niederländische Schulinspektion angenommenen Wirkungsmechanismen mittels einer policy-orientierten Ex-Ante-Evaluation der zugrunde liegenden Programmtheorie herausgearbeitet. Gleichwohl gilt dieses Modell nur für den speziellen Fall der niederländischen Schulinspektion, die den Ansatz einer Risikoanalyse nutzt. Das bislang einzige Modell mit Anspruch auf Generalisierbarkeit, das eine Beschreibung der Wirkungsmechanismen von Schulinspektion im Allgemeinen ermöglichen soll, haben Ehren und Visscher (2006) vorgeschlagen. Diesem Modell zufolge entstehen Wirkungen von Schulinspektionen als Folgen einer kausalen Wirkungskette aus a) Merkmalen des Schulinspektionsprozesses, b) Reaktionen der Schulen auf den Prozess und die Ergebnisse der Inspektion und

c) den hieraus resultierenden Effekten. Darüber hinaus zeigen die Autoren entsprechend Befunden von Leithwood, Jantzi und Mascall (2002), dass mit Blick darauf, wie Schulen mit den zurückgemeldeten Ergebnissen umgehen, innerschulische Merkmale sowie externe Impulse und Unterstützungsmaßnahmen eine wichtige Rolle spielen.

Aufseiten der Inspektion haben demnach vor allem a) die reziproke Beziehung, die im Evaluationsprozess zwischen den Mitarbeitern der Schulinspektion und den Schulbeteiligten aufgebaut wird, b) der Kommunikationsstil der Inspektoren und c) die Art und Form der Rückmeldungen (spiegelt die Rückmeldung z. B. ein akkurates Bild der evaluierten Schule wider?) einen Einfluss darauf, ob die Inspektion Wirkungen entfaltet. Mit Blick auf den Umgang der Schule mit den Ergebnissen aus der Evaluation ist es dabei einerseits aufseiten der Schulen wichtig, in welchem Maße die Schulbeteiligten bereit für Veränderungen sind und sich die Schule als lernende Organisation versteht. Andererseits ist es aufseiten der externen Impulsgebung relevant, in welchem Maße z. B. finanzielle oder personelle Unterstützungsmaßnahmen für Schulen bereitgestellt werden oder aber auch Druck ausgeübt wird, um Veränderungen zu forcieren. Diese Faktoren beeinflussen letztlich wiederum, wie Schulen mit den Ergebnissen aus Schulinspektionsverfahren umgehen und ob intendierte Entwicklungsziele auch tatsächlich erreicht werden oder gar nicht-intendierte Nebeneffekte aus der Evaluation durch Schulinspektion resultieren.

Bislang nutzen nur sehr wenige Autoren ein solch ausdifferenziertes Design, um die Wirkungsmechanismen von Schulinspektionen zu beschreiben. In der Regel wird auf einfach-lineare Modelle der Kausalanalyse zurückgegriffen. Dass ein solch simpler Ansatz jedoch problematisch ist, zeigen Ehren und Visscher (2008) in einer Fallstudie an zehn niederländischen Schulen, die anhand ihres Innovationspotenzials ausgewählt wurden, auf. Hier konnten die Autoren deutlich machen, dass Veränderungen auf Schulebene als Reaktion auf eine Schulinspektion nur dann entstehen, wenn komplexe Bedingungsgefüge gewährleistet sind und z. B. sowohl ein hohes Innovationspotenzial der Schule als auch ein inspektionsseitiges Rückmeldeformat, welches auf die Schwächen der inspizierten Schule fokussiert, beobachtet werden können. Entsprechend kritisieren die Autoren die derzeit gängige Annahme, dass die Intervention durch Schulinspektionen allein und direkt zu einer Qualitätssteigerung an Schulen führe, als naiv, da die gängigen Analysemodelle der Wirksamkeitsforschung und die damit einhergehenden Wirksamkeitsannahmen zu unterkomplex seien und insbesondere Kontextbedingungen sowie Interdependenzen nicht berücksichtigen würden. Ehren et al. (2005) wiederum verweisen darauf, dass zwischen den bildungspolitischen Erwartungen, die an Schulinspektionsverfahren gestellt werden, und der konkreten Implementierung der Verfahren eine solch erhebliche Diskrepanz bestehe, dass es notwendig erscheine, die realen (und nicht die hypothetischen Ideal-)Bedingungen auszumodellieren, unter denen Schulinspektion Wirkung erzielen muss, um so die Wirksamkeit von Schulinspektionsverfahren im Kontext ihrer realen und nicht ihrer hypothetischen Möglichkeiten zu prüfen. Entsprechend können Ehren und Honingh (2011) im Rahmen der Ex-ante-Rekonstruktion der Programmtheorie der niederländischen Schulinspektion nachweisen, dass die Modellannahmen zur Wirksamkeit von Schulinspektionen, in Abhängigkeit von den jeweils genutzten bildungspolitischen Paradigmen, systematisch variieren und verschiedene Faktoren je nach Inspektionsmodell einen unterschiedlich starken Einfluss darauf haben, ob Inspektionen kausale Effekte nach sich ziehen können oder nicht.

Programmfehler als Gegenstand der Evaluation von Inspektionseffekten

Studien, die sich mit Programm- oder Implementationsfehlern beschäftigen, gehen der Frage nach, ob ein implementiertes Programm nicht gut oder nicht korrekt durchgeführt wird. Es geht hier darum, zu untersuchen, inwieweit es gelingt, ein (Bildungs-) Programm in seinem spezifischen Kontext praktisch umzusetzen. Im Fokus stehen dabei die Akkuratheit, mit der ein theoretisch begründetes Programm im Alltag umgesetzt wird und die dabei beobachteten Abweichungen von normativen und empirisch begründeten Vorgaben (vgl. Leeuw, 2012). Nicht nachweisbare Effekte werden in diesem Ansatz entsprechend auf eine mangelhafte oder ineffiziente Durchführungs- respektive Implementationspraxis zurückgeführt (vgl. Stame, 2010). Bei Studien zur Evaluation von Programmfehlern geht es entsprechend entweder darum herauszufinden, ob und in welchem Ausmaß eine Maßnahme Effekte nach sich zieht, um auf diesem Wege auf die Akkuratheit der Umsetzungspraxis zurückzuschließen (vgl. ebd.). Oder es geht darum, einzelne Aspekte des implementierten Programms en détail auf Abweichungen von normativen Vorgaben zu prüfen, um auf diesem Wege systematisch zu evaluieren, welcher Teilaspekt ursächlich für einen Effekt oder für ein Ausbleiben desselbigen ist (vgl. Scriven, 2008).

Wird von Ergebnissen auf potenzielle Fehler des Programms geschlossen, so ist eine explizite Ausmodellierung des (gegebenenfalls komplexen) Wirkungszusammenhangs nicht immer notwendig. Inferenzstatistisch kann mittels mehr oder weniger elaborierter statistischer Verfahren darauf geschlossen werden, ob ein Treatment einen intendierten Effekt nach sich zieht (vgl. Scriven, 1994). Leitend ist dabei die Fragestellung: Was wäre geschehen, wenn es kein Treatment gegeben hätte? Diese vergleichsweise atheoretische Herangehensweise an Evaluation nimmt dementsprechend Inputs und Outputs in den Fokus, ohne jedoch die Transformationsmechanismen, mittels derer Inputs in Outputs verwandelt werden, zu berücksichtigen (vgl. Chen, 1990). Eine derartige Kausalanalyse ermöglicht herauszufinden, ob eine Intervention eine empirisch nachweisbare Wirkung nach sich zieht, jedoch nicht zu klären, wie und warum dies geschieht. *Blackbox*-Evaluations-Modelle dienen entsprechend ausschließlich dazu, abzuschätzen, ob ein Programm in Gänze gelingt oder misslingt.

Verschiedene Autoren haben in den vergangenen Jahren jedoch betont, dass bei kausalen Fragestellungen eine Alternative zu einem experimentellen Input-Output-Forschungsdesign darin besteht, eine Mischung aus theoriebasierter Evaluation und dem systematischen Ausschluss alternativer Erklärungen zu nutzen (vgl. Rogers, 2008; Scriven, 2008, 2009, Stame, 2010, Weiss, 2002). Um mittels dieses Ansatzes festzustellen, ob zwischen einer Intervention und einer abhängigen Variable ein kausaler Zusammenhang besteht – Effekte in der abhängigen Variablen also der Intervention zugeordnet (*attributed*) werden können –, sind vier Bedingungen zu erfüllen (Weiss, 2002, S. 217):

> (1) responsiveness of the outcome (that is, the outcome follows the putative cause), (2) the elimination of plausible alternative explanations, (3) identification of the mechanism that yields the outcome, and (4) replication of the results.

Insbesondere in komplexeren Situationen ist dieses Verfahren sinnvoll, um systematisch zu prüfen, welcher mögliche Grund tatsächlich ursächlich für einen Effekt ist (vgl. Scriven, 2008, 2009). Hierfür muss dann, so Scriven (2008), a) eine Übersicht erstellt werden, die mögliche Ursachen für Effekte beinhaltet, b) im Sinne eines Modus Operandi beschrieben werden, wie diese möglichen Ursachen zu Effekten führen und c) im Rahmen der empirischen Analyse das Vorhandensein respektive das Nicht-Vorhandensein der einzelnen Modi Operandi geprüft werden. Diejenigen Modi Operandi, die als mögliche Ursache übrig bleiben, sind dann mit hoher Wahrscheinlichkeit verantwortlich für den beobachteten Effekt. Erst wenn alle Modi Operandi einer potenziellen Ursache auffindbar sind, ist jedoch davon auszugehen, dass diese wirklich den beobachteten Effekt nach sich ziehen. Bei der Evaluation von Interventionen mittels eines solchen Ansatzes muss es dann entsprechend darum gehen, auf Basis eines theoretischen Modells durch den systematischen, empirisch gestützten Ausschluss rivalisierender Erklärungen inferentiell auf die wahrscheinlichste und somit beste Erklärung für Effekte zu schließen (vgl. Cook, Scriven, Coryn & Evergreen, 2010).

Mit Blick auf die genutzten kausalanalytischen Forschungsmodelle zur Wirksamkeit von Schulinspektion werden derzeit vor allem *Blackbox*-Verfahren eingesetzt. Es wird geprüft, ob die Intervention Schulinspektion einen direkten Einfluss auf bestimmte Effektvariablen hat. Untersucht werden Effekte auf Einstellungen und Handlungsabsichten von Schulbeteiligten, direkte Reaktionen auf eine Schulinspektion auf Ebene der Einzelschule und die Veränderung von Schülerleistungen (vgl. Husfeldt, 2011; Wolf & Janssens, 2007). Eine Studie, die ausgewählte Modi Operandi aufseiten der Schulinspektion untersucht, wurde von Pietsch (2011) vorgelegt. Komplexere Analysen, die die einzelnen Teilbereiche z. B. mittels Strukturgleichungsmodellen kombinieren, liegen derzeit nicht vor. Mit Blick auf die Entwicklung von Schülerleistungen werden in der Regel Testleistungen re-analysiert. Insgesamt liegen bislang sechs Studien zum Zusammenhang von Schulinspektionsinterventionen und Schülerleistungen vor (vgl. Cullingford & Daniels, 1999; Luginbuhl et al., 2009; Matthews & Sammons, 2004; Rosenthal, 2004; Shaw et al., 2003; Wilcox & Gray, 1996). So zeigten Cullingford und Daniels (1999), Shaw et al. (2003) und Rosenthal (2004) mithilfe von Daten des General Certificate in Education (GCSE), dass Schulinspektionen in England einen negativen Einfluss auf die Schülerleistungen im Jahr nach einer Inspektion ausüben. Zwar berichten Matthews und Sammons (2004) einen gegenteiligen Befund, der jedoch auf die allgemeine Steigerung der Schülerleistungen im englischen Schulsystem und nicht auf die Intervention Schulinspektion zurückzuführen ist (vgl. Husfeldt, 2011). Für die Niederlande wiederum wiesen Luginbuhl et al. (2009) nach, dass Schulinspektionen gegebenenfalls zu leichten Steigerungen der Schülerleistungen in der Größenordnung von zwei bis drei Prozent einer Standardabweichung in den ersten zwei Jahren nach einer Schulinspektion führen könnten. Nachdem die Autoren dieser Studie jedoch ein Alternativdesign anlegten, um eventuelle Stichprobenverzerrungen zu kontrollieren, ließen sich auch hier keine Effekte von Schulinspektionsbesuchen auf Schülerleistungen mehr nachweisen.

Methodenfehler als Gegenstand der Evaluation von Inspektionseffekten

Studien zu Methodenfehlern setzen sich mit der Fragestellung auseinander, ob ein theoretisch angenommener Kausaleffekt nicht nachgewiesen werden kann, weil der Zusam-

menhang mit unzureichenden oder unangemessenen empirischen Methoden untersucht wurde. Ziel dieser Untersuchungen ist es zu klären, ob die genutzten Analysemethoden geeignet sind, um nachzuweisen, dass potenzielle Veränderungen in der interessierenden Effektvariablen durch eine Intervention erzeugt wurden (vgl. Stame, 2010). Hier stellen sich Fragen zu guter wissenschaftlicher Praxis, die sowohl Anlage als auch Durchführung einer Untersuchung sowie die nachfolgende Datenanalyse, Interpretation und Berichtlegung der erhobenen Daten betreffen, wie sie z. B. in den Standards des Joint Committee on Standards for Educational Evaluation (1999) definiert sind. Es geht daher beispielsweise um Fragen wie: „Ermöglicht das genutzte Stichprobendesign verlässliche Aussagen auf Populationsebene?", „Erfolgte die Datenerhebung in Subpopulationen der Stichprobe zu gleichen Bedingungen?" und „Werden Kausalanalysen nach dem methodischen State-of-the-Art durchgeführt?". Bei Studien, die anfallende Daten – wie in der Auseinandersetzung mit Schulinspektionseffekten üblich – re-analysieren, ist vor allem der letzte Punkt relevant, da sich die Planung und Durchführung der Intervention den Forschenden zumeist entzieht (vgl. Wolf & Janssens, 2007). Im Fokus stehen dabei in der Regel zum einen Fragen nach der statistischen Signifikanz und zum anderen der Umgang mit Selektionseffekten in Stichprobendesigns. So kann z. B. eine große Stichprobe dazu führen, das ein Typ-I-Fehler (die Nullhypothese wird abgelehnt; es wird berichtet, dass ein Programm funktioniert, obwohl es nicht funktioniert) auftritt (vgl. z. B. Stame, 2010), während eine komplexe Modellierung des Kausalzusammenhangs mit messfehlerbehafteten Variablen Typ-II-Fehler (die Nullhypothese wird angenommen; es wird berichtet, dass ein Programm nicht funktioniert, obwohl es funktioniert) nach sich ziehen kann (vgl. z. B. Weiss, 1998).

Bei Fragen zu Selektionseffekten wiederum geht es darum, herauszufinden, ob Stichproben zufallsbasiert erhoben wurden, und falls nicht, ob die Nichtzufälligkeiten im Rahmen der Analysen angemessen ausmodelliert wurden. Mit Blick auf die Wirksamkeit von Schulinspektionen auf schulische Outputs steht grundsätzlich die Frage im Hintergrund, ob experimentelle oder quasi-experimentelle Attributionsanalysen (*attribution analysis,* vgl. White, 2010) durchgeführt wurden, um zu klären, ob, und falls ja, in welchem Maße, messbare Veränderungen in Schülerleistungen auf die Intervention Schulinspektion zurückzuführen sind. Ein solcher Ansatz der Kausalanalyse beruht generell auf der kontrafaktischen Annahme (vgl. ebd.), wonach eine Wirkung definiert wird als der Unterschied in der interessierenden Variablen Y mit der Intervention Y_1 und ohne die Intervention Y_0. Mit anderen Worten: Für jede Schule gibt es zwei potenziell mögliche Ergebnisse: y_i^1, wenn eine Schulinspektion durchgeführt wurde und y_i^0, wenn keine Schulinspektion stattgefunden hat. Für jede Schule lässt sich dann die Wirkung der Schulinspektion als Differenz zwischen den beiden potenziellen Ergebnissen y_i^1-y_i^0 definieren. Wichtig ist hierbei, dass die potenziellen Ergebnisse der einzelnen Schule nur davon abhängen, ob die jeweilige Schule inspiziert wurde oder nicht, aber nicht davon, ob auch die anderen Schulen inspiziert wurden (*stable unit treatment assumption* – SUTVA, vgl. Morgan & Winship, 2007).

Das grundlegende Problem der Kausalanalyse besteht jedoch darin, dass nicht beide potenziellen Ergebnisse gleichzeitig auftreten können – für jede Schule lässt sich nur eines der theoretisch möglichen Ergebnisse beobachten, entweder es wurde im entsprechenden Zeitraum eine Schulinspektion an der Schule durchgeführt oder eben nicht.

Bei dem jeweils anderen Ergebnis handelt es sich daher um ein „unbeobachtetes, *kontrafaktisches* Ergebnis im Sinne einer ‚was-wäre-wenn-Frage'" (Legewie, 2012, S. 127). So lässt sich die Wirkung der Schulinspektion auf die einzelne Schule niemals direkt messen. Man kann dieses Problem am einfachsten umgehen, indem man die Ergebnisse von Schulen *mit* einem Besuch durch die Schulinspektion mit denen von Schulen *ohne* Besuch durch die Schulinspektion vergleicht. Die Leistungsunterschiede zwischen diesen beiden Gruppen kann man dann als durchschnittlichen kausalen Effekt interpretieren. Voraussetzung für diese Interpretation ist allerdings, dass es zwischen den Schulen mit und ohne Schulinspektion keine weiteren Unterschiede gibt, die auch mit der abhängigen Variable, also der durchschnittlichen Schülerleistung, in Zusammenhang stehen. Diese Annahme ist bei Beobachtungsdaten in der Regel nicht vertretbar. Eine Lösung für dieses Selektionsproblem bieten Zufallsexperimente.

Zufallsexperimente sind dadurch gekennzeichnet, dass der Treatmentstatus (in diesem Fall: die Durchführung der Schulinspektion) den Beobachtungseinheiten (also den Schulen) zufällig zugewiesen wird und somit jede Schule die gleiche Chance hat, zur Kontroll- oder zur Treatmentgruppe zu gehören. Auf diese Weise sollten sich die Kontroll- und Treatmentgruppe – abgesehen von zufälligen Differenzen – nur im Hinblick auf den Treatmentstatus unterscheiden. Die Entscheidung, an welcher Schule eine Schulinspektion im betreffenden Zeitraum durchgeführt wurde, sollte nicht aufgrund der Leistungen, die in dieser Schule vorher erbracht wurden oder auf der Grundlage anderer Merkmale der Schule gefällt worden sein, sondern rein zufällig. In diesem Fall lässt sich der durchschnittliche kausale Effekt der Schulinspektion durch einen einfachen Mittelwertvergleich zwischen den Schulen mit und ohne Besuch durch die Schulinspektion berechnen. Stehen keine experimentellen Daten zur Verfügung, können verschiedene statistische Verfahren angewandt werden, um zufällige Variationen im Treatmentstatus zu identifizieren und auf der Grundlage dieser Variationen den kausalen Effekt zu schätzen. Dies wird üblicherweise durch Schätzung des Treatmenteffekts nach der Kontrolle beobachtbarer Variablen erreicht und umfasst Verfahren wie z. B. das aktuell viel beachtete Propensity-Score-Matching (vgl. Rosenbaum & Rubin, 1983).

Studien, die sich mit der methodischen Angemessenheit von Evaluationen zur Wirksamkeit von Schulinspektionsverfahren auseinandersetzen, sind ausgesprochen rar. Einzig Wolf und Janssens (2007) haben sich der Frage gewidmet, ob die Methoden, mit deren Hilfe in den letzten Jahren auf die Wirksamkeit von Schulinspektion geschlossen wurde, überhaupt geeignet sind, um kausale Zusammenhänge zu modellieren. Dabei kamen die Autoren zu dem Schluss, dass es in allen bis zum damaligen Zeitpunkt durchgeführten Studien problematisch ist, dass die genutzten Daten nicht auf Zufallsstichproben von Schulen basieren. Entsprechend konnte in diesen Untersuchungen nicht abgesichert werden, dass Befunde nicht arbiträr und dem Vorhandensein unzureichender Stichproben geschuldet waren. Darüber hinaus wurden auch nahezu nirgends statistische Verfahren zur Schätzung des Treatmenteffekts nach der Kontrolle beobachtbarer Variablen eingesetzt. Einzig die Studie von Rosenthal (2004), die Ergebnisse im Sinne eines Regressions-Ansatzes generiert, indem kontrolliert wird, ob ein systematischer Zusammenhang zwischen der Auswahl einer Schule für die Inspektion und den Schülerleistungen im Jahr vor der Inspektion besteht, erfüllt den Autoren zufolge grundlegende Kriterien

an eine Kausalanalyse, wenngleich auch diese Studie methodische Schwächen aufweise. Entsprechend resümieren Wolf und Janssens (2007, S. 392):

> A problem with the existing studies into the effects ... is that (a) the findings are ambiguous, (b) the research methodology varies substantially and is not always appropriate for testing causal effects and (c) the findings appear to be closely linked to the research methodology used ... Exposing a causal relationship ... makes considerable demands on the research design.

Seit der Veröffentlichung des Reviews von Wolf und Janssens (2007) ist nur eine weitere Studie hinzugekommen, die den Zusammenhang zwischen Schulinspektionen und Schülerleistungen modelliert. Luginbuhl et al. (2009) haben hier einerseits ein Design gewählt, das demjenigen der Rosenthal-Studie ähnelt, wobei jedoch ein Fixed-Effekt-Modell genutzt wurde, um zu kontrollieren, ob die Inspektion einer Schule von der Ausgangsleistung der jeweiligen Schülerschaft abhängt. Andererseits nutzten die Autoren eine kleine Zufallsstichprobe der niederländischen Schulinspektion, um Effekte im Sinne eines klassischen Zufallsexperiments zu analysieren. Während die Autoren mithilfe des Fixed-Effekt-Modells zu dem Ergebnis kamen, dass Schulinspektion in den Niederlanden positive Ergebnisse nach sich zieht, konnte dieser Befund mittels des Zufallsexperiments nicht bestätigt werden. Da das genutzte Zufallsdesign besser geeignet sei, um Kausaleffekte zu testen, so Luginbuhl et al. (2009), müsse man konstatieren, dass die Durchführung von Schulinspektionen zwar nicht schaden würde, jedoch keinen oder nur einen sehr geringen Einfluss auf Schülerleistungen habe.

3 Synthese

Obwohl mittlerweile etliche Studien zur Arbeit und Wirksamkeit von Schulinspektionen vorliegen, sind die vorhandenen Befunde zur Wirksamkeit auf Schülerleistungen von sehr heterogener Qualität, Aussagekraft und Belastbarkeit und ergeben entsprechend kein kohärentes Ganzes. Mit Blick auf den Bereich der genutzten Theorien zeigt sich, dass diese bislang nicht ausreichend ausmodelliert wurden und vor allem die Komplexität und die Kompliziertheit der Intervention unterschätzen. Programmfehler wurden zwar untersucht, die durchgeführten Studien genügen jedoch dem methodischen State-of-the-Art empirischer Kausalanalysen zumeist nicht, und die Befunde können somit ebenso gut methodisch-bedingten Selektionseffekten wie tatsächlichen Programmfehlern geschuldet sein.

Dass die Vernachlässigung von Komplexität und Kompliziertheit der Intervention Schulinspektion in den bislang vorliegenden Studien zur Wirkung von Schulinspektionen nicht berücksichtigt wurde, könnte dabei durchaus erklären, warum in der Literatur zur messbaren Wirkung von Schulinspektionen in der Regel keine oder nur geringe Effekte auf Schülerleistungen berichtet werden. Die wenigen Studien, die sich bisher mit dieser Problematik im Rahmen von Schulinspektionsverfahren beschäftigt haben, deuten dann auch darauf hin, dass bei Berücksichtigung dieser kontextuellen Bedingungen zwar durchaus damit zu rechnen wäre, dass Schulinspektionen Wirkung zeigen, jedoch

nur unter bestimmten Voraussetzungen und bei unterstützenden Begleitumständen. Verwunderlich ist dies nicht: Denn wie Glouberman und Zimmerman (2002) betonen, ist insbesondere der Bereich des organisationalen Lernens komplexer und nicht-linearer Natur und erfordert eine entsprechende Formulierung und Modellierung in Kausalanalysen.

Vergleicht man die Forschung zur Wirksamkeit von Schulinspektion auf Schülerleistungen mit anderen, bereits etablierten Feldern der bildungswissenschaftlichen Wirkungsforschung, so muss man daher wohl Nachfolgendes konstatieren: Die Forschung zur Effektivität von Schulinspektionsverfahren befindet sich derzeit unserer Einschätzung nach in etwa im Stadium der Schuleffektivitätsforschung der späten 1960er bis frühen 1970er Jahre, in der mittels ökonomischer Studien – die einem reinen Input-Output-Paradigma folgten und weder Prozess- noch Kontextvariablen berücksichtigten (vgl. Reynolds & Teddlie, 2000) – diagnostiziert wurde, dass „schools bring little influence to bear on a child's achievement" (vgl. Coleman et al., 1966, S. 325).

4 Fazit und Diskussion

Die Probleme der bislang vorliegenden Studien liegen vor allem im Bereich der Theorien und der angewandten Methoden: a) die aktuell genutzten Programmtheorien sind zu unterkomplex und berücksichtigen darüber hinaus kaum die Kompliziertheit der Intervention, b) die eingesetzten Forschungsmethoden wiederum sind kausalanalytischen Fragestellungen zumeist nicht angemessen und berücksichtigen bekannte methodologische Fallstricke nicht oder nur ungenügend. Dies wiederum hat zur Folge, dass Aussagen zu Programmfehlern nur schwer möglich sind und die bislang vorliegenden Befunde wenig verlässliche Aussagen zur Wirksamkeit von Schulinspektionsinterventionen auf Schülerleistungen ermöglichen.

Um zu evaluieren, ob Schulinspektion auf Schülerleistungen wirkt oder nicht, sollten daher zukünftig zweierlei Desiderata angegangen werden. Zum einen bedarf es einer Zusammenführung der bislang vorliegenden theoretischen Annahmen zur Wirkungsweise von Schulinspektion mit weiteren theoretischen Annahmen, wie sie z. B. aus der Schul- und Unterrichtsentwicklung aber auch aus der Forschung zur Rezeption und Verarbeitung von Rückmeldungen aus Evaluationen vorliegen. Dabei sollte unbedingt beachtet werden, dass es wichtig ist, dass *„theory-driven evaluators do not replace substantive social and behavioral science theory with a focus only on putative mechanisms"* (Astbury & Leeuw, 2010, S. 375). Eine handhabbare Technik, um die benötigten Modelle sukzessive zu entwickeln und auszudifferenzieren, kann dabei Scrivens (2008, 2009) *General-Elemination*-Methode (GEM) darstellen. Ein besonderer Vorteil dieses Ansatzes besteht dabei darin, dass er es ermöglicht, Befunde aus qualitativen und quantitativen Studien integrativ zu nutzen (Reichardt, 2011).

Zum anderen ist es wichtig, in Zukunft verstärkt belastbare empirische Evidenz zu generieren, um Hinweise darauf zu erhalten, ob die bundesweit eingesetzten Programme zur externen Evaluation von Schulen in Gänze eher miss- oder gelingen. Solange keine ausgearbeiteten Programmtheorien zur Prüfung vorliegen, scheint es geboten, hierfür *Blackbox*-Verfahren einzusetzen. Denn einerseits ist es aus einer abnehmerorientier-

ten Nutzenperspektive weniger relevant, wie Schulinspektion wirkt, sondern ob sie ihre Funktion (z. B. Schulinspektion soll zu verbesserten Schülerleistungen führen) erfüllt (vgl. Scriven, 1994, 1998). Andererseits besteht die Gefahr, dass, solange keine ausgearbeiteten, validen und empirisch prüfbaren Modelle vorliegen, im Rahmen von Studien zur Wirksamkeit von Schulinspektionsverfahren ad-hoc-Theorien aufgestellt werden, die dem Untersuchungsgegenstand nicht angemessen sind und entsprechend zu Fehlschlüssen führen können (vgl. Stufflebeam & Shinkfield, 2007). In einem solch methodengesteuerten (*method driven*) Evaluationsansatz kommt dann jedoch dem angemessenen Einsatz kausalanalytischer Forschungsmethoden eine besondere Bedeutung zu.

Literatur

Astbury, B. & Leeuw, F.L. (2010). Unpacking black boxes. Mechanisms and theory building in evaluation. *American Journal of Evaluation, 31* (3), 363–381.

Böttcher, W. & Kotthoff, H.-G. (2007). Gelingensbedingungen einer qualitätsoptimierenden Schulinspektion. In W. Böttcher & H.-G. Kotthoff (Hrsg.), *Schulinspektionen: Evaluation, Rechenschaftslegung und Qualitätsentwicklung* (S. 223–230). Münster: Waxmann.

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., Holtappels, H.-G. & Rösner, E. (2006). Schulinspektionen in den deutschen Bundesländern – eine Baustellenbeschreibung. In W. Bos, H.G. Holtappels, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung, Bd. 14. Daten, Beispiele und Perspektiven* (S. 81–124). Weinheim: Juventa.

Chen, H.-T. (1990). *Theory-driven evaluations.* Newbury Park: Sage.

Chen, H.-T. (2005). *Practical program evaluation: Assessing and improving, planning, implementation and effectiveness.* Thousand Oaks: Sage.

Chen, H.-T. (2006). A theory-driven evaluation perspective on mixed methods research. *Research in the Schools, 13* (1), 75–83.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. & York, R. (1966). *Equality of educational opportunity.* Washington: U.S. Government Printing Office.

Cook, T.D., Scriven, M., Coryn, C.L.S. & Evergreen, S.D.H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31* (1), 105–117.

Coryn, C.L.S., Noakes, L.A., Westine, C.D. & Schröter, D.C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32* (2), 199–226.

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

Ditton, H. (2010). Evaluation und Qualitätssicherung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 607–623). Wiesbaden: VS Verlag für Sozialwissenschaften.

Döbert, H., Rürup, M. & Dedering, K. (2008). Externe Evaluation von Schulen in Deutschland – die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede. In H. Dö-

bert & K. Dedering (Hrsg.), *Externe Evaluation von Schulen. Historische, rechtliche und vergleichende Aspekte* (S. 63–152). Münster: Waxmann.

Ehren, M.C.M. & Honingh, M.E. (2011). Risk-based school inspections in the Netherlands: A critical reflection on intended effects and mechanisms. *Studies in Educational Evaluation, 37* (4), 239–248.

Ehren, M.C.M., Leeuw, F.S. & Scheerens, J. (2005). On the impact of the Dutch educational supervision act: Analysing assumptions concerning the inspection of primary schools. *American Journal of Evaluation, 26* (1), 60–76.

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.

Ehren, M.C.M. & Visscher, A.J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies, 56* (2), 205–227.

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.

Glouberman, S. & Zimmerman, B. (2002). *Complicated and complex systems: What would successful reform of medicare look like?* Ottawa: Commission on the Future of Health Care in Canada.

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

Janssens, F.J.G. & Wolf, I.F. de (2009). Analyzing the assumptions of a policy program: An ex-ante evaluation of 'Educational Governance' in the Netherlands. *American Journal of Evaluation, 30* (3), 411–425.

Joint Committee on Standards for Educational Evaluation. (1999). *Handbuch der Evaluationsstandards.* Opladen: Leske + Budrich.

Köller, O. (2009). Evaluation pädagogisch-psychologischer Maßnahmen. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 331–351). Heidelberg: Springer.

Leeuw, F.L. (2012). Linking theory-based evaluation and contribution analysis: Three problems and a few solutions. *Evaluation, 18* (3), 348–363.

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.

Leithwood, K., Jantzi, D. & Mascall, B. (2002). A framework for research in large scale reform. *Journal of Educational Change, 3* (1), 7–33.

Lipsey, M., Crosse, S., Dunkle, J., Pollard, J. & Stobart, G. (1985). Evaluation: the state of the art and the sorry state of the science. In D.S. Cordray (Ed.), *Utilizing Prior Research in Evaluation Planning* (pp. 7–28). San Francisco: Jossey-Bass.

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

Maritzen, N. (2007). Schulinspektion – ein neues Element der Systemsteuerung. *Journal für Schulentwicklung, 11* (3), 6–14.

Maritzen, N. (2008). Schulinspektionen – zur Transformation von Governance-Strukturen im Schulwesen. *Die Deutsche Schule, 100* (1), 85–96.

Maritzen, N. (2009). Schulinspektion und Schulaufsicht. In H. Buchen & H.-G. Rolff (Hrsg.), *Professionswissen Schulleitung* (S. 1368–1392). Weinheim: Beltz.

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.

Müller, S. (2010). Erste Effekte von Schulinspektion – eine Zwischenbilanz. In N. Berkemeyer, W. Bos, H.G. Holtappels, N. McElvany & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung, Bd. 16. Daten, Beispiele und Perspektiven* (S. 289–308). Weinheim: Juventa.

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

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.

Raudenbush, S.W. & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1* (2), 138–154.

Reichardt, C.S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation, 32* (2), 246–272.

Reynolds, D. & Teddlie, C. (2000). An introduction into school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 3–25). London: Falmer Press.

Rogers, P.A. (2002). Program theory: Not whether programs work but how they work. In D.L. Stufflebeam, G.F. Madaus & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 209–232). New York: Kluwer.

Rogers, P.A. (2008). Using program theories to evaluate complicated and complex aspects of interventions. *Evaluation, 14* (1), 29–48.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* (1), 41–55.

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.

Scheerens, J., Glas, C. & Thomas, S.M. (2003). *Educational evaluation assessment and monitoring. A systematic approach.* Lisse: Swets & Zeitlinger.

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

Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation, 19* (1), 57–70.

Scriven, M. (2008). A summative evaluation of RCT methodology & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5* (9), 11–24.

Scriven, M. (2009). Demythologizing causation and evidence. In S. Donaldson, C. Christie & M.M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 134–152). Los Angeles: Sage.

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.

Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation, 16* (4), 371–387.

Stufflebeam, D.L. & Shinkfield, A.J. (2007). *Evaluation theory, models, & applications.* San Francisco: Jossey-Bass.

Suchman, E.A. (1969). Evaluating educational programs. *Urban Review, 3* (4), 15–17.

van Ackeren, I. & Klemm, K. (2009). *Entstehung, Struktur und Steuerung des deutschen Schulsystems.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Weber, K. (2006). From nuts and bolts to toolkits: Theorizing with mechanisms. *Journal of Management Inquiry, 15* (2), 119–123.

Weiss, C.H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21* (4), 501–524.

Weiss, C.H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19* (1), 21–33.

Weiss, C.H. (2002). What to do until the random assigner comes? In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 98–224). Washington: Brookings Institution Press.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation, 16* (2), 153–164.

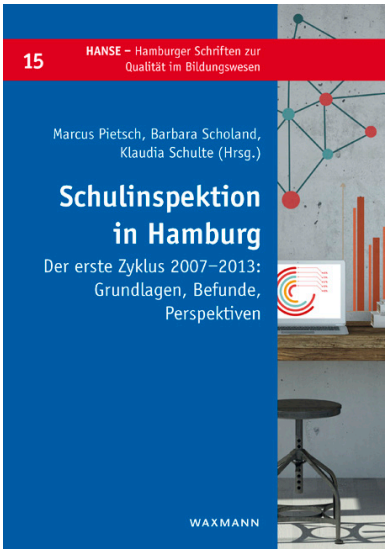
Wilcox, B. & Gray, J. (1996). *Inspecting schools: Holding schools to account and helping schools to improve.* Buckingham: Open University Press.

Wolf, I.F. de & 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.

Marcus Pietsch, Barbara Scholand,
Klaudia Schulte (Hrsg.)

Schulinspektion in Hamburg

Der erste Zyklus 2007–2013: Grundlagen, Befunde und Perspektiven



Das gedruckte Buch finden Sie hier.

Waxmann 2015
Münster • New York

Bibliografische Informationen der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten
sind im Internet über http://dnb.d-nb.de abrufbar.

HANSE – Hamburger Schriften zur Qualität im Bildungswesen, Band 15
ISSN 1864-2225
ISBN 978-3-8309-3278-9

Steinfurter Straße 555, 48159 Münster

www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Pleßmann Design, Ascheberg
Umschlagfoto: © Robert Kneschke – Fotolia.de
Lektorat und Satz: Judith Zimmer, Hamburg
Druck: Mediaprint, Paderborn

Gedruckt auf alterungsbeständigem Papier,
säurefrei gemäß ISO 9706

Wirkungen von Schulinspektion: Ein Rahmen zur theoriegeleiteten Analyse von Schulinspektionseffekten

Marcus Pietsch, Ann-Katrin van den Ham & Olaf Köller

Zusammenfassung

Schulinspektionen sollen die Qualität von Schule und Unterricht verbessern und hierdurch zur Steigerung von Schülerleistungen beitragen. Um die derzeitig gängigen Schulinspektionsverfahren wissensbasiert weiterzuentwickeln, intendierte Wirkungen mit höherer Wahrscheinlichkeit als bislang zu erzielen und somit die Wirksamkeit der Verfahren zu erhöhen, ist es notwendig, empirische Evidenz zu wirksamkeitsförderlichen Bedingungen und Vorgehensweisen zu sammeln. Eine gute Möglichkeit, dies zu erforschen, bieten theoriebasierte Evaluationen. Da die bisherige Forschung zu Wirkungen von Schulinspektionen eher kursorisch erfolgt, wird im Rahmen des vorliegenden Beitrages ein dreischrittiges Verfahren zur systematischen Evaluation von Schulinspektionseffekten vorgeschlagen. Hierbei wird insbesondere dafür plädiert, dass im Rahmen theoriegeleiteter Evaluation von Schulinspektionseffekten die Komplexität und die Kompliziertheit der Intervention stärker berücksichtigt werden als bisher und die im Rahmen programmtheoretisch fundierter Analysen generierten Befunde mittels des Ansatzes des Interpretation/Use Argument *nach Kane validiert werden, um so die Belastbarkeit der Studien und der daraus abzuleitenden Konsequenzen für die Weiterentwicklung von Inspektionsverfahren zu erhöhen.*

1. Einleitung

Schulinspektionen sollen dazu beitragen, Schülerleistungen zu verbessern. Dies, so die Erwartung, soll gelingen, indem sie Schulen durch die Rückmeldung beobachteter Stärken und Schwächen Impulse für die Weiterentwicklung von Schule und Unterricht liefern. Hierfür wird nach normativen Vorgaben, die in der Regel in landesspezifischen Qualitätsrahmen oder Qualitätstableaus formuliert wurden, durch Schulinspektoren extern an Schulen evaluiert. Es wird dann davon ausgegangen, dass Schulverantwortliche die zurückgemeldeten Informationen für eine wissensbasierte Schul- und Unterrichtsentwicklung nutzen, die sich wiederum in verbesserten Schülerleistungen niederschlägt. Wie Pietsch, Schulze, Schnack und Krause (2011) herausstellen, knüpfen die diesbezüglichen Wirksamkeitserwartungen an Schulinspektionen vor allem an die Forschung zum zielorientierten

Feedback (vgl. Kluger & DeNisi 1996; Ramaprasad 1983; Visscher & Coe 2003) an. Entsprechend wird erwartet, dass das empirisch begründete Aufzeigen von Differenzen zwischen normativ vorgegebenen Soll- und empirisch beobachteten Ist-Zuständen dazu führt, dass in extern evaluierten Schulen infolge der Rückmeldung eine Handlungsoptimierung geplant wird, die es ermöglicht, anzustrebende Ziele in Zukunft besser zu erreichen.

Eine Vielzahl von Studien zur Wirksamkeit von Schulinspektionen zeigt jedoch, dass dieses Ziel häufig nicht erreicht wird oder sogar nicht-intendierte Nebeneffekte beobachtbar sind. Die Befundlage ist hierbei sehr heterogen. So zeigen einzelne Studien, dass sich Effekte auf die innerschulische Qualitätsentwicklung finden lassen, und andere Studien negieren genau diesen Befund (vgl. zur Übersicht z. B. de Wolf & Janssens 2007; Dedering 2012; Husfeldt 2011). Auch mit Blick auf die Verbesserung von Schülerleistungen weisen bislang nur zwei Studien aus den Niederlanden und Deutschland positive Effekte nach (vgl. Pietsch et al. 2014; Luginbuhl et al. 2009). Diejenigen Studien hingegen, die sich auf Effekte der englischen Schulinspektion OFSTED *(The Office for Standards in Education)* beziehen, zeigen, 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). Darüber hinaus wird in einzelnen Studien von nicht-intendierten Nebenwirkungen bei Schulinspektionen berichtet. Demnach steigt den Befunden zufolge das Stresserleben bei Lehrkräften während einer Inspektion (vgl. z. B. Brunsden et al. 2006) oder es findet eine überzogen positive Selbstdarstellung der Schule *(Window Dressing)* statt (vgl. de Wolf & Janssens 2007).

Insgesamt findet sich somit auch zehn Jahre nach Einführung der ersten Schulinspektorate in Deutschland eine uneindeutige Befundlage, die keine kohärente Einschätzung darüber ermöglicht, ob Schulinspektionen ein Instrument sind, von dem erwartet werden kann, dass es zu einer Verbesserung von Schul- und Unterrichtsqualität und infolge dessen zu einer Steigerung von Schülerleistungen führt. Pietsch et al. (2014) haben vorgeschlagen, diesem Problem mithilfe einer Kombination kausalanalytischer Blackbox-Verfahren und theoriegeleiteter Evaluationen zu begegnen. Während die erste Herangehensweise Aufschluss darüber geben kann, ob Inspektion grundsätzlich die intendierten Effekte erzielt oder nicht, ist letztere insbesondere dann sinnvoll, wenn es um Entscheidungen zur weiteren Gestaltung, zu Optimierung, Veränderung oder Verbesserung von Inspektionsverfahren geht. Einen Punkt, den bereits Cronbach et al. (1980, S. 251) als besonders relevantes Kernelement von Evaluationen herausstellen:

> „Knowing this week's score does not tell the coach how to prepare next week's game. The information that an intervention had satisfactory or unsatisfactory outcomes is of little use by itself; users […] need to know what led to success or failure."

Während die Evaluation mithilfe von Blackbox-Verfahren zwar methodisch anspruchsvoll, ansonsten jedoch geradlinig und überschaubar ist, erfordert eine theoriegeleitete Evaluation von Inspektionswirkungen eine komplexere Herangehensweise, um zu belastbaren Resultaten zu gelangen (vgl. Astbury & Leeuw 2010). Im Folgenden wird daher ein Rahmen für eine theoriegeleitete Analyse von Schulinspektionseffekten vorgestellt, die es ermöglichen soll, Befunde zur Wirkung von Schulinspektionen systematisch zu generieren. Hierzu werden zuerst Annahmen zur Wirksamkeit und den Funktionen von Schulinspektionen beschrieben. Anschließend wird dargestellt, welche Gründe es aus programmtheoretischer Perspektive dafür geben kann, dass Schulinspektionen ggf. nicht wirksam werden, um anschließend auf ein strukturiertes Vorgehen zur Evaluation von Schulinspektionswirkungen einzugehen.

2. Annahmen zur Wirksamkeit und zu Funktionen bei Schulinspektionen

2.1 Funktionen von Schulinspektionen

Schulinspektion hat die Aufgabe, Informationen aus dem Bereich Bildung und Erziehung systematisch zu beschaffen und aufzubereiten, um auf diesem Wege Akkreditierungen, Rechenschaftslegungen und Diagnosen für systemisches Lernen im Bildungssystem zu ermöglichen. Schulinspektionen sollen entsprechend dazu beitragen, elementare Standards von Bildungsqualität zu gewährleisten, einen verbesserten Service für das einzelschulische Qualitätsmanagement zu bieten und unterschiedlichen Akteuren im Bildungssystem verwertbare Informationen zu schulischen Prozessqualitäten bereitzustellen (vgl. Döbert et al. 2008; Maritzen in diesem Band; Pietsch et al. 2009, 2013). Dabei liegt der Schwerpunkt der Inspektionsarbeit auf der Evaluation einzelschulischer Prozesse (vgl. Böttcher & Kotthoff 2010; Döbert et al. 2008; Maritzen in diesem Band; van Ackern & Klemm 2009), wobei die Inspektionen in den Ländern trotz augenscheinlicher Unterschiede in Ausstattung, Ausgestaltung und ministerieller Anbindung „auf Grundlage von Verfahren mit wiederkehrenden Standardelementen“ erfolgen (Maritzen in diesem Band, S. 16).

Grundlage für die externe Einzelschulevaluation bilden dabei in allen Ländern die landesspezifischen Qualitätsrahmen oder Qualitätstableaus, die im Sinne eines Input-Prozess-Output-Modells (teilweise unter Berücksichtigung von Kontextfaktoren) Anforderungen an Schulqualität definieren (vgl. Ehren & Scheerens in diesem Band; Müller 2010). In allen Inspektionen wurden, basierend auf diesen Qualitätskatalogen, Leistungsindikatoren definiert, die es ermöglichen sollen, die Qualität einzelner Prozesse möglichst differenziert zu erfassen (vgl. Böttcher

& Kotthoff 2007). Geschulte Experten – meist Lehrkräfte, Schulleitungen oder Schulaufsichten – evaluieren anhand dieser Indikatoren die Schulen in Teams von zwei bis vier Personen (vgl. Döbert et al. 2008; Müller 2010). Die Datenerhebung selbst erfolgt über ein umfangreiches Repertoire an Methoden. Fragebögen und Interviews, gerichtet an verschiedene Schulbeteiligte, sowie Schulbegehungen und die Begutachtung schulinterner Dokumente und Statistiken gehören derzeit ebenso wie die Beobachtung von Unterrichtseinheiten zum länderübergreifenden Methodenstandard (vgl. Döbert et al. 2008). Die derart ermittelten Befunde werden in einem Bericht zusammengefasst und den Schulen sowie ggf. weiteren Beteiligten, wie z. B. Schulaufsichten, übergeben. In regelmäßigem Turnus werden darüber hinaus Berichte verfasst, die Befunde der Schulinspektionen auf Systemebene präsentieren.

Durch diese Rückmeldungen aus Schulinspektionen erhalten die Schulbeteiligten, ebenso wie Bildungsadministration und Bildungspolitik, Informationen zur extern wahrgenommenen Qualität innerschulischer Prozesse auf verschiedenen Ebenen. Obwohl dies häufig nicht explizit benannt wird, steht hinter dem Vorgehen, aufbereitete Informationen an verschiedene Gruppen auf Schul- und Schulsystemebene zurückzumelden, die Idee, dass die Rückmeldungen von Evaluationsergebnissen als Grundlage für Entwicklungen an den evaluierten Schulen genutzt werden, die in ihrer Konsequenz in verbesserte fachliche Schülerleistungen münden (vgl. de Wolf & Janssens 2007; Ehren & Visscher 2006; Ehren et al. 2005). Insbesondere in den deutschen Ländern zeichnet sich diesbezüglich aktuell ein Trend ab (vgl. Böttcher & Kotthoff 2010, S. 307), „demzufolge Schulinspektionen […] als ein Unterstützungs- und Dienstleistungsangebot für die Schule verstanden werden, deren Erkenntnisse wirksam für die interne Schulentwicklung genutzt werden können bzw. sollen". Folgerichtig betont Köller (2009), dass der Wert von Schulinspektionen insbesondere darin läge, dass sie Hinweise auf die Prozesse vor Ort gäben, mithilfe derer Hypothesen darüber generiert werden könnten, welche Veränderungen von Schule und Unterricht zu höheren Lernerträgen der Schülerinnen und Schüler führen können.

2.2 Annahmen zur Nutzung von Inspektion als Grundlage für Schulentwicklung

Entsprechend wird davon ausgegangen, dass Schulverantwortliche die Inspektion und ihre Berichte nutzen, um hieraus konkrete Maßnahmen für die Schul- und Unterrichtsentwicklung abzuleiten (vgl. Ehren & Visscher 2006, 2008). Im Zentrum eines solch entwicklungsfokussierten, anwenderorientierten Verständnisses von Evaluation, wie sie den Schulinspektionsverfahren in den deutschen Bundesländern zugrunde liegt, steht daher zwangsläufig der Evaluationsnutzen, der sich aus

der Untersuchung der Qualität von Schule und Unterricht für die Schulbeteiligten ergibt. Dabei ist es von grundlegender Bedeutung zu wissen, in welcher Art und Weise Informationen aus Schulinspektionsverfahren innerhalb von Schulen genutzt werden können, wenn das Ziel Qualitäts- und Schulentwicklung heißt, da „das Verwendungsinteresse der […] Anwendersysteme […] unterschiedlich ist, je nachdem ob die Evaluation eher zur Überprüfung/Kontrolle, zur Legitimation, zur Innovation/Entwicklung oder lediglich zur Bestätigung durchgeführt wird" (vgl. Stamm 2003, S. 195 f.). Diesbezüglich lassen sich mit Blick auf die Nutzung von Evaluationsergebnissen grundsätzlich drei Nutzungsformen voneinander unterscheiden, die je einen spezifischen Zweck verfolgen (vgl. King & Pechman 1984; Weiss 1998):

1) Instrumentelle Nutzung *(Instrumental Use)*: Die instrumentelle Nutzung von Informationen aus Evaluationen bezieht sich darauf, konkrete Probleme zu lösen bzw. konkrete Entscheidung zu treffen, die den Evaluationsgegenstand betreffen.
2) Konzeptionelle Nutzung *(Conceptual Use)*: Die konzeptionelle Nutzung von Informationen aus Evaluationen findet statt, wenn diese indirekt genutzt werden, um das Wissen bzgl. des Evaluationsgegenstandes zu erweitern.
3) Symbolische Nutzung *(Symbolic Use)*: Die symbolische Nutzung von Informationen aus Evaluationen findet statt, wenn Befunde zum Evaluationsgegenstand eingesetzt werden, um bereits getroffene Entscheidungen gegenüber Dritten zu legitimieren.

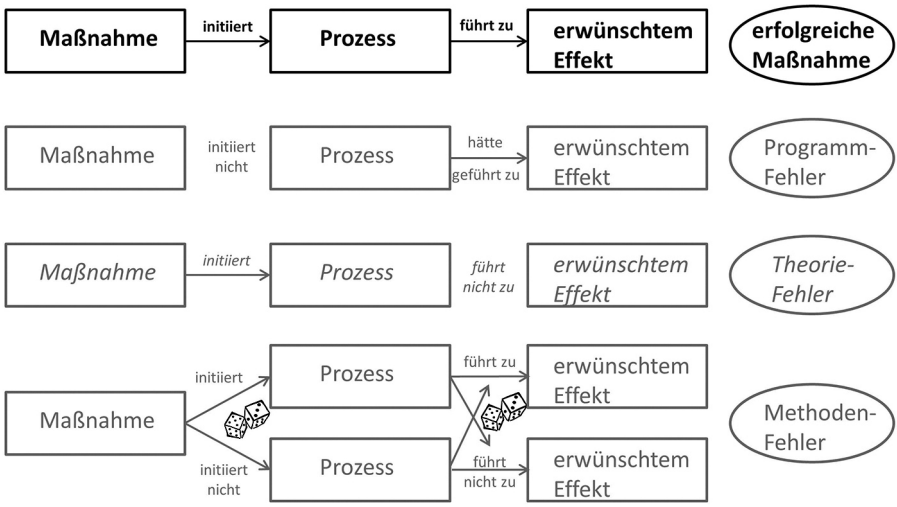
Dabei dient die instrumentelle Nutzung primär dem Zweck der konkreten und zielgerichteten Weiterentwicklung, die konzeptionelle Nutzung wiederum hat die indirekte Veränderung von Einstellungen, Meinungen und Kognitionen der an der Evaluation Beteiligten zum Ziel und die symbolische Nutzung dient der Bekräftigung von Positionen in Diskursen und argumentativen Auseinandersetzungen (vgl. Weiss 1998). Für die Nutzer der Evaluationsbefunde auf Ebene der Einzelschule, müssen die evaluativ gewonnen Informationen daher, wenn es um die konkrete Weiterentwicklung von Schule und Unterricht geht, in erster Linie einen instrumentellen, einen direkten praxisbezogenen Nutzen zur Entscheidungsfindung erbringen (vgl. Scheerens 2007). Einen Nutzen also, der es ermöglicht, inkrementelle Entscheidungen zu treffen, um so z. B. ein bestimmtes Programm zu beenden, es zu modifizieren, Aktivitäten zu verändern oder Weiterbildungsmaßnahmen zu gestalten (vgl. Weiss 1998). Eine solche instrumentelle Nutzung von Evaluationsbefunden kann jedoch nur unter einer der vier folgenden Voraussetzungen stattfinden (vgl. Weiss 1980): 1) wenn die Implikationen, die sich aus den Befunden ergeben, nicht kontrovers sind; 2) wenn die geplanten Veränderungen sich innerhalb des bisher geplanten Programms bewegen und nicht allzu umfassend sind; 3) wenn das

Umfeld des Programms stabil ist, es keine größeren Veränderungen im Bereich von Leitung, Finanzen und Teilnehmern gibt oder wenn 4) ein Programm in der Krise ist und niemand weiß, was getan werden muss. Die direkte instrumentelle Nutzung, die in Veränderungen und Entwicklung mündet, findet somit vor allem dort statt, wo der bestehende Status quo nicht gestört wird (vgl. Nutley et al. 2003).

3. Problemstellung

Wirksamkeits- und Wirkungsstudien untersuchen nun, ob Schulinspektionen durch Schulverantwortliche genutzt werden und welche Wirkungen erreicht werden. Intendiert ist eine instrumentelle Nutzung, die zu einer Verbesserung von Schülerleistungen und diesbezüglich vermittelnden Faktoren der Schul- und Unterrichtsqualität führt. Lassen sich diese Zusammenhänge nicht nachweisen, so kann dies theoretisch auf drei Gründe zurückgeführt werden (vgl. Pietsch et al. 2013; Stame 2010): Erstens kann ein Programm- resp. Implementations-, zweitens ein Theorie- und/oder drittens ein Methodenfehler vorliegen (vgl. Abb. 1).

Abbildung 1: Wirkungskette und mögliche Ursachen für Unwirksamkeit aus programmtheoretischer Perspektive



Während ein Programmfehler vorhanden ist, wenn es mittels einer Intervention nicht gelingt, eine intendierte Wirkung nachweisbar zu erzeugen, liegt ein Theoriefehler vor, sofern ein theoretisch postulierter Wirkungszusammenhang nicht ausreichend valide begründet wurde. Darüber hinaus kann eine nicht nachweisbare

Wirkung letztlich auch darauf zurückzuführen sein, dass Methodenfehler vorliegen. Die Nicht-Wirksamkeit einer Intervention rührt in diesem Falle daher, dass es im Rahmen der empirischen Kausalstudie aufgrund methodologischer Unzulänglichkeiten nicht gelingt, den angenommenen Wirkungszusammenhang empirisch verlässlich zu überprüfen.

Empirische Untersuchungen lassen sich zu allen drei Bereichen finden, wobei derzeit ein grundsätzliches Problem darin besteht, dass Programmtheorien, die eine theoriebasierte Evaluation von Inspektionswirksamkeit ermöglichen sollen, zu unterkomplex sind und die Kompliziertheit der Intervention Schulinspektion nicht berücksichtigen (vgl. de Wolf & Janssens 2007; Pietsch et al. 2013). Dies wiederum hat zur Folge, dass Aussagen zu Programmfehlern – also dazu, ob Schulinspektionen ggf. nicht zufriedenstellend implementiert sind und/oder ihrer Aufgabe nur unzureichend nachkommen – nur schwer möglich sind und die bislang vorliegenden Befunde daher nur wenig verlässliche Aussagen zur Wirksamkeit von Schulinspektionen auf die Entwicklung von Schule und Unterricht sowie auf Schülerleistungen ermöglichen. Entscheidungen zum weiteren Umgang mit Inspektionsverfahren sind daher aktuell nur schwer möglich.

Um die Wirkungen von Inspektionen systematisch zu erforschen, schlagen wir daher ein dreischrittiges Verfahren vor, das der Komplexität des Untersuchungsgegenstandes gerecht zu werden versucht. Hierbei soll in einem ersten Schritt gewährleistet werden, dass das Zielkriterium, der Effekt, um den es geht, sowie der logisch angenommene Wirkungszusammenhang vor Beginn einer Untersuchung klar und eindeutig definiert wird. Im zweiten Schritt geht es darum, die Wirkmechanismen im Rahmen eines logischen Modells zu beschreiben und zu untersuchen, das der Komplexität und der Kompliziertheit des Untersuchungsgegenstandes gerecht wird. In einem dritten Schritt sollen die Ergebnisse dieser Untersuchungen zusammengeführt werden, um hieraus belastbare Konsequenzen auf Systemebene, aber auch auf Ebene der einzelnen Inspektion ziehen zu können.

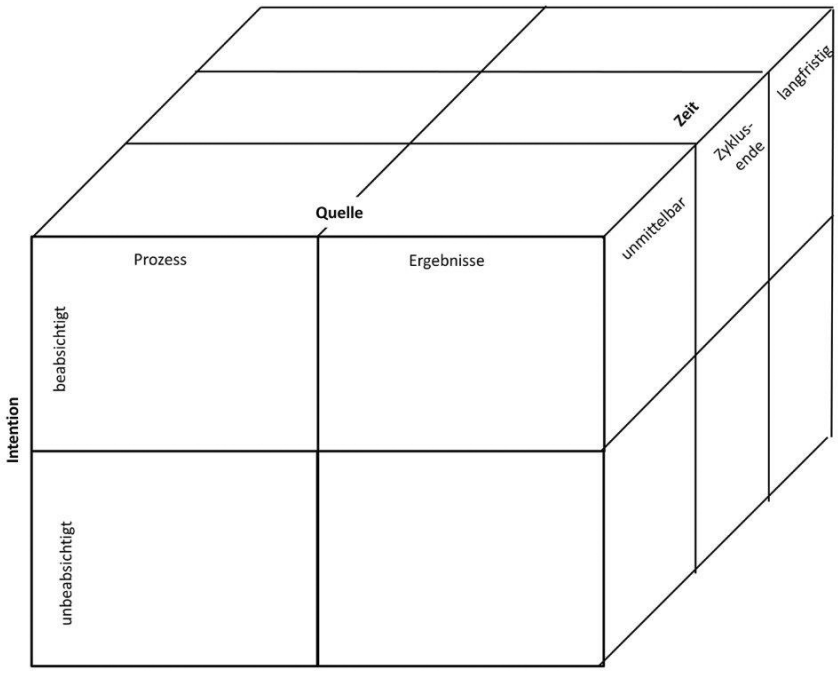
4. Ein konzeptioneller Analyserahmen zur Evaluation von Schulinspektionswirkung

4.1 Grundlage: Definition von Inspektionswirkungen mithilfe eines mehrdimensionalen Modells

Schulinspektionen können, wie beschrieben, sowohl intendierte als auch nicht-intendierte Wirkungen nach sich ziehen, aber auch vollkommen wirkungslos bleiben. Dabei geht es um den Einfluss (im fachwissenschaftlichen Diskurs wird diesbezüglich von *‚Influence‘* gesprochen, vgl. z. B. Herbert 2014), den Inspektio-

nen auf Schule, Unterricht und Schülerleistungen haben. Daher ist es im Rahmen einer theoriegeleiteten Evaluation von Schulinspektionswirkungen in einem ersten Schritt notwendig, zu klären, welche Wirkung in welchem Zeitraum generiert und durch welchen Prozess (Durchführung einer Inspektion oder Rückmeldung von Ergebnissen) erwartet wird. Möchte man also den Einfluss von Schulinspektionen auf Schülerleistungen sowie vermittelnde Faktoren wie Schul- und Unterrichtsentwicklung analysieren, benötigt man entsprechend elaborierte Modelle. Ebensolche haben Kirkhart (2000) sowie Alkin und Taut (2003) vorgeschlagen. Mithilfe dieser dreidimensionalen Modelle lässt sich die Wirkung von Evaluationen – und entsprechend auch von Schulinspektionen – fassen (vgl. Abb. 2).

Abbildung 2: Modell zur Definition von Inspektionswirkungen (nach Kirkhart 2000)



Der Vorteil eines solchen Schemas liegt vor allem darin, dass er über den Begriff der Nutzung hinaus geht und es vor allem auch erlaubt, nicht-intendierte Wirkungen in die Analysen mit einzubeziehen und darüber hinaus Effekte resp. Veränderungen – und somit Wirkungen, nicht die Nutzung von Evaluationen – in den Mittelpunkt stellt. Weiterhin ermöglicht dieses Schema, zu berücksichtigen, dass Schulinspektionen, anders als z. B. Vergleichsarbeiten und Lernstandserhebungen, auch einen Einfluss durch das Wirken von Inspektorinnen und Inspektoren vor Ort

haben und nicht ausschließlich Effekte durch die Rückmeldung von Ergebnissen erzielen. Die Stärken eines solchen Modells stellt Herbert (2014, S. 394) entsprechend heraus:

- „influence provides a definition and a framework that reflects the full impact of evaluation and a cohesive way to organize theoretical and empirical knowledge of the effect evaluation can have on programs;
- by adopting this more comprehensive view, influence allows for the study of implicit mechanisms that affect change, including processes at the individual, interpersonal, and collective levels;
- influence frameworks are oriented around linkages to more developed constructs in other fields of literature such as attitude change, priming, skill acquisition, and persuasion;
- shifting to an influence framework allows for the study of pathways of influence and the study of situations where evaluation failed to affect change; and influence is built around social betterment as the ultimate goal of evaluation, rather than use.“

Untersucht werden kann in einem solchen Design somit: a) ob intendierte Effekte durch die Evaluation erzielt wurden oder nicht *(Intention),* b) ob diese durch die Ergebnisrückmeldung oder den Evaluationsprozess zustande gekommen sind *(Quelle)* und c) ob die Effekte unmittelbar, zum Ende eines Zyklus oder aber erst langfristig nachweisbar sind *(Zeit)*.

4.2 Analyse: Theoriebasierte Evaluation mithilfe komplexer logischer Modelle

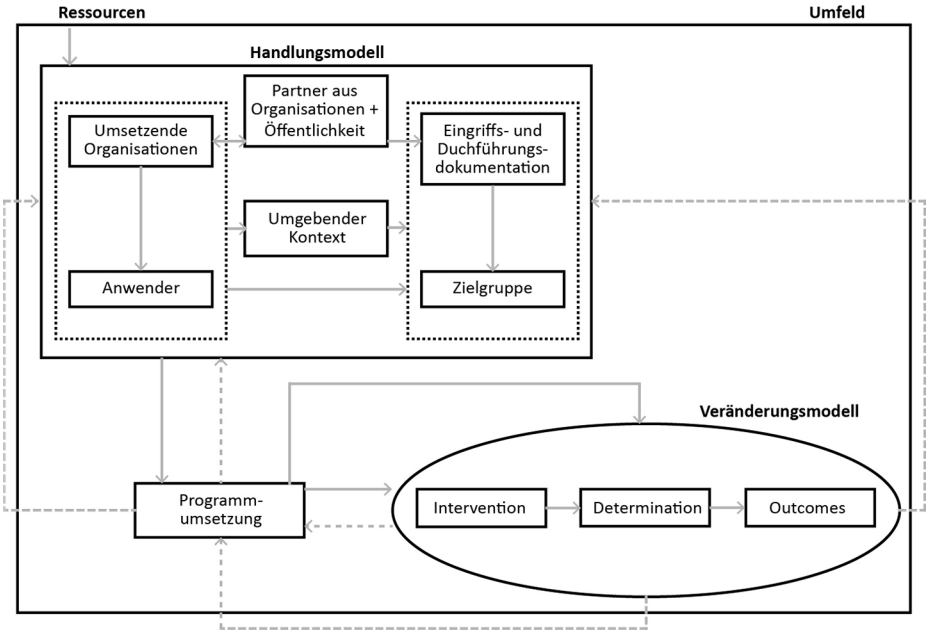
In einem zweiten Schritt geht es im Rahmen einer theoriegeleiteten Evaluation zu Wirkungen von Schulinspektion dann darum zu beschreiben, mithilfe welcher Mechanismen, auf welche Art und Weise die definierten Wirkungen durch ein Schulinspektionsverfahren generiert werden sollen. Von einfach-linearen Modellen wird abgeraten, da diese für die Analyse von Schulinspektionswirkungen zu unterkomplex sind: „The complexity and situatedness of much organizational activity begs for a style of theory that preserves some ambiguity.“ (vgl. Weber 2006, S. 120)

Komplexe programmtheoretische Modelle bestehen in der Regel (vgl. Chen 2006; Coryn et al. 2011) aus mindestens zwei Teilen (vgl. Abb. 3), die eng miteinander verknüpft sind und es u. a. ermöglichen, Kontextbedingungen und/oder rekursive Wirksamkeitsmechanismen zu inspizieren. Chen (2006) unterscheidet diesbezüglich zwischen einem Veränderungs- und einem Handlungsmodell:

> „The action model and change model are closely related to each other and are essential for the success of a program. On the one hand, a change model is needed to justify the selection of an intervention for achieving the goals and/or outcomes and it provides a basis for developing the action model. On the other hand, the action model provides a blueprint to organize program activities and to activate and energize the change model for achieving program goals."

Dabei umfasst das *Veränderungsmodell* drei Komponenten: a) die Intervention, die sich auf das Programm bezieht, mit dessen Hilfe Veränderungen bewirkt werden sollen, b) die Determination, die diejenigen Prozesse und Mechanismen umfasst, die den Input durch die Intervention, mit Blick auf mögliche Effekte, vermitteln, und c) die Outcomes, womit die antizipierten Effekte des Programms gemeint sind. Das Veränderungsmodell unterstellt entsprechend, dass die Implementation einer Intervention diejenigen Determinanten kausal beeinflusst, die ihrerseits zu Veränderungen in den Outputs führt.

Abbildung 3: Modell zur Evaluation von Inspektionswirkungen (nach Chen 2006)



Das *Handlungsmodell* hingegen beschreibt die Voraussetzungen und Annahmen der Intervention selber, stellt also dar, mit welchen Mitteln, unter welchen Voraus-

setzungen und bei welchen Adressaten die Intervention Wirkung erzielen soll. Das Handlungsmodell umfasst dabei sechs Bereiche:

1) *Umsetzende Organisationen:* Die Organisationen, die für die Koordination von Personal und die Ressourcenallokation bzgl. der Intervention verantwortlich sind.
2) *Anwender:* Die Personen, die für die konkrete Umsetzung resp. Implementierung des Programms verantwortlich sind.
3) *Partner aus Organisationen und Öffentlichkeit:* Die Organisationen oder Partner, die zusätzlich zur Primärorganisation notwendig sind, um die Intervention zielgerichtet durchzuführen.
4) *Umgebender Kontext:* Der ökologische Kontext, der direkt mit dem zu implementierenden Programm interagiert, z. B. in Form von Zielvorgaben, Normen, Gesetzen.
5) *Eingriffs- und Durchführungsdokumentation:* Der Umsetzungsplan, der die Intervention inkl. ihrer spezifischen Inhalte und Aktivitäten beschreibt und dabei die Schrittfolge definiert, mit der die Intervention im Feld umgesetzt werden soll.
6) *Zielgruppe:* Die Zielgruppe, für die das intervenierende Programm gedacht ist, insbesondere das Vorhandensein von klar definierten Ansprechpartnern und die Einbindung in das Programm.

4.3 Konsequenzen: Validierung mithilfe von *Interpretation/Use*-Argumentationen

Die Ergebnisse, die mit diesen Modellen generiert werden können, müssen in einem letzten Schritt sinnvoll zusammengeführt und validiert werden, um hieraus Konsequenzen auf Systemebene, aber auch auf Ebene der einzelnen Inspektion zu ziehen. Scriven (2012, S. 2 ff.) betont konsequenterweise, dass Validität im Rahmen von Evaluation das zentrale Kriterium ist:

> „Validity: This is the key criterion – the matter of truth. […] It requires some evidence of ‚real' value, which usually (not quite always) means visible or directly testable evidence somewhere along the line of implications of the evaluation."

Hierbei geht es dann vor allem darum, die Annahmen zur Wirkungsweise von Inspektionen zu prüfen. Was aber meint Validierung hier konkret? In den Standards für Pädagogisches und Psychologisches Testen von 1999 und 2014, die fachlich-methodische und ethische Richtlinien unter anderem für Validität beinhalten, wird Validität als das Ausmaß definiert, in dem empirische Nachweise und Theorien

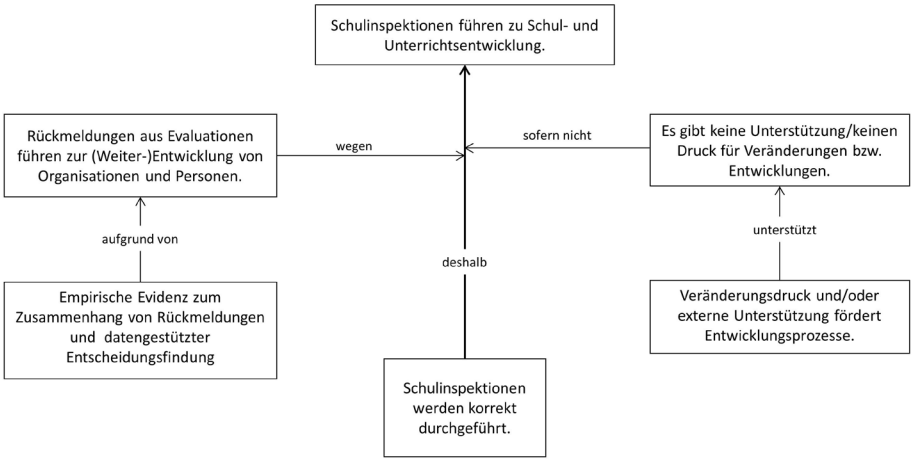
die Interpretation von Testergebnissen für eine bestimmte Nutzung stützen. Die Standards unterscheiden fünf Kategorien der Evidenz (Testinhalt, Antwortprozesse, Interne Struktur, Zusammenhänge mit anderen Variablen und Testkonsequenzen), die bei der Validierung evaluiert werden können. Jedoch wird Validität als ein einheitliches Konzept definiert (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; S. 14):

> „It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use."

Zu validieren bedeutete in diesem Sinne, Argumente für und gegen die geplante Interpretation von Testergebnissen oder, im Falle von empirischen Untersuchungen zur Wirksamkeit von Schulinspektionen, von Studienergebnissen, zu konstruieren und zu evaluieren. Damit spiegeln die Standards den professionellen Konsens darüber wider, dass Validieren das Sammeln von Evidenz für eine wissenschaftlich solide Basis der Testwertinterpretationen und Schlussfolgerungen beinhaltet. In diesem Sinne ist Validität graduell. Für unterschiedliche Tests werden unterschiedliche Arten der Evidenz benötigt, wobei Validität ein einheitliches Konstrukt bleibt. Dabei ist Validität keine Eigenschaft des Tests oder der Untersuchung, sondern bezieht sich auf deren Interpretationen (vgl. Kane 2006; Shaw & Crisp 2012; Messick 1989; American Educational Research Association, American Psychological Association, National Council on Measurement in Education 1999 und 2014).

Neuere Validierungsansätze, die Validitätsnachweise in Form von Validitätsargumenten organisieren, wie der *Argument Based Approach* von Kane (2013), das *Evidence-Centered Design* von Mislevy (2003) und das *Assessment Use Argument* von Bachman (2005), basieren auf Toulmins Beschreibung von „Informellen Argumenten" (1958, 2003). Die informellen Argumente werden in Form einer Argumentationskette aufgebaut, die für eine bestimmte Schlussfolgerung plädiert. In dieser Argumentationskette werden die Plausibilität von Behauptungen und Beobachtungen begründet und Zusammenhänge zwischen diesen bewiesen. In seinem Schema schließt der Argumentierende aufgrund von Daten und mithilfe einer Schlussregel auf die Konklusion. Dabei liefert eine Stützung Beweise für die Schlussfolgerung, Qualifikatoren bringen die Stärke der Behauptung zum Ausdruck und eine Ausnahme beschränkt oder verstärkt die Behauptung.

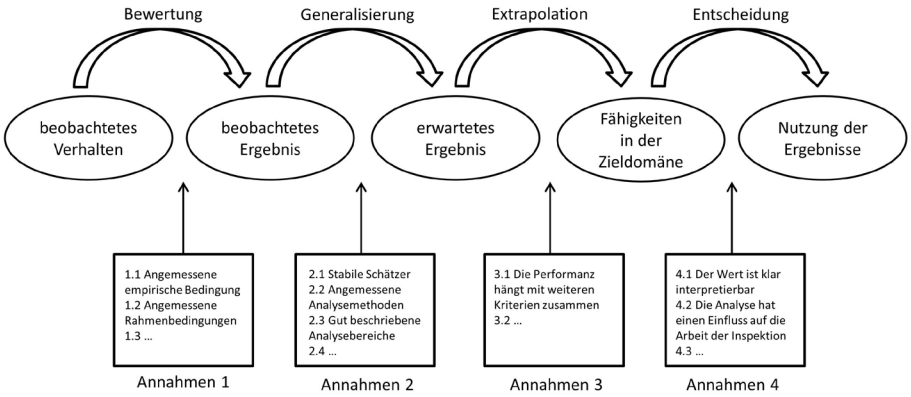
Abbildung 4: Modell zur argumentativen Validierung von Schulinspektionswirkungen (nach Toulmin 2003)



Ein Beispiel für so eine Behauptung wäre, dass Schulinspektion zu einer Verbesserung der Schul- und Unterrichtsqualität führt (siehe Abb. 4). Das Argument hierfür lautet, dass sie korrekt durchgeführt werden. Die Schlussregel ist eine Generalisierung, die genutzt wird, um die spezifischen Daten mit der spezifischen Behauptung zu verbinden. Im Falle des Beispiels in Abbildung 4 lautet die Schlussregel, dass Schulinspektionen, die korrekt durchgeführt werden, in der Regel zu einer Verbesserung der Schul- und Unterrichtsqualität an evaluierten Schulen führen. Die Stützung liefert Beweise für die Schlussfolgerung in Form von Theorien, Untersuchungen, Erfahrung etc.; alternative Erklärungen relativieren eine Behauptung und werden durch Ausnahmen beschränkt oder verstärkt.

Kane geht noch weiter als Toulmin (2003) bzw. Mislevy (2003) und argumentiert, dass für die Verbindung der Beobachtung mit der angestrebten Behauptung mehrere Arten von Schlussfolgerungen notwendig sind. In Abbildung 5 wird das sogenannte *Interpretation/Use Argument* (IUA) dargestellt, welches die multiplen Schlussfolgerungen und Ergebnisnutzungen spezifiziert. Jede dieser Schlussfolgerungen basiert auf Annahmen, die Stützung benötigen. Die erste Schlussfolgerung „Bewertung“ geschieht bei der Ergebnisermittlung. Das beobachtete Verhalten bei oder infolge der Durchführung einer Inspektion wird in ein beobachtetes Ergebnis umgesetzt.

Abbildung 5: Interpretation/Use Argument zur Validierung von Schulinspektionswirkungen (nach Kane 2002)



Dabei wird beispielsweise angenommen, dass die Bewertungsprozeduren angemessen sind, korrekt angewendet wurden und frei von offenkundigen Fehlern sind. Die zweite Schlussfolgerung bezieht sich auf die Generalisierung der Ergebnisse aus einer ausgewählten Messung/Studie (mit bestimmten Kriterien, an einem bestimmten Tag, zu einer bestimmten Zeit) auf andere, vergleichbare Situationen. Dabei wird unter anderem davon ausgegangen, dass die Stichprobe der Umstände, unter denen gemessen wird, repräsentativ ist. Oft werden die Interpretationen auf weitere Bereiche wie z. B. die Realsituation erweitert. Die hierauf gründenden Schlussfolgerungen ermöglichen eine deskriptive Interpretation der Ergebnisse. Die Einbeziehung von Verwendungen und Entscheidungen führt schließlich zu einer entscheidungsbasierten Interpretation (Kane 2002). Bei der Schlussfolgerung „Entscheidung“ werden die Ergebnisse für das Treffen von Konsequenzen verwendet (Kane 2013).

5. Fazit

Vor rund zehn Jahren wurden die ersten Schulinspektorate in Deutschland eingeführt. Die Frage, welche Wirkungen Schulinspektionen nach sich ziehen, ist bislang jedoch nur in Ansätzen und darüber hinaus sehr unsystematisch erforscht. Befunde zur Wirksamkeit, also bezogen darauf, ob Schulinspektionen intendierte Wirkungen erzielen, liegen noch seltener vor. Problematisch ist darüber hinaus, dass die Befunde häufig widersprüchlich sind und keine Generalisierung ermöglichen. Evidenzbasierte Entscheidungen zum weiteren Umgang mit Inspektionen, zu

Veränderungs- und Verbesserungsmöglichkeiten, aber auch zur Einbettung in die länderspezifischen Bildungssysteme sind daher kaum möglich. Aktuelle Veränderungen basieren entsprechend eher auf anekdotischer Evidenz und auf Alltagserfahrungen, denn auf empirischer Evidenz.

Um die Wirksamkeit von Inspektionen jedoch nachhaltig und erfolgreich zu optimieren, sind Entscheidungsträger auf belastbare Befunde aus empirischen Untersuchungen und Evaluationen zu Wirkungen von Inspektion auf erwartete Zielkriterien wie Schul- und Unterrichtsqualität, aber auch Schülerleistungen angewiesen. Wir denken, dass der von uns vorgestellte konzeptionelle Rahmen zur Evaluation von Schulinspektionswirkungen dazu beitragen kann, derartige Informationen zu generieren, sofern er systematisch, im Sinne eines Forschungsprogramms angewendet wird.

Besonders wichtig erscheint es uns dabei, an die aktuelle Validitätsdebatte in der empirischen Bildungsforschung anzuknüpfen und das *Interpretation/Use Argument* von Kane im Rahmen einer theoriegeleiteten Evaluation von Schulinspektionswirkungen anzuwenden. Dieses Modell kann unserer Ansicht nach zur Evaluation der Wirksamkeit von Schulinspektionen genutzt werden und Programm-, Theorie- und Methodenfehler aufdecken bzw. vorbeugen. Dabei wird die Argumentationskette zweimal durchlaufen. In der ersten Phase wird die Maßnahme, in diesem Fall die Schulinspektion, evaluiert. Bei der Bewertung kann überprüft werden, ob mit den Instrumenten ermittelte Ergebnisse das Verhalten der Schule tatsächlich repräsentieren oder ob beispielsweise die Bewertung der Fragebögen, Interviews, Schulbegehungen etc. subjektiv oder fehlerbehaftet ist. Bei der Generalisierung kann untersucht werden, ob sich die Schulinspektion auf andere Facetten und Kontexte der Durchführung übertragen lässt. Hier stellt sich unter anderem die Frage, ob eine Inspektion mit anderen Inspektoren, zu einem anderen Zeitpunkt, in anderen Unterrichtsstunden oder bei anderen Lehrkräften ein gleichwertiges Ergebnis produziert hätte. Bei der Extrapolation ist es möglich, zu analysieren, ob das Konstrukt einer „guten Schule“ empirisch bestätigt werden kann und ob die Erfüllung dieser Normen tatsächlich eine „gute Schule“ in der realen Welt bedingt. Lassen sich keine Argumente oder sogar Gegenargumente dafür finden, dass das Ergebnis der Schulinspektion die reale Schulwirklichkeit beschreibt, kann dies keinen oder einen unerwünschten Prozess zur Folge haben. Gründe können ein Programm- oder ein Methodenfehler in der Schulinspektion sein. Können jedoch Argumente gefunden werden, dass das Ergebnis der Schulinspektion die reale Schulwirklichkeit beschreibt, so liegt eine solide Basis für die Schulentwicklung vor. Auf dieser Basis können dann Entscheidungen getroffen und ein gewünschter Prozess initiiert werden. Der Prozess führt wiederum zu einem eventuell veränderten beobachtbaren Verhalten der Schule. Dieses Verhalten kann erneut mithilfe von Instrumenten gemessen werden, womit das Modell von Kane ein zweites Mal durchlaufen werden kann. Bei der Bewertung kann

erneut untersucht werden, ob die mit den Instrumenten generierten Ergebnisse das Verhalten der Schule korrekt abbilden. Bei der Generalisierung ist es möglich, festzustellen, ob die Befunde aus verschiedenen Untersuchungen einheitlich und übertragbar sind. Bei der Extrapolation kann wiederum evaluiert werden, ob sich die Theorie über den Prozess durch die Testergebnisse bestätigen lässt und ob der Prozess zu erwünschten Effekten in der Wirklichkeit führt. Hierbei können Theorien und Ergebnisse aus dem Handlungsmodell und Veränderungsmodell (Chen 2006) integriert werden. Deuten die Argumente darauf, dass mit dem Instrument nicht die gewünschten Effekte des Prozesses in der Wirklichkeit repräsentiert werden, so kann es sich um einen Theoriefehler des Prozesses oder einen Methodenfehler in der Messung des Prozesses handeln. Bei der Entscheidung geht es wiederum um die Nutzung der Ergebnisse. In diesem Schritt stellt sich u. a. die Frage, ob die Maßnahme als erfolgreich gewertet werden kann, ob sie positive oder negative Konsequenzen für die Schule, die Lehrkräfte oder Schülerinnen und Schüler hat, etc. In diesem Schritt können auch die Aspekte aus dem Modell zur Kategorisierung von Inspektionswirkungen berücksichtigt werden, beispielsweise durch die Kategorisierung von beabsichtigten und unbeabsichtigten Prozessen und Ergebnissen über die Zeit.

Auf diese Weise bietet das Modell von Kane die Möglichkeit, Schulinspektion und ihre Konsequenzen umfassend zu evaluieren. Die Evaluation der einzelnen Schritte ist transparent und offenbart die evidenzbasierten Argumente, jedoch auch Argumente mit schwacher Evidenz, Gegenargumente und Ausnahmen. Auf diesem Wege wird es möglich, wissensbasierte Entscheidungen zur weiteren Ausgestaltung von Schulinspektionsverfahren zu treffen und das Verfahren so zu optimieren, dass es die beabsichtigten Wirkungen auch tatsächlich entfalten kann, also wirksam wird.

Literatur

van Ackeren, I. & Klemm, K. (2009). *Entstehung, Struktur und Steuerung des deutschen Schulsystems*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Alkin, M. C. & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation, 29,* 1–12.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Astbury, B. & Leeuw, F. L. (2010). Unpacking black boxes. Mechanisms and theory building in evaluation. *American Journal of Evaluation, 31* (3), 363–381.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2* (1), 1–34. doi:10.1207/s15434311laq0201_1

Böttcher, W. & Kotthoff, H.-G. (2007). Gelingensbedingungen einer qualitätsoptimierenden Schulinspektion. In W. Böttcher & H.-G. Kotthoff (Hrsg.), *Schulinspektionen: Evaluation, Rechenschaftslegung und Qualitätsentwicklung* (S. 223–230). Münster: Waxmann.

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.

Brunsden, V., Davies, M. & Shevlin, M. (2006). Anxiety and stress in educational professionals in relation to OFSTED. *Education Today, 56* (1), 24–31.

Chen, H. T. (2006). A theory-driven evaluation perspective on mixed methods research. *Research in the Schools, 13* (1), 75–83.

Coryn, C. L. S., Noakes, L. A., Westine, C. D. & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32* (2), 199–226.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik Phillips, D. C., Walker, D. F. & Weiner, S. S. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Dedering, K. (2012). Schulinspektion als wirksamer Weg der Systemsteuerung? *Zeitschrift für Pädagogik, 58* (1), 69–88.

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

Döbert, H., Rürup, M. & Dedering, K. (2008). Externe Evaluation von Schulen in Deutschland – die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede. In H. Döbert & K. Dedering (Hrsg.), *Externe Evaluation von Schulen. Historische, rechtliche und vergleichende Aspekte* (S. 63–152). Münster: Waxmann.

Ehren, M. C. M., Leeuw, F. S. & Scheerens, J. (2005). On the impact of the Dutch educational supervision act: Analysing assumptions concerning the inspection of primary schools. *American Journal of Evaluation, 26* (1), 60–76.

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.

Ehren, M. C. M. & Visscher, A. J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies, 56* (2), 205–227.

Herbert, J. L. (2014). Researching evaluation influence: A review of the literature. *Evaluation Review, 38* (5), 388–419.

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

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21* (1), 31–41.

Kane, M. T. (2006). In Praise of Pluralism. A Comment on Borsboom. *Psychometrika, 71* (3), 441–445. Verfügbar unter: http://doi.org/10.1007/s11336-006-1491-2 [Zugriff am 30.03.2015]

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50* (1), 1–73. Verfügbar unter: http://doi.org/10.1111/jedm.12000 [Zugriff am 30.03.2015]

King, J. A. & Pechman, E. M. (1984). Pinning a wave to the shore: Conceptualizing evaluation use in school systems. *Educational Evaluation and Policy Analysis, 6* (3), 241–451.

Kirkhart, K. E. (2000). Reconceptualising evaluation use: An integrated theory of influence. *New Directions for Evaluation, 88,* 5–23.

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.

Köller, O. (2009). Evaluation pädagogisch-psychologischer Maßnahmen. In E. Wild & J. Möller (Hrsg.), Pädagogische Psychologie (S. 331–351). Heidelberg: Springer.

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

Maritzen, N. (2009). Schulinspektion und Schulaufsicht. In H. Buchen & H.-G. Rolff (Hrsg.), *Professionswissen Schulleitung* (S. 1368–1392). Weinheim: Beltz.

Messick, S. (1998). Test Validity: A Matter of Consequence. *Social Indicators Research, 45* (1/3), 35–44. Verfügabr unter: http://doi.org/10.1023/A:1006964925094 [Zugriff am 30.03.2015]

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2,* 237–258.

Müller, S. (2010). Erste Effekte von Schulinspektion – eine Zwischenbilanz. In N. Berkemeyer, W. Bos, W., H. G. Holtappels, N. McElvany & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung Band 16. Daten, Beispiele und Perspektiven* (S. 289–308). Weinheim: Juventa.

Nutley, S., Walter, I. & Davies, H. T. O. (2003). From Knowing to doing: A framework for understanding the evidence-into-practice agenda. *Evaluation, 9* (2), 125–148.

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., Janke, N. & Mohr, I. (2014). Führt Schulinspektion zu besseren Schülerleistungen? Difference-in-Differences-Studien zu Effekten der Schulinspektion

Hamburg auf Lernzuwächse und Leistungstrends. *Zeitschrift für Pädagogik, 60* (3), 446–470.

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.

Ramaprasad, A. (1983). On the Definition of Feedback. *Behavioral Science, 28* (1), 4–13.

Scheerens, J. (2007). The case of evaluation and accountability provisions in education as an area for the development of policy malleable system indicators. In H.-H. Krüger, T. Rauschenbach, U. Sander (Hrsg.), *Bildungs- und Sozialberichterstattung* (Zeitschrift für Erziehungswissenschaft: Sonderheft 6, S. 207–224). Wiesbaden: VS Verlag für Sozialwissenschaften.

Scriven, M. (2012). *Evaluating Evaluations: A Meta-Evaluation Checklist.* Verfügbar unter: http://michaelscriven.info/images/EVALUATING_EVALUATIONS_8.1.11.pdf [Zugriff am 30.03.2015]

Shaw, S. & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication* (Special Issue 3). Verfügbar unhter: http://www.cambridgeassessment.org.uk/images/110003-research-matters-special-issue-3-an-approach-to-validation.pdf [Zugriff am 30.03.2015]

Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation, 16* (4), 371–387.

Stamm, M. (2003). Evaluation im Spiegel ihrer Nutzung: Grande idée oder grande illusion des 21. Jahrhunderts? *Zeitschrift für Evaluation, 2* (2), 183–200.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The uses of argument* (aktualisierte Aufl.). Cambridge: Cambridge University Press.

Visscher, A. J. & Coe, R. (2003). School Performance Feedback Systems: Conceptualisation, Analysis and Reflection. *School Effectiveness and School Improvement, 14* (3), 321–349.

Weber, K. (2006). From nuts and bolts to toolkits: theorizing with mechanisms. *Journal of Management Inquiry, 15* (2), 119–123.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge: Creation, Diffusion, Utilization, 1* (3), 381–404.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19* (1), 21–33.

Melanie C.M. Ehren
Editor

Methods and Modalities of Effective School Inspections

Springer

Chapter 3
Validation of Inspection Frameworks and Methods

Melanie C.M. Ehren and Marcus Pietsch

Abstract This chapter explores the issues of reliability and validity of inspection frameworks and methods, and challenges and tensions in inspection frameworks and methods. Validity is an important aspect of thinking about effective inspection system as invalid inspection systems may lead to flawed judgments which will misguide administrative interventions and policy decisions and which are likely to have a negative impact on schools and teachers. We will introduce Kane's (J Educ Meas 50(1):1–73, 2013) notion of argument-based approaches to evaluate the validity of inspection frameworks and provide two examples of how such an argument can be constructed. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. American Educational Research Association, Washington, DC, 1999; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. American Educational Research Association, Washington, DC, 2014) will then be used to describe five types of evidence to evaluate the validity of these arguments. For each of these sources of evidence we present examples of available studies.

3.1 Introduction

Chapter 2 suggested a range of standards to include in inspection frameworks, particularly of those Inspectorates of Education aiming to improve school quality. The example from different countries highlighted how similar standards can be

M.C.M. Ehren (✉)
Reader in Educational Accountability and Improvement,
UCL Institute of Education, University College London,
20 Bedford Way, London WC1H 0AL, UK
e-mail: m.ehren@ucl.ac.uk

M. Pietsch
Institut für Bildungswissenschaft, Leuphana Universität Lüneburg,
Scharnhorststr.1, 21335 Lüneburg, Germany
e-mail: pietsch@leuphana.de

M.C.M. Ehren (ed.), *Methods and Modalities of Effective School Inspections*,
Accountability and Educational Improvement,
DOI 10.1007/978-3-319-31003-9_3

measured in different ways but will generally include standardized protocols and assessment guidelines, which often result in reports and/or graded judgements. The approach to making such judgements was originally however largely connoisseurial and it was not until the 1980s that Inspectorates of Education (e.g. in the Netherlands and England) started using a standard grading scale and required their inspectors to define criteria by which the quality of schools could be judged. It was around this time that performance tables of aggregated results from student examinations were also first established, enabling school inspectors to substantiate their analysis of school documentation, interviews with school staff and stakeholders, and observation of lessons with more 'objective' school data to inform their assessment. Structured protocols and decision rules were put in place to guide lesson observations, interviews and document analyses and to guide the aggregation of scores of individual lesson observations into a score for indicators and standards and often also an overall assessment of the school as providing insufficient, sufficient or good educational quality. Inspection standards thus operationalise 'education quality' as attributes that are measurable in classrooms and schools.

The development and use of such measures automatically raises the question of the validity of the use of these measures. How do we know that school inspections, and the measures used by school inspectors in such inspections provide an accurate assessment of the quality of a school? Do the standards, indicators and measures in the (inspection) framework measure what they intend to measure (e.g. quality of schools) within a specific context of high stakes accountability and/or improvement? Validity is, for many authors, central to having a credible and fair inspection system and is therefore seen as the most important technical criterion for defending the quality of such inspection systems (see for example Marion and Gong 2003; Scheerens et al. 2005; Lane et al. 1998; Haertel 2002; Lane and Stone 2002). Invalid inspection systems may lead to flawed judgments which will misguide administrative interventions and policy decisions and which are likely to have a negative impact on schools and teachers.

Validity of inspection results is not self-evident as many inspection systems suffer from rapid development and many press releases and articles indicate that inspection is often not underpinned by agreed standards and facts while inspection evidence and interpretation of evidence are hotly disputed. Morrison (1996) for example describes how a sample of a few hours of high stress teaching on an atypical and disruptive day of teaching (as is usually the case when schools are inspected) and a summary judgement of 'good', 'very good' etc. to summarize school quality stretches validity beyond credibility.

These claims of invalid inspection results are however often circumstantial and a more nuanced study and overview of the validity of the interpretation of inspection results is needed. This chapter aims to fill this gap by discussing important notions around validity in the context of school inspections, and providing an overview of the research methods available to establish the validity of inspection results, as well as providing examples of validity studies on school inspections.

3.2 Validity in the Context of School Inspections

Validity was originally developed in the context of test construction to measure some trait of students (e.g. math skills) but can equally be applied to evaluate the inspection measures used to assess traits (e.g. quality/output) of schools. The Standards for Educational and Psychological Testing (AERA 1999, 2014) define validity as 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (p. 9), which not only refers to test characteristics but also to the appropriateness of test use and to the accuracy of the inferences made on the basis of test scores (Sireci and Sucin 2013). Validity covers the essential question of 'how well a test does the job it is employed to do' (p. 1; see also Kane 2006, p. 21), 'is the test any good as a measure of the characteristic that it purports to assess (scientific questions)?' and 'should the test be used for its present purpose (ethical questions)?'(Newton and Shaw (2014, p. 131).[1] In more recent literature (Newton and Shaw 2014) validity also includes the reliability of tests where the consistency of outcomes is treated as a technical facet of a broader concept of measurement quality. Kane (2013) explains how reliability is a special case of validity as evidence for the generalizability (or reliability) of scores over conditions of observation is generally necessary in making a case for validity. Validity is treated here as a unitary concept that cannot be distinguished into different kinds of validity (e.g. content validity or predictive validity), but should be treated as a 'whole'.

These notions of validity are different from how validity is defined in the context of research where different expressions and categorizations such as internal versus external validity are used to analyse the degree to which a research study measures what it intends to measure.

An important notion here is that one does not validate a test or measure, but an interpretation of data arising from the test or measure (see Kane 2006, p. 2). As Newton and Shaw (2013, p. 304) state: it is never the test that is to be judged valid or invalid, in a general sense, but the interpretation of test scores as measures of a specific attribute under specified conditions. Different conclusions concerning validity of a test might follow for different groups of individuals, or for different situations within which individuals or groups found themselves. When applying these notions of validity to school inspections we come up with the following description:

[1] 'Test' can, according to Newton and Shaw (2013), be interpreted as 'measurement procedure' which can include the elements of the measurement procedure ('the test item is valid'), the measurement procedure ('the test is valid'), the decision procedure ('the use of the test is valid'), and the testing policy ('the system is valid'). Each level requires different kinds of conclusions, derived from different kinds of evidence and analysis to establish whether: the item is fit to be used in the measurement procedure, the measurement procedure is fit to be used in the decision procedure, the decision procedure is fit to be used in the testing policy, and the testing policy is fit to be used in the construction of a good education system. This paper focuses on the measurement and decision procedure. An inspection measurement procedure is also a multi-level construct as inspection frameworks are often hierarchical in nature and assessments on lower levels in the framework are aggregated to assess a construct on a higher level in the framework.

> Validity in the context of school inspections entails the extent to which inspection frameworks, guidelines, protocols used in the assessment of schools are a good measure of the characteristic that it purports to assess (e.g. school quality), and whether these frameworks, guidelines and protocols should be used for its present purpose (e.g. control, improvement, liaison).

Validity involves the inferences school inspectors are drawing from the information on educational processes and/or output of schools (captured in inspection indicators and collected through standardized observation protocols, interview guidelines etc.), whether these inferences are consistent with the actual functioning and results of the school and whether the use of inspection results (by schools and stakeholders) fits the intended purpose of the measures. Relevant questions are: What indicators are included in the inspection system and how is each indicator used? Are measures of the indicators well operationalized? Is the methodology for measuring the indicators appropriate? How well do the definitions of these indicators capture what is intended? Were the 'right schools' identified for rewards, sanctions, and interventions? And 'do the decision rules distinguish between failing and good schools?'

3.3 Interpretation/Use Argument

Making a case for the validity of any measure should start, according to Kane (2013) and Sireci and Sukin (2013), with making the reasoning inherent in proposed interpretations and uses of tests explicit so that it can be better understood and evaluated. Such a reasoning or rationale is called an 'interpretation/use argument' and involves an outline of the proposed interpretations and uses of the scores generated by the testing programme as applied to some population over the range of contexts in which it is to be used.

Toulmin's model of argumentation (1958) provides the building blocks to construct such an argument. He explains how a good argument to succeed, needs to provide good justification for a claim. This, he believed, will ensure it stands up to criticism and earns a favourable verdict. In The Uses of Argument (1958), Toulmin proposed a layout containing six interrelated components for analysing arguments:

- Claim (Conclusion): a conclusion whose merit must be established. In argumentative essays, it may be called the thesis [10]. For example, if a person tries to convince a listener that he is a British citizen, the claim would be "I am a British citizen" (1).
- Ground (Fact, Evidence, Data): a fact one appeals to as a foundation for the claim. For example, the person introduced in 1 can support his claim with the supporting data "I was born in Bermuda" (2).
- Warrant: a statement authorizing movement from the ground to the claim. In order to move from the ground established in 2, "I was born in Bermuda," to the claim in 1, "I am a British citizen," the person must supply a warrant to bridge the gap between 1 and 2 with the statement "A man born in Bermuda will legally be a British citizen" (3).

- Backing: credentials designed to certify the statement expressed in the warrant; backing must be introduced when the warrant itself is not convincing enough to the readers or the listeners. For example, if the listener does not deem the warrant in 3 as credible, the speaker will supply the legal provisions: "I trained as a barrister in London, specialising in citizenship, so I know that a man born in Bermuda will legally be a British citizen."
- Rebuttal: statements recognizing the restrictions which may legitimately be applied to the claim. The rebuttal is exemplified as follows: "A man born in Bermuda will legally be a British citizen, unless he has betrayed Britain and has become a spy for another country."
- Qualifier: words or phrases expressing the speaker's degree of force or certainty concerning the claim. Such words or phrases include "probably," "possible," "impossible," "certainly," "presumably," "as far as the evidence goes," and "necessarily." The claim "I am definitely a British citizen" has a greater degree of force than the claim "I am a British citizen, presumably."
- The first three elements, "claim," "data," and "warrant," are considered as the essential components of practical arguments, while the second triad, "qualifier," "backing," and "rebuttal," may not be needed in some arguments.

The argument in the context of school inspections would include a set of interrelated components to argue how school inspections lead to improvement of schools. Such an argument should include an outline of the construct being measured (e.g. what is it that Inspectorates of Education are measuring, how are they defining 'educational quality' of the education system/schools and/or teachers?), how the construct is operationalized in an inspection framework used to assess schools, and whether the measurement and how it is communicated to schools and the wider public is fit for purpose in reaching the intended aims of inspections (control, support/improvement, liaison).

Following Toulmin's framework, the claim is the overall statement, such as 'annual school inspections lead to improvement of schools'. The ground is the evidence (data and facts) to support the claim; e.g. all schools are inspected annually'. The warrant is the principle, provision or chain of reasoning that connects the grounds to the claim, such annual inspection assessments generate feedback which schools use to implement improvements. The backing (or *support*) for an argument gives additional support to the warrant and could include statements such as 75 % of the inspected schools indicate they accept inspection feedback. The qualifier (or *modal qualifier*) indicates the strength of the leap from the data to the warrant and may limit how universally the claim applies; a qualifier for example specifies that inspections only lead to improvement in schools that have been assessed as failing.

Despite the careful construction of the argument, there may still be counterarguments that can be used and these would be included in the 'rebuttal'. A rebuttal may include a statement saying that schools that were inspected last year show no improvement in test scores in the year after the visit. Figure 3.1 shows an example

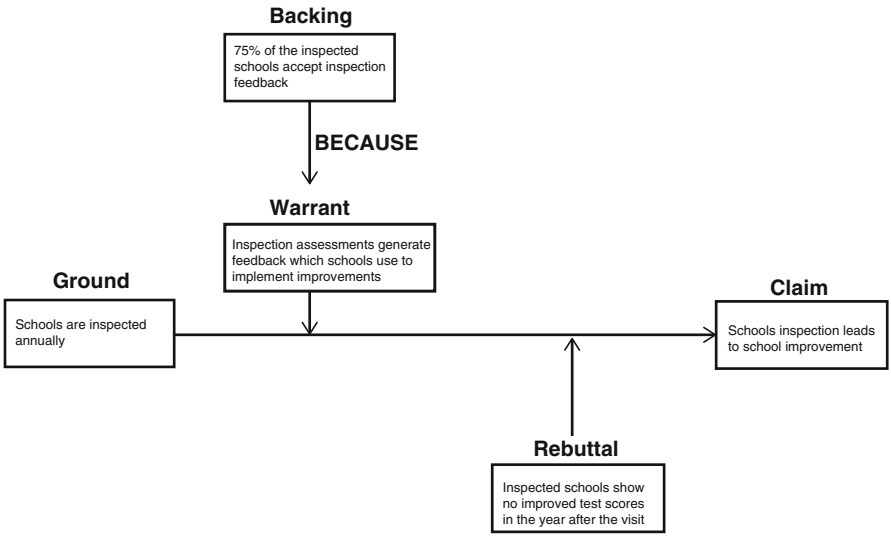


Fig. 3.1 Example structure of an interpretative argument about the impact of school inspections

of an argument, consisting of one claim. An interpretation/use argument will generally include a set of interconnected claims.

Other approaches to inform an 'interpretation/use argument' have been developed by Ehren et al. (2013) in their comparative EU-study on the impact of school inspections in six countries. They used a policy scientific approach (see Leeuw 2003; Ehren et al. 2005), including interviews with inspection officials in six European countries and additional document analyses, to reconstruct Inspectorates' theory of action on how such inspections of educational quality are typically expected to lead to improvement of schools and student achievement, such as when schools use inspection feedback to address weaknesses, when they use the standards to inform school organisation and teaching practices ('setting expectations'), or when stakeholders use inspection reports to exercise voice, choice or exit. This theory of action was summarized by these authors in Fig. 3.2.

Validation should then be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. The first step in such an evaluation includes, according to Kane (2013) a conceptual analysis of the clarity, coherence, and plausibility of the argument, while in a second step evidence is collected to test the claims in the interpretation/use argument and to establish the consistency of the evidence with the definition of the construct. The following section focuses on the second step and follows the Standards for Educational and Psychological testing to describe five potential types of evidence that can be used to test a validity argument. We also provide examples for the validation of school inspection measures.

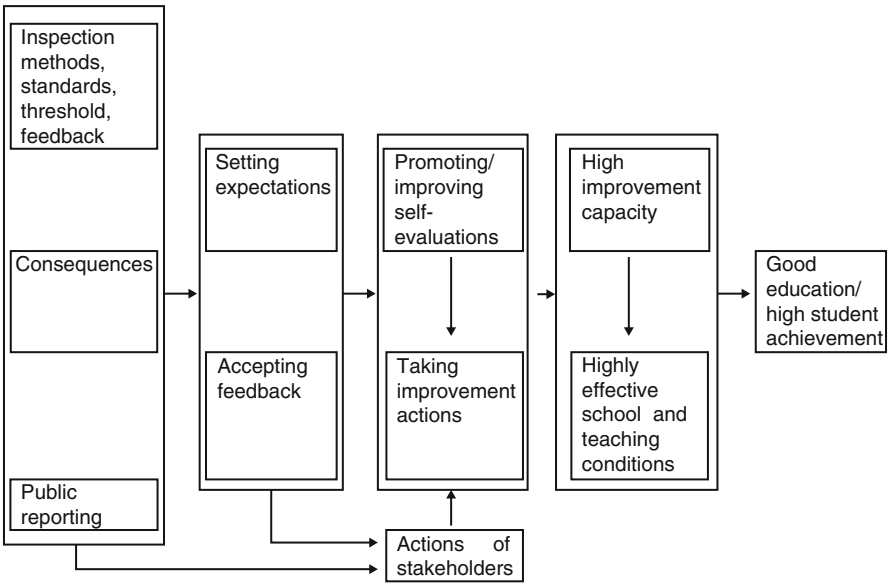


Fig. 3.2 Example of a theory of action about the impact of school inspections

3.4 Sources of Validity Evidence

The *Standards for Educational and Psychological Testing* (AERA et al. 1999, 2014) describe how a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. The Standards stipulate five sources of validity evidence "that might be used in evaluating a proposed interpretation of test scores for particular purposes" (AERA et al. 1999, p. 11): evidence based on the content of tests, on relations to other variables, on internal structure, on response processes and on consequences of testing. The interpretation/use argument can be used to generate hypotheses that can be tested with those five sources of evidence. The validity of a proposed interpretation or use depends on how well the evidence supports the claims being made.

3.4.1 Validity Evidence Based on Test Content

Validity evidence based on test content starts with a description of what is being measured: the construct (e.g. ability in mathematics) or the content domain (e.g. addition and subtraction). The central question is: 'to what extent does the content of the test, including items, subscales, formats and scoring rubrics, adequately and comprehensively represent the content domain? Potential sources of invalidity are

therefore under-representation of the construct (e.g. when performance elements are given less emphasis in the test than the intended inferences warrant) and construct-irrelevant variance (when one or more irrelevant constructs are being assessed in addition to the intended construct; e.g. vocabulary when using word problems to test mathematical problem-solving).

Relevant hypotheses and questions in the context of school inspection are 'Do school inspection frameworks include the relevant and representative aspects of school quality?' 'Do assessment instruments reflect the conceptual framework of reference?' 'Is the sample of a school's reality as assessed in school inspections relevant and representative for school quality and instructional quality?' 'On what rational basis can standards of good school and instructional quality be established?' (Pant 2010; Gaertner and Pant 2011).

Content validity evidence is, according to Sireci and Sucin (2013) usually gathered by having experts review test items and make judgments regarding the relevance of each item to the construct measured and the degree to which the items adequately and fully represent the construct. Evaluators typically focus on whether any important areas are omitted from the specifications or whether superfluous areas are included. Such processes can be structured by using pre-set methods of standard setting in which panels of experts follow a pre-set method (e.g. Angoff, Nedelsky, Ebel; see Hambleton and Pitoniak 2006).

In Germany Pietsch (2010) used such a procedure to set standards for judging the quality of classroom teaching within a school inspection. First, after collecting ratings of classroom observations, the items of an observation record were scaled by using Item Response Theory. Second, a bookmark equivalent, item-centred expert study was conducted to define cut-scores on the latent continuum "Teaching Quality". Thus an expert panel was employed to evaluate each item of the classroom observation sheet and associate the item with a criterion behaviour (precisely: the quality of teaching regarding the optimal standard as defined within the Hamburgian framework for inspection). Third, these cut-scores were adjusted with respect to the distribution of the items, for maximising the classification accuracy of teaching quality within the strata, which are describing four different levels of teaching quality: IV Optimal Standard, III Average Expected Standard, II Minimal Standard, I Below Minimal Standard. Afterwards a person-centred contrasting group study, in which the quality of teaching at the level of a school were categorised by school inspectors who previously inspected the sample schools, according to the description of the strata, were conducted for validating the results. Thus the panellists finally decided the pass score by detecting it on the score scale by classifying inspected schools in two groups: those who reached a specific standard and those who did not reach it.

Another strategy to collect and analyse validity evidence based on content is the comparison of inspection frameworks with scientific research. An example of such a strategy was introduced in the Chap. 2 and implemented more rigorously is a study by Scheerens et al. (2005). These authors validated the framework of the Dutch Inspectorate of Education by making a connection between the indicators in the framework and the research literature on school and instructional effectiveness.

Their study evaluated the scientific basis of the framework by answering questions such as 'Are the process indicators on teaching and learning in the Dutch Inspection Frameworks supported by the knowledge base on school and instructional effectiveness?', 'How feasible is the idea of proportional supervision, given the possibilities and state of the art of school self-evaluation in the Netherlands?', and 'Do the Inspection Frameworks manifest defensible choices with respect to outcome and process indicators, and strategic applications, given current perspectives on educational governance and modern interpretations of the core societal functions of education?' They conclude that the process indicators in the Inspection Framework can be seen as effectiveness enhancing conditions that are inspired by empirical school effectiveness research and by an international consensus on good teaching practice. Correspondence is largest for indicators on the classroom level, and less so for indicators on the school level which suggests that particularly classroom level indicators in the Dutch inspection framework contribute to valid assessments of a school's effectiveness.

3.4.2 Validity Evidence Based on Relations to Other Variables

Validity evidence based on relationships with other variables addresses the type and extent of the relationship between test scores and other variables the test is expected to correlate with or predict (Sireci and Sucin 2013; Brown 2010). These 'other variables' may include performance of a test taker on a later date to evaluate the effectiveness and accuracy of test scores in selection, classification and placement decisions. Test scores are for example assumed to predict future performance on the labour market or academic success in college. Confirmation of such theoretical relationships can reinforce the interpretations and uses that are intended to result from a score on a given instrument, according to Sireci and Sucin (2013).

Relevant other variables to validate inspection measures are for example student outcomes in domains not included in inspection assessments or aspects of school quality not specifically measured. Inspectorates of Education often assume that their assessments of schools, using for example student achievement scores in mathematics and literacy, also reflect the quality of a school in other areas and in the future. Such claims are specifically made by Inspectorates of Education using early warning analyses to target potentially failing schools for inspection visits for 1 or more years (e.g. the Netherlands and England). Schools with high student achievement results in mathematics and literacy are in principal not scheduled for inspection visits, assuming their teaching and school organisational quality and student achievement in other domains is also high and will remain high until the next round of inspections.

Such claims can, according to Sireci and Sucin (2013) be evaluated by using a number of statistical methods to analyse the relationship between measures and other variables, such as correlation coefficients and multitrait–multimethod correlations to assess both convergent and discriminant relationships, where one would

expect high correlations for the same construct on different measures (convergent relationships) and noticeably lower correlations among measures of different constructs (discriminant relationships). Multiple regression allows for gauging the predictive accuracy of test scores as well as the relative predictive utility of test scores when used in conjunction with other predictor variables, whereas hierarchical linear modelling can be used to calculate different predictor equations for the multiple groups or, experimental and quasi experimental designs and meta-analyses.

Examples of studies about relations between inspection measures and other variables are provided by the Hamburg School Inspection who highlight a significant correlation of inspection measures for quality of teaching and the results of school leaving examinations (Schwippert 2015).

3.4.3 Validity Evidence Based on Internal Structure

The internal structure of a test refers, according to Sireci and Sucin (2013), to the dimensionality or underlying factor structure of an assessment; an assessment of self-concept may for example hypothesize separate dimensions for academic self-concept and social self-concept. Important issues to consider and clarify in gathering validity evidence based on the internal structure are, according to Sireci and Sucin (2013, p. 72), the type of information gained from collecting this kind of evidence, the scoring model for the assessment, the declaration of dimensionality, and the decision to report subtest scores, composite scores, or both. Validity investigations based on internal structure should also be analyzed across subgroups of examinees to evaluate whether the test as a whole, subtests, or individual items are invariant across relevant subgroups of test takers.

The internal structure of scores on inspection measures refers to how school quality is separated into dimensions, such as teaching quality, quality of school leadership and quality of assessments of students. It reflects the hierarchical nature of the inspection framework into indicators, substandards and standards and how judgements on these levels are aggregated to higher levels to come to an overall judgement of school quality. Inspection frameworks for example include standards, which are detailed in sub standards, which are detailed in indicators. Indicators are scored, using for example a four-point grading scale from outstanding to inadequate. Illustrations or grade descriptors of when an indicator is outstanding or inadequate are provided to support school inspectors in making adequate judgements.

Validity evidence based on the internal structure asks whether the measurement model conforms with the structure of the construct, which is, in the context of school inspections, the multilevel nature of school quality where teachers within classes/grades/years are nested within schools and quality on these levels includes different elements and conceptualizations. The structure of the measurement needs to reflect this multileveled nature. Relevant questions to ask are: are the computa-

tional and aggregational steps made during an inspection evaluation sound? And should judgements about school quality be reported on the level of indicators, standards or as an overall summary grade for the entire school, or for different grade levels or subject departments? The brief analysis of inspection frameworks in six European countries in Chap. 2 explained how many frameworks ignore the interrelatedness of conditions of school quality.

The internal structure of inspection assessments can be further validated by running statistical analysis on actual inspection assessments and on how the functioning of teachers/grades and aspects of the school organisation are scored and aggregated into an overall assessment of school quality. Statistical analysis, such as exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and IRT residual analysis can be used to analyse variation in the data (e.g. across subgroups of schools or for specific measures and indicator sets) and determine how many dimensions (or factors) are needed to characterize the variation in the data and to evaluate whether these observed dimensions are congruent with the theoretical, hypothesized dimensions. Using factor analysis, the empirical structure that underlies the inspectors' scoring can for example be compared with the intended structure of the framework for assessing school performance, looking at the extent to which the dimensions (factors) from a factor analysis are similar to the theoretical standards in the inspection framework. Relevant studies would, according to Pant (2010) also ask whether the steps of information processing in inspections– such as the combination of individual evaluations into an overall evaluation – are always admissible, or indeed whether these steps can be related to the general construct of school and instructional quality.

An example of such a study is Matthews et al. (1998) who investigated the suitability of the set of indicators for use as a measuring instrument. They calculated the discrimination of each indicator – the correlation (ris) between the score on the indicator and the sum score over the set of indicators –to show the strength of its relationship with all the other indicators. Another example is a study by Gaertner and Pant (2011) who examined, based upon ratings of schools inspectors, the factorial structure of the Brandenburg framework for inspection, finding extensive structural differences between the theoretical hierarchical structure of inspection assessments and the actual inspection judgments (Pietsch 2010; Sommer 2011). Another approach to analysing the internal structure of inspection measures was applied by Schwippert (2015) who used a hierarchically multilevel design to calculate the interrelation of school leadership and job satisfaction of teachers as measured within an inspection. He found that about 20–25 % of the variance of both measures could be explained by the levels of school and school type, 75 % of the variation lies within a single school and that two examined measures are not correlated within all schools. Hence he questions current common practice in school inspections of aggregating individual teacher data to inform school level assessments without taking into account the variability of these scores.

3.4.4 Validity Evidence Based on Response Processes

Gathering validity evidence based on response processes involves demonstrating that the type of performances or responses of the individuals completing the test fit the intended construct being measured or evaluated. It involves, according to Sireci and Sucin (2013), a demonstration of whether examinees are invoking the hypothesized constructs the test is designed to measure in responding to test items. As the *Standards* (AERA et al. 1999, p. 12) describe, "Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees".

Gathering this type of evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items. It is however important in order to establish a fit between the construct (e.g. ability in solving mathematics problems) and the nature of performance or response actually engaged in by examinees (Standards, AERA 1999). The evidence should rule out specific construct irrelevant strategies such as guessing or test wiseness. A number of authors have for example studied teachers' and students' responses to high stakes testing and have found that certain strategies (e.g. drill and practice of high stakes test items) lead to a misfit between the nature of the response and the construct being measured: high test scores do not represent high ability in solving math problems anymore. Sireci and Sucin (2013) also refer to the criteria used by judges when they score performance tasks as an important element of response processes.

Response processes in the context of school inspections particularly refer to how school inspectors interpret and use inspection guidelines, handbooks and frameworks to judge school quality, as well as to responses of schools to these measures (e.g. when filling in inspection surveys or responding to interview questions during visits). Both types of responses (of inspectors and schools) are relevant as they are part of the actual measurement process of school quality, and affect how well the construct (school quality) is measured. Studies on strategic behaviour of schools (see De Wolf et al. 2007) for example show how schools manipulate data to improve the outcome of the inspection, while personal bias of school inspectors may lead to an over emphasis on some of the indicators in the inspection framework.

The handbooks and scoring guidelines developed and used by many Inspectorates of Education aim to support inspectors in their use of inspections frameworks when evaluating schools. Many Inspectorates of Education also provide structured training and professional development to improve school inspectors' scoring of schools. Training often includes case study material related to the specific phases of data collection and analysis, such as videos of lessons to help standardise the evidence inspectors record and the judgements they make in the classroom. Inspection handbooks illustrate how the inspection standards and evaluation criteria should be interpreted and provide examples of specific practices that would fit each of the categories on the grading scale (e.g. outstanding/good/weak/failing). Handbooks often also specify how the overall judgement, as reflected by the grade, must arise from

weighing up strengths and weaknesses and evaluating how well their combination promotes high educational outcomes, in particular high attainment, good progress, and a positive response from pupils (Matthews et al. 1998). Specific guidance and attention is given to evidence of practices around the 'failure boundaries' and these practices are then discussed in consensus/moderation meetings to increase consistency of assessments.

Quality assurance procedures, including complaint procedures are additionally implemented to provide checks on the validity of school inspection assessments. These procedures particularly include protocols for cross-checking the judgement of failing schools who go into special measures and face high stakes consequences. In England, senior HMI inspectors always have to confirm judgements of additional inspectors of failing schools. In the Netherlands, a back office check of judgements of failing schools against collected evidence is required before inspection reports of such schools are finalized.

Most studies of response processes collect evidence through participants' responses to performance tasks using participant debriefing interviews, or investigations of the ways that raters, observers, interviewers, and judges collect and interpret data, evaluating respondents' accuracy when scores on different measures are combined into an aggregate grade. Other studies look at observer agreement on individual classroom observations and around specific failure boundaries, or look at agreement between recorded evidence and final judgements.

Examples of studies about response processes of school inspectors are Matthews et al. (1998) who calculated correlations to evaluate the extent to which pairs of inspectors base their grades on the same recorded evidence, and the extent to which teaching grades awarded by inspectors match their recorded evidence. Matthews et al. (1998) also asked two trained inspectors from England and the Netherlands, to independently observe the same lesson and evaluate and grade the quality of teaching. A total of 173 pairs of observations were received from 100 inspections, representing about 13 % of the inspections conducted during the period of the study. The lessons used for dual observation were chosen by mutual agreement between inspectors within a team. Inspectors were advised to choose subjects or areas which both of the pair felt competent and confident to inspect. In each lesson observed, inspectors had to assess the quality of teaching; the pupils' response, such as pupils' attitudes to learning and their behaviour, the pupils' attainment level in comparison with national standards, and the progress the pupils are making. The research addressed the extent to which pairs of inspectors observing the same lesson agree about the grade awarded to the teacher and the level of agreement between two inspectors at particular grade boundaries. The research found that in 33 % of cases, the pairs of inspectors awarded different grades after observing the same lesson. In the majority of these cases, the pairs of inspectors arrived at judgements which were one grade apart, for example, one graded a lesson '3' another '4'. However, in 3 % of cases the difference was two grades. The statistical correlation between the two sets of inspectors' judgements was $r = 0.81$. At the 'failure boundary' between grade 4 (satisfactory) and grade 5 (less than satisfactory), only two thirds of inspectors in OFSTED's research arrived at the same judgement. Matthews et al. (1998) however

found that two inspectors are likely to identify the same strengths and weaknesses in the teaching and to arrive at similar conclusions about its overall quality.

A study by Pietsch and Tosana (2008) also examined the extent to which the results of classroom observation can be generalised beyond an individual inspector. A generalisability study was used to simultaneously analyse the effect of several different factors on the evaluation (the rating of a lesson, of specific substandards by specific observers or even the interaction between these factors). Pietsch and Tosana conclude that it is mainly an interaction between an observer and an item that causes bias in evaluation, if only to a moderate degree (about 9 % of total variation). Further, their analyses reveal leniency/severity effect tendencies among the inspectors. Although these evaluation tendencies are quite slight, the authors note that these biases persist despite intensive inspector training (Pietsch and Tosana 2008). De Jong and Reezigt (2007) have reached similar findings with regard to inspectors' rating tendencies for Dutch inspections.

Sinkinson and Jones (2001) examine whether or not it is possible to distinguish properly between consistency of application of the published inspection framework by school inspectors and the loading given to any particular criterion within the framework. They analysed 64 OFSTED reports on secondary mathematics ITE (initial teacher education) courses to look for evidence of the demarcation of grade boundaries, particularly what distinguishes one grade from another, examine the tone of reports to see whether descriptions of weaknesses outweigh expressions of course strength and investigate the reliability of the judgements by searching for possible inconsistencies across equal grades. Their study shows that the length of text in reports on different indicators, as well as the evidence provided alongside judgements, differs substantially. They also found inconsistencies in evidence, as well as a considerable disparity in the descriptors used to validate judgements. A comparison of grades under an old and revised inspection framework also indicated substantial differences in the distribution of grades, particularly with respect to grades 1 and 2; it appeared to be much harder to be awarded a grade 1 in 1997/1998 than in 1996/1997.

A similar, but more qualitative approach was used by Sowada (2010) who investigated how school inspectors arrive at their judgements given that they are required to draw on and synthesise diverse forms of evidence stemming from observations, interviews, surveys, work samples, performance data, school policies and self-evaluation, as well as other forms of documentation. He investigated how inspectors individually and cooperatively aggregate various 'minor' judgements into higher-order judgements. His study included three methods: shadowing school inspectors during inspection visits and analysing how they discuss evidence and judgements with each other and with stakeholders of the inspected school; conducting interviews before and after an inspection to facilitate experience-near accounts; and providing anonymised primary school data and sources that are usually used to inform inspection assessments and having school inspectors analyse the data and how they would assess the school.

A different methodological design was used by Lankes et al. (2013) to evaluate to what extent information from observations and surveys during an inspection

contribute to the final rating results of inspectors. The analysis demonstrates that the primary data source used for judging the quality of teaching are classroom observations and that questionnaires do not explain additional variance between judgments. Thus, it seems that inspectors perceive self-collected data, data generated by a collection process in which they have been actively involved, to be more reliable than other data sources. Another result of this study was that 60–85 % of the variance in judgements could not be explained by the data sources at all. Like Sowada and Dedering (2014) found within another qualitative study this may be a result of discretion effects – inaccurate ratings are a result of indeterminacy (e.g. Sadler 2009) – during an inspection. Exploring the discretion that school inspectors have in their interpretations and decision-making, these authors found that applying discretion while judging the quality of schools and teaching may be beneficial in stimulating school improvement through inspection and that inspectors therefore adapt their judgments to the specific situation of the inspected school within a discursive validation.

Other methods suggested by Matthews et al. (1998) to evaluate inter-rater reliability of assessments include the proportional agreement (pa) between pairs of inspectors (evaluating the proportion of dual observations for which the two teaching grades are identical), calculating a Gower coefficient (a measure of average agreement per observation between inspectors who observe the same lesson dually).

Studies about responses of schools to inspection measures include descriptions of how schools prepare for visits, and potentially try to game or manipulate inspections to improve their rating. A categorization of such behaviours has been published by De Wolf and Janssens (2007) who distinguish between intended and unintended strategic behaviour of schools. Unintended strategic behavior arises when behavior in schools (and as a result the content and organization of education) is influenced by the assessor and/or by the method of working used for the assessment. This means that schools unintentionally focus on only the elements that are assessed in school inspections. Daily practice is morphed into something that is measurable, transparent and auditable. Intended strategic behavior occurs when schools try to improve their status on the measures used in school inspections without creating a commensurate improvement in the educational processes or output the data are intended to measure. Schools for example manipulate records, or put up a show during lesson observations. Such behaviors invalidate inspection assessments as they will not be an accurate reflection of the school's quality assurance. These responses will be explained in more detail in Chap. 5.

3.4.5 Validity Evidence Based on Consequences of Testing

Validity evidence based on consequences of testing refers to evaluating both the intended and the unintended consequences associated with a testing program. An evaluation of consequences is relevant, according to Kane (2006) as it can pinpoint weaknesses or problems in the measurement procedure and is therefore closely

connected to the previous category. When tests for example have high stakes consequences they may lead to the strategic responses described in the previous section and invalidate the measures. Consequences include teachers', students', and administrators' interpretations of inspection assessments, as well as the actions they take when interpreting inspection results (e.g. improving instruction and learning, changing policy, closing a school). Evidence about consequences can indicate sources of invalidity such as construct underrepresentation or construct irrelevant components. Following Lane and Stone (2002) and CCSSO (2004), the overarching question to collecting evidence based on consequences of inspection systems is:

- Does this inspection system do what it is intended to do?
- To what extent are the anticipated benefits of inspections realized?
- To what extent do unanticipated benefits, both negative and positive, occur?
- Is the design of the inspection system and measures fit for how it is, and how it should be used?

An important starting point for gathering consequential evidence is the vision/mission of Inspectorates of Education, and the goals they have formulated for their inspections of schools. These goals often reflect (as we explained previously) improvement of schools on the standards in the inspection framework, but may also reflect a broader contribution to the functioning of the education system; e.g. by means of informing parental school choice or informing national policy.

An evaluation of these propositions typically includes collecting evidence (e.g. in the form of surveys, or more direct measures, such as classroom artefacts or classroom observations) from different stakeholders (e.g. teachers, principals and students), testing the relationship between improvement of schools and teachers' classroom instruction practices (assuming these are the intended outcomes) and the schools' scores on the inspection measures and how schools and teachers were assessed (see Linn 2000; Nelson and Ehren 2014).

An example of such an approach was used by a number of researchers in a recent EU study (see Ehren et al. 2013[2]) to evaluate the impact of school inspections. These authors used a survey of principals (and in England and the Netherlands also teachers) for three consecutive years to test assumptions about the impact of school inspections. An overview of other studies and their results are provided in Chaps. 4 and 5.

3.5 Challenges and Tensions in Inspection Frameworks, Methods and Processes

Despite these efforts to ensure the validity of school inspections, there are two common challenges and tensions we want to address at the end of this chapter.

The first tensions involves the seemingly irreconcilability of procedures to ensure both the technical accuracy of inspection outcomes, as well as the intended out-

[2] See also www.schoolinspections.eu for two literature reviews on effects and side effects of school inspections.

comes of inspections. Inspection frameworks and protocols guide inspectors on what is required to assess a school as sufficient or outstanding and what a failing school generally looks like. Although these protocols are not intended to act as frameworks for school improvement and school planning, schools have a strong incentive to familiarise themselves with these documents and to make sure they have required practices in place of what a good school looks like. The UK select Committee (1999) explains how schools tend to revert to inspection templates when developing their school internal processes and much circumstantial evidence reports of headteachers in England referring to 'Ofsted wants to see…' or 'Ofsted expects…', who feel that they will be marked down if they do not do things in what they perceive to be the 'Ofsted way'. As a result, inspection protocols become self-fulfilling prophecies. There seems to be a difficult interaction and trade-off between the level of specificity of the inspection criteria (which would enhance validity of assessments), and the subsequent manipulation and gaming of the same assessments by schools (which would consequently invalidate the inspection assessments).

A second, related issue is the standardized nature of inspection frameworks which assumes that the same standards can accurately and unambiguously reflect every school. The difficulty here is, according to Gilroy and Wilcox (1997), in understanding how something as necessarily varied as the social practices created by the different social contexts within which schools operate can be pinned down by one set of inspection criteria. The immense variety across schools also creates difficulties in how school inspectors interpret the meaning of inspection criteria, how they apply them when assessing schools (are all criteria of equal value and importance across all schools?), and how they aggregate observations to provide an overall objective judgement of a school and its teaching.

What is needed, according to Gilroy and Wilcox (1997) is to embed inspection frameworks in a particular context, so that its meaning can be understood by seeing it in operation. This would require standard setting procedures to identify what is meant by effective teaching in a particular school, with its particular blend of children, with their particular social background, where one particular practice appears to work well, whereas another does not. Such procedures should allow for an open discussion around common assumptions about what is 'reasonable' and how effective teachers are expected to behave, and how that behaviour may vary across different contexts. Following Gilroy and Wilcox (1997), standard setting procedures must accept the essential ambiguity of social judgements and build in protocols that allow for firmly locating inspection judgements within their appropriate context; acknowledging that what counts as 'reasonable' or 'effective' has to be understood by reference to a particular context. This notion of contextualized assessments was also described by Messick (1994) who suggests a number of ways of coping with interactions with context when measuring skills. Approaches include inferences about a particular construct from consistencies in behaviour across context, or across varied tasks within context. One could also treat a construct (or skill) as revealed in different contexts as qualitatively different constructs or skills, looking at all potentially relevant construct-context combinations.

Unfortunately time constraints of current inspection practices, prescriptive and detailed frameworks, lack of expertise of inspectors and fear of schools complaining about unfair assessments when they are assessed differently often don't allow for or limit such a substantive reflection. A high trust environment seems an important condition for any kind of validity. As Professor Robin Alexander points out in the fourth report of the UK Select Committee 'the best inspections use the framework as just that, a framework around which they gather evidence and reflect before delivering the most fitting judgement'.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *Hong Kong Journal of Occupational Therapy, 20*(1), 30–42.

Council of Chief State School Officers (CCSSO). (2004). *A framework for examining validity in state accountability systems*. Washington, DC: CCSSO.

De Jong, R., & Reezigt, G. (2007). *Interraterreliability of inspectors*. Herne: Internationale Tagung 'Validität von Daten im Rahmen von Schulinspektion'.

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

Ehren, M. C. M., Leeuw, F. L., & Scheerens, J. (2005). On the impact of the Dutch Educational Supervision Act. Analyzing assumptions concerning the inspection of primary education. *American Journal of Evaluation, 26*(1), 60–77.

Ehren, M. C. M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of school inspections on teaching and learning – Describing assumptions on causal mechanisms in six European countries. *Educational Assessment, Evaluation and Accountability, 25*(1), 3–43.

Gaertner, H., & Pant, H. A. (2011). How valid are school inspections? Problems and strategies for validating processes and results. *Studies in educational evaluation, 37*(2), 85–93.

Gilroy, P., & Wilcox, B. (1997). Ofsted, criteria and the nature of social understanding: A Wittgensteinian critique of the practice of educational judgement'. *British Journal of Educational Studies, 45*(1), 22–38. http://elibrary.ioe.ac.uk/login?url=http://dx.doi.org/10.1111/1467-8527.00034

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Practice and Issues, 21*(1), 16–22.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *Educational measurement, 4*, 433–470.

Kane, M. T. (2006). Validation. In: R. L. Brennan (Ed.), Educational measurement (4th ed.). (pp. 17–64). Westport: American Council on Education/Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Measurement: Issues and Practice, 21*(1), 23–30. http://elibrary.ioe.ac.uk/login?url=http://dx.doi.org/10.1111/j.1745-3992.2002.tb00082.x

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28.

Lankes, E. M., Vaccaro, D., & Gegenfurtner, A. (2013). Wie kommen die Evaluationsteams zu ihrer Einschätzung der Unterrichtsqualität bei der Externen Evaluation? *Unterrichtswissenschaft, 41*, 197–215.

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation, 24*(1), 5–20.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(4), 4–16.

Marion, S., & Gong, B. (2003) Evaluating the validity of state accountability systems. The 2003 Reidy Interactive Lecture Series. http://www.nciea.org/publications/RILS2003_BGSM03.pdf

Matthews, P., Roger Holmes, J., Vickers, P., & Corporaal, B. (1998). Aspects of the reliability and validity of school inspection judgements of teaching quality. *Educational Research and Evaluation, 4*(2), 167–188.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Morrison, K. (1996). Why present school inspections are unethical. *Forum, 38*(3), 79–80.

Nelson, R., & Ehren, M. C. M. (2014). *Review and synthesis of evidence on the (mechanisms of) impact of school inspections*. http://schoolinspections.eu/wp-content/uploads/downloads/2014/02/Review-and-synthesis-of-evidence-on-the-mechanisms-of-impact-of-school-inspections.pdf

Newton, P., & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301–319.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Thousand Oaks: Sage.

Pant. (2010). *How valid are school inspections? Problems and strategies for validating processes and results*. SICI conference. Retrieved August 2014, http://www.stebis.de/en/publikationen/Pr__sentation_How_valid_are_school_inspections_gaertner__pant.pdf?1374053858

Pietsch, M. (2010). Evaluation von Unterrichtsstandards. *Zeitschrift für Erziehungswissenschaft, 13*, 121–148.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179.

Scheerens, J., Seidel, T., Witziers, B., Hendriks, M., & Doornekamp, G. (2005). *Positioning and validating the supervision framework. Positioning the supervision frameworks for primary and secondary education of the Dutch Educational Inspectorate in current educational discourse and validating core indicators against the knowledge base of educational effectiveness research*. Enschede/Kiel: University of Twente/IPN

Schwippert, K. (2015). Daten für die Schulentwicklung – auf die Perspektive kommt es an. In M. Pietsch, B. Scholand, & K. Schulte (Eds.), *Schulinspektion in Hamburg, der erste Zyklus 2007–2013: Grundlagen, Befunde, Perspektiven* (pp. 157–176). Münster: Waxmann.

Sinkinson, A., & Jones, K. (2001). The validity and reliability of Ofsted judgements of the quality of secondary mathematics initial teacher education courses. *Cambridge Journal of Education, 31*(2), 221–237.

Sireci, S. G., & Sucin, T. (2013). Chapter 4. Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology* (Test theory and testing and assessment in industrial and organizational psychology. APA handbooks in psychology, Vol. 1, pp. 61–84). Washington, DC: American Psychological Association.

Sommer, N. (2011). Wie beurteilen schulische Gruppen die erlebte Schulinspektion? In S. Müller, M. Pietsch, & W. Bos (Eds.), *Schulinspektion in Deutschland. Eine Zwischenbilanz aus empirischer Sicht* (pp. 137–164). Münster: Waxmann.

Sowada, M. (2010). *School inspection: Evaluative judgements and how they can be investigated. BERA presentation 0691*. School inspection: Evaluative judgements and how they can be investigated. http://www.beraconference.co.uk/2010/downloads/abstracts/pdf/BERA2010_0691.pdf

Sowada, M. G., & Dedering, K. (2014). Ermessensspielräume in der Bewertungsarbeit von Schulinspektor/innen. *Zeitschrift für Bildungsforschung, 4*(2), 119–135.

Pietsch and Tosana (2008). In: http://www.sciencedirect.com/science/article/pii/S0191491X11000368

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

UK Select Committee on Education and Employment. (1999). The work of Ofsted. http://www.publications.parliament.uk/pa/cm199899/cmselect/cmeduemp/62/6212.htm

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 & 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 Schulinspektionsprozesses und (b) Re-

	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 Informa ionen		
		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 Mo ivation und das Commitment
			Schulleitung
			Stabilität des Kollegiums
			Zeit, Veränderungen umzusetzen
Merkmale des Individuums		Vorwissen/Expertise	
		Motivation, Emotion, Volition	
	Merkmale der Entscheidungsträger	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

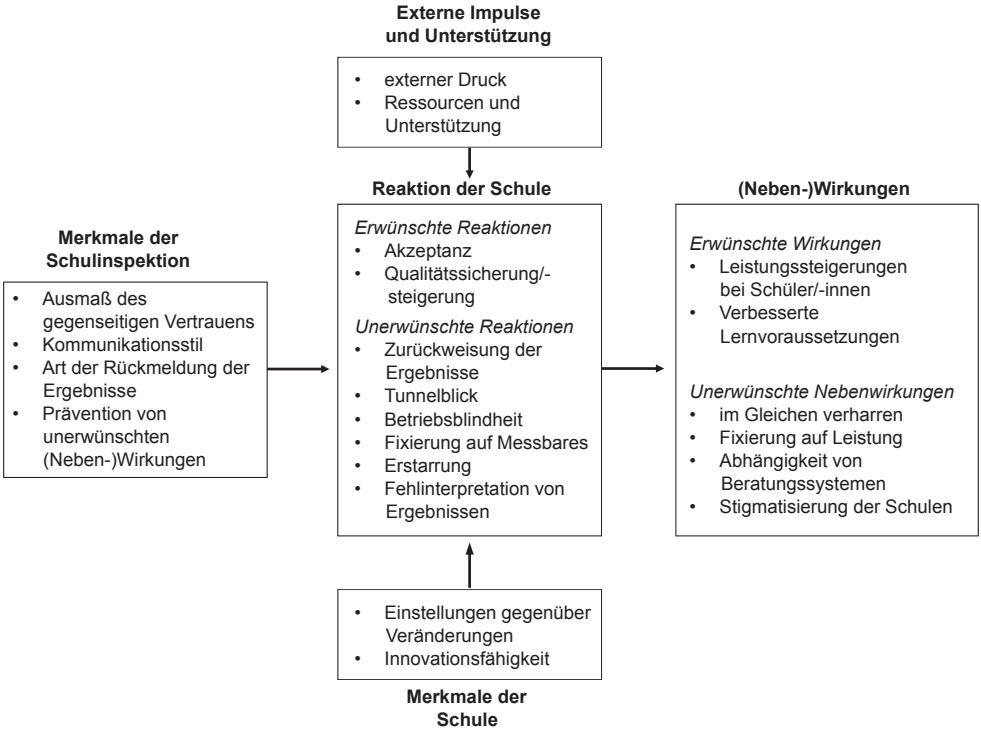


Abb. 1: Modell zur Wirkungsweise von Schulinspektion (vgl. Ehren & Visscher, 2006)

aktionen 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). Hier 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 Schulbeteiligter, 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 Schulbesuches, (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 aus 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* der Schulinspektion als CB (Kontrollgruppe – before) und den Mittelwert der Schülerleistungen der nicht-inspizierten Schulen *nach* 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

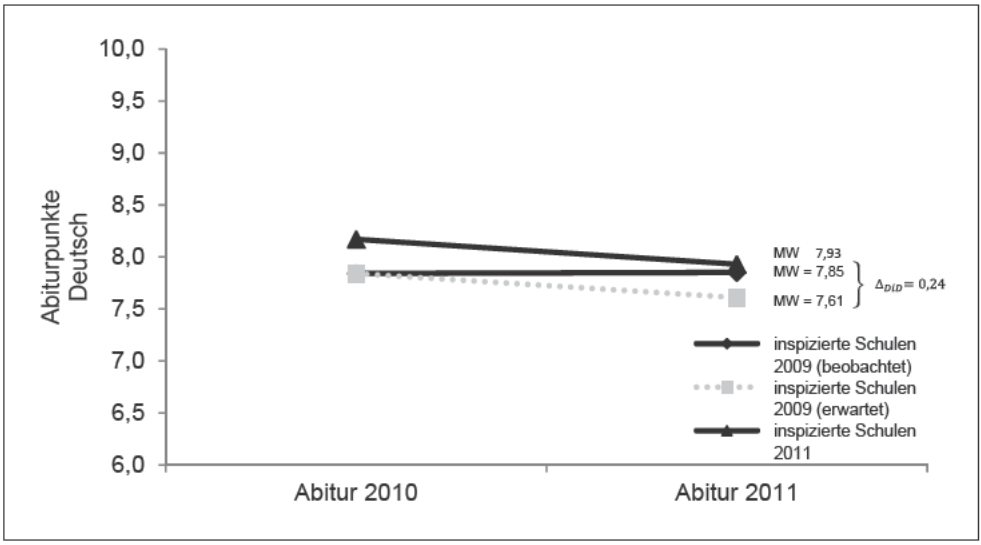
einer zweiten Fachlehrkraft durchgesehen, die sich entweder der ersten Bewertung anschließt oder ein ergänzendes Gutachten mit Bewertung anfertigt. Abschließend legt die oder der Vorsitzende des Prüfungsausschusses die endgültige Punktzahl der Abiturarbeit fest. Beträgt die Differenz der im Erstgutachten 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.



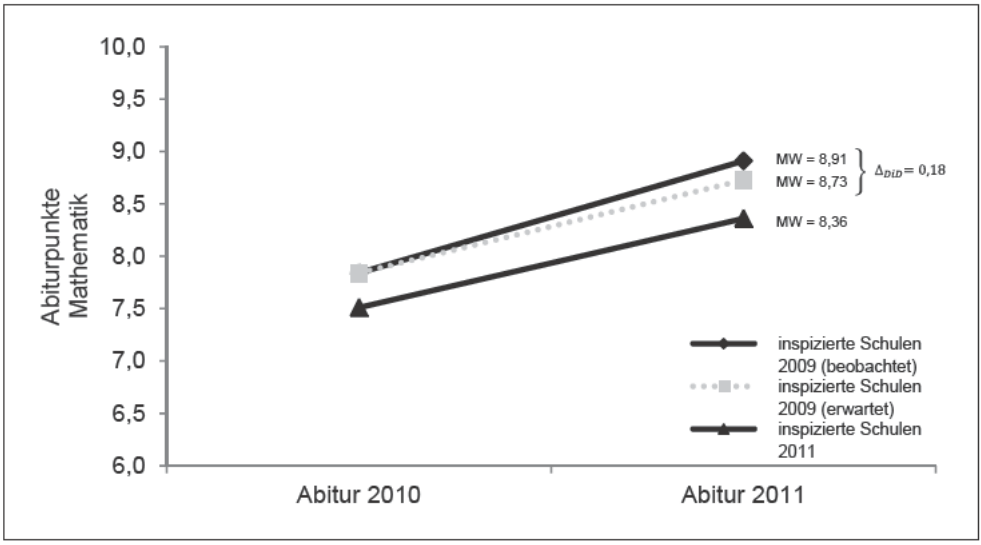


Abb. 2 und 3: Trends im Zentralabitur Deutsch (Abb. 2) und Mathematik (Abb. 3) für die Treatment- (beobachtet und erwartet) und Vergleichsgruppe (beobachtet)

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

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.[3] 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[4] 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 Domänen Leseverstehen und Mathematik zu allen Messzeitpunkten von allen Schülerinnen und Schülern bearbeitet wurden. Die Tests der KESS-Untersuchung waren als ro-

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

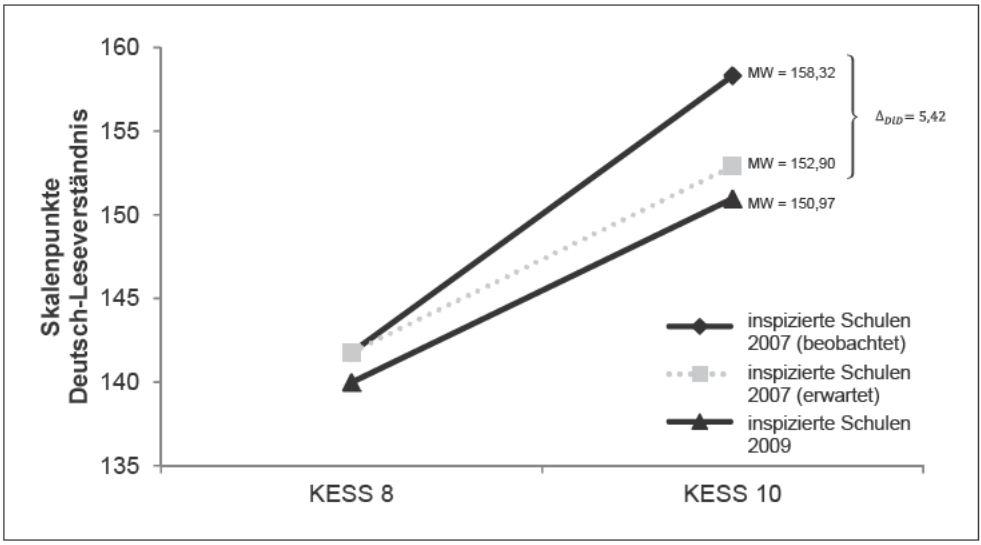
4 Eine Ausnahme bildet die Erhebung KESS 10/11, in der die Teilnahmequoten – trotz Teilnahmepflicht – nur im Bereich von 83 bis 86 Prozent lagen.

tiertes 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 ($\Delta_{DiD} = 0.31$). Schulinspektion, verstanden als Intervention auf Schulebene, hat demnach keinen Einfluss auf die Lernentwicklung von Schülerinnen und Schülern im Fach Mathematik.



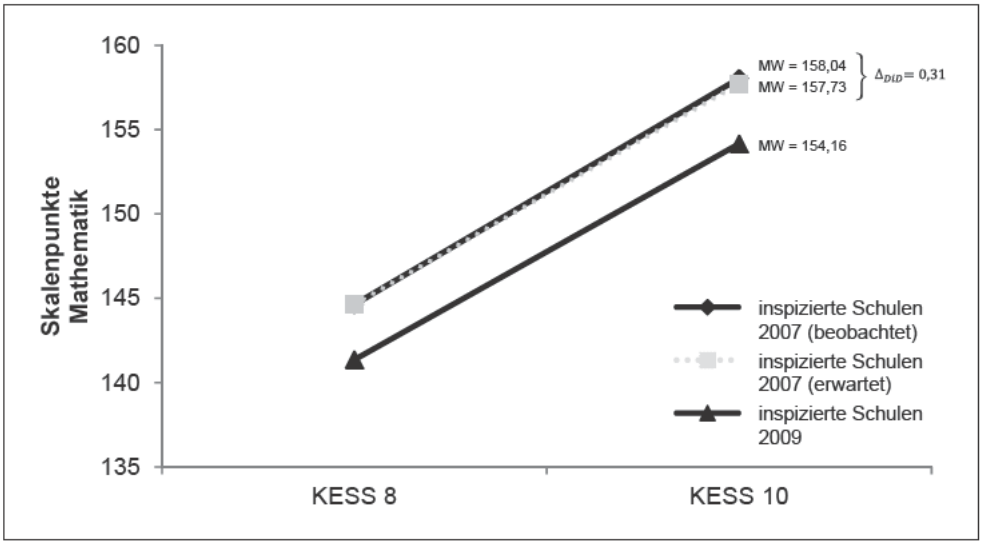


Abb. 4 und 5: Lernentwicklungen von Schülerinnen und Schülern im Deutsch-Leseverständnis (Abb. 4) und in Mathematik (Abb. 5) für die Treatment- (beobachtet und erwartet) und die Vergleichsgruppe (beobachtet)

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 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

Arbeitsgruppe Schulinspektion (Hrsg.)

Schulinspektion als Steuerungsimpuls?

Ergebnisse aus Forschungsprojekten

Die Arbeitsgruppe Schulinspektion ist vertreten durch Oliver Böhm-Kasper, Thomas Brüsemeister, Fabian Dietrich, Lisa Gromala, Martin Heinrich, Maike Lambrecht, Bianca Preuß, Matthias Rürup, Odette Selders und Jochen Wissinger

Springer VS

Herausgeber
Arbeitsgruppe Schulinspektion
Gießen, Deutschland

Arbeitsgruppe Schulinspektion

Oliver Böhm-Kasper, Universität Bielefeld, Deutschland
Thomas Brüsemeister, Justus-Liebig-Universität Gießen, Deutschland
Fabian Dietrich, Leibniz Universität Hannover, Deutschland
Lisa Gromala, Justus-Liebig-Universität Gießen, Deutschland
Martin Heinrich, Universität Bielefeld, Deutschland
Maike Lambrecht, Universität Bielefeld, Deutschland
Bianca E. Preuß, Justus-Liebig-Universität Gießen, Deutschland
Matthias Rürup, Bergische Universität Wuppertal, Deutschland
Odette Selders, Universität Bielefeld, Deutschland
Jochen Wissinger, Justus-Liebig-Universität Gießen, Deutschland

Educational Governance
ISBN 978-3-658-10871-7 ISBN 978-3-658-10872-4 (eBook)
DOI 10.1007/978-3-658-10872-4

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer VS

Lektorat: Stefanie Laux, Daniel Hawig

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Von der Schulinspektion zur Schulentwicklung

Welche Rolle spielen innerschulische Voraussetzungen?

Marcus Pietsch, Tobias Feldhoff und Lina Sophie Petersen

Zusammenfassung

Schulinspektionen in Deutschland sollen, als Teil einer evidenzbasierten Outputsteuerung, vor allem die Schul- und Unterrichtsentwicklung stimulieren. Hierbei wird erwartet, dass schulische Akteure die Informationen, die im Rahmen der Inspektion generiert werden, aktiv nutzen, um wissensbasierte Entscheidungen zur weiteren Ausgestaltung von Schule und Unterricht zu treffen. Bisher jedoch konnte kein eindeutiger Zusammenhang von Schulinspektion und Schulentwicklungsaktivitäten, bei allerdings hoher Akzeptanz der Schulinspektion, festgestellt werden. Dabei wird deutlich, dass den innerschulischen Verarbeitungsprozessen in Schulen bisher zu wenig Beachtung geschenkt wurde. Die vorliegende Studie versucht diese Lücke zu schließen, indem sie die Voraussetzungen, unter denen an Schulen mit Informationen aus Schulinspektionsverfahren umgegangen wird, näher in den Blick nimmt und diese in Beziehung zu konkreten Schulentwicklungsmaßnahmen an Schulen setzt. Hierfür wird das Modell der Kapazitäten organisationalen Lernens genutzt und Daten von 49 Schulen, die infolge der Schulinspektion in Hamburg Entwicklungsmaßnahmen ergriffen haben, analysiert. Die Ergebnisse zeigen, dass a) Schulen infolge einer Inspektion Entwicklungsmaßnahmen generell vor allem im Bereich des Unterrichts sowie der Schul- und Unterrichtsorganisation ergreifen, wobei innerschulische Voraussetzungen diesbezüglich keinen Unterschied zwischen Schulen machen, b) an Schulen mit guten innerschulischen Voraussetzungen für die Schulentwicklung infolge einer Inspektion vor allem Maßnahmen im Bereich von Schulleitung und Schulmanagement ergriffen werden und dass c) Schulen mit geringen Kapazitäten infolge einer Inspektion vor allem auf möglichst viele symbolische Maßnahmen setzen, die vor allem dazu beitragen, die Außenwirkung der Schule zu verbessern. Die Studie macht

A. Schulinspektion (Hrsg.), *Schulinspektion als Steuerungsimpuls?,* Educational Governance 25,
DOI 10.1007/978-3-658-10872-4_7,

darüber hinaus jedoch deutlich, dass selbst das hier genutzte Forschungsdesign nur bedingt ausreicht, um den Einfluss von Schulinspektionen auf die Schulentwicklung zu analysieren. Um diesen Zusammenhang zukünftig besser zu verstehen, scheint es sinnvoll, Schulen im Längsschnitt direkt im Anschluss an die Inspektion oder gar schon im Vorfeld des Inspektionsbesuchs bis zum nächsten Inspektionszyklus wissenschaftlich zu begleiten.

1 Einleitung

Schulinspektionen in Deutschland sollen, als Teil einer evidenzbasierten Outputsteuerung (vgl. Altrichter und Maag Merki 2010), vor allem die Schul- und Unterrichtsentwicklung stimulieren. Hierbei wird erwartet, dass schulische Akteure die Informationen, die im Rahmen der Inspektion generiert werden, aktiv nutzen, um wissensbasierte Entscheidungen zur weiteren Ausgestaltung von Schule und Unterricht zu treffen. Neuere Studien in Deutschland (vgl. z. B. Heinrich et al. 2014; Gärtner et al. 2009; Dedering et al. 2012; Böhm-Kasper und Selders 2013) zeigen, dass diese Erwartungen an die Rezeption und Nutzung von Inspektionsdaten für Schul- und Unterrichtsentwicklung eher nicht erfüllt werden. Bisher konnte kein eindeutiger Zusammenhang von Schulinspektion und Schulentwicklungsaktivitäten, bei allerdings hoher Akzeptanz der Schulinspektion, festgestellt werden. Dabei wird deutlich, dass den innerschulischen Verarbeitungsprozessen in Schulen bisher zu wenig Beachtung geschenkt wurde. So zeigen Studien beispielweise, dass die Rückmeldungen der Schulinspektion für viele Schulen gerade nicht evident, im Sinne von unmittelbar einsichtig, sind (vgl. Heinrich et al. 2014), sondern eine aktive Rekontextualisierungsleistung erfordert, bei der die Ergebnisse der Schulinspektion mit der eigenen Einschätzung zur Qualität der Schule und dem schulischen Umfeld in Beziehung gesetzt werden müssen. Diesbezüglich kann die soziale Interaktion zwischen Schulinspektorinnen und Schulinspektoren und innerschulischen Akteuren während einer Inspektion wesentlich dazu beitragen Veränderungs- und Innovationsprozesse in Gang zu setzen. Entsprechend betont Sowada (2015, S. 151), dass es – bezogen auf die Entwicklungsfunktion von Schulinspektion – wichtig ist, dass es „Inspektorinnen und Inspektoren gelingt, den schulischen Akteuren zu neuen Einsichten zu verhelfen, für Veränderungsbedarf zu sensibilisieren und für Entwicklungsaufgaben zu motivieren." Welche schulischen Verarbeitungsprozesse konkret bei der Rezeption und Nutzung der Inspektionsdaten für Schul- und Unterrichtsentwicklung eine Rolle spielen und welche Voraussetzung im Sinne einer Schulentwicklungskapazität hierfür notwendig sind, ist bisher weitgehend unklar.

An diesem Punkt setzt dieser Beitrag an. Anhand eines theoretischen Schulentwicklungsmodells wird untersucht, ob Schulen mit unterschiedlichen Schulentwicklungskapazitäten gemessen mithilfe der Bewertung der Schulinspektion sich in Bezug auf initiierte und umgesetzte Maßnahmen der Schul- und Unterrichtsentwicklung im Anschluss an die Schulinspektion unterscheiden. Hierfür werden in Abschnitt 2 zunächst die allgemeinen Annahmen zur Wirkungsweise von Inspektionen sowie ihre Grenzen aufgezeigt. Anschließend werden in Abschnitt 3 die Kapazitäten organisationalen Lernens als Modell einer evidenzbasierten Schulentwicklung vorgestellt. In Abschnitt 4 werden schließlich das Forschungsdesiderat und die Fragestellung skizziert. Die Evaluation von Inspektionswirkungen am Beispiel der Schulinspektion Hamburg erfolgt in Abschnitt 5. Hier wird zunächst das Rückmeldeverfahren der Hamburger Schulinspektion erläutert (5.1). Anschließend werden die Qualitätsbereiche des Hamburger Orientierungsrahmens mit den Dimensionen der Kapazitäten verglichen (5.2). Danach wir die Anlage und Durchführung der Untersuchung mit dem methodischen Vorgehen und der Stichprobe erläutert (5.3). Dann werden die Befunde einzeln dargestellt (5.4). Zum Schluss werden die Befunde in Abschnitt 6 noch einmal zusammengefasst und anhand des Modells der Kapazitäten organisationalen Lernens sowie dem Forschungsstand zur Schulinspektion in Deutschland kritisch diskutiert sowie ein Ausblick für zukünftige Forschungsvorhaben gegeben.

2 Annahmen zur Wirkungsweise von Inspektionen und ihre Grenzen

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 und Klemm 2009). Die Schulinspektion hat dabei vier Funktionen (vgl. Landwehr 2011):

1. **Katalysefunktion**: Schulinspektion soll die Schulentwicklung fördern, indem sie durch die Rückmeldung Handlungsfelder identifiziert und nächste Entwicklungsschritte aufzeigt. Durch die Rückmeldungen und Berichte kann und soll der innerschulischen Diskussions- und Entwicklungsprozess stimuliert werden.

2. **Rechenschaftsfunktion**: Schulinspektion leistet einen Beitrag zur staatlichen Gewährleistung, indem sie schulische Qualität sichtbar macht und (Mindest-) Standards sichert.
3. **Normendurchsetzungsfunktion**: Schulinspektion transportiert und vermittelt die Inhalte von Referenz- oder Orientierungsrahmen. Die dort formulierten normativen Erwartungen an schulische Qualität werden in den Schulen vor allem mit Blick auf eine anstehende Evaluation verarbeitet und so aktiv aufgenommen.
4. **Erkenntnisfunktion**: Schulinspektion leistet einen Beitrag zum Bildungsmonitoring, indem sie die einzelschulischen Befunde zu Aussagen über die Qualität des Gesamtsystems verdichtet und Steuerungserfordernisse offenlegt. Diese Informationen dienen der Administration und der Politik als Grundlage für die Systemsteuerung.

Schulinspektionen in Deutschland verfolgen dabei derzeit primär das Ziel, Schul- und Unterrichtsentwicklung mittels der Rückmeldung von Informationen zur extern wahrgenommenen Qualität von Schule und Unterricht zu stimulieren (vgl. Böttcher und Kotthoff, 2010). Diesbezügliche Wirksamkeitserwartungen an Schulinspektionen knüpfen dabei vor allem an die Forschung zum zielorientierten Feedback an (vgl. Kluger und DeNisi 1996; Visscher und Coe 2003). Entsprechend wird erwartet, dass das Aufzeigen von Differenzen zwischen normativ vorgegebenen Soll- und empirisch beobachteten Ist-Ständen dazu führt, 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 pädagogische Verarbeitungsprozesse in den Blick nehmen, haben Cousins und 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 stimulieren soll und kann. Eine solche Annahme zur Nutzung von Evaluationsbefunden unterstellt dann, dass Entscheidungen rational, auf Basis bereitgestellter Informationen in einem prozessualen Ablauf getroffen werden (vgl. Hyyryläinen und 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 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 Reichweite. Pietsch et al. (2014) zeigen diesbezüglich, dass Reezigt und Creemers (2005) vor allem innerschulische Aspekte der Schulkultur in den Blick nehmen, Helmke und Hosenfeld (2005) darüber hinaus auch auf die Relevanz individueller Persönlichkeitsmerkmale von Lehrenden und schulischen Entscheidungsträgern im Verarbeitungsprozess hinweisen und Cousins und Leithwood (1993) auch noch die soziale Interaktion innerhalb der Schule sowie zwischen Schulbeteiligten und Evaluatoren als Merkmal wirksamer Evaluationen betrachten.

Welche Merkmale die Wirksamkeit speziell von Schulinspektionen konkret moderieren, haben Ehren und Visscher (2006, vgl. Abb. 1), basierend auf Visscher und Coes (2003) Arbeiten zur Nutzung von Schul-Performance-Feedback-Systemen, herausgearbeitet.

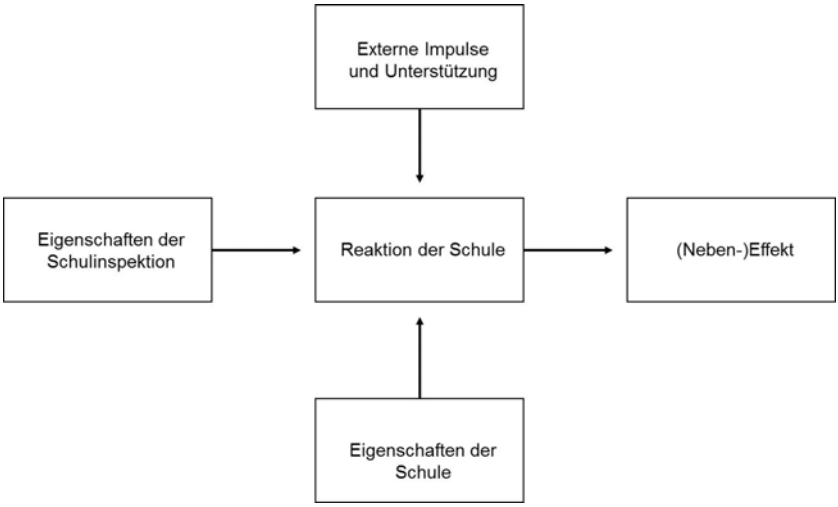


Abb. 1 Modell zur Wirkungsweise von Schulinspektion

Grundsätzlich gehen Ehren und Visscher davon aus, dass Wirkungen von Schulinspektionen als Folgen eine kausalen Wirkungskette aus (a) Merkmalen des Schulinspektionsprozesses und (b) Reaktionen der Schulen auf den Prozess und die Ergebnisse der Inspektion entstehen, wobei neben Unterstützungsmaßnahmen aus dem weiteren Bildungssystem vor allem, auch innerschulische Voraussetzungen der Schule wichtige Determinanten für den Umgang und die Nutzung von Schulinspektionsbefunden Grundlage für die Schul- und Unterrichtsentwicklung sind. Insbesondere die Haltungen, Fähigkeiten und Kompetenzen der Schulverantwortlichen und Lehrerschaft an der evaluierten Schule spielen letztlich eine wichtige Rolle dabei, ob Inspektionsbefunde für die Weiterentwicklung von Unterricht und Schule genutzt werden.

Verschiedene Untersuchungen machen nun jedoch deutlich, dass die gängigen Modelle zur Beschreibung von Schulinspektionswirksamkeit zu unterkomplex sind, um den Untersuchungsgegenstand angemessen zu beschreiben. So zeigt eine Untersuchung von Pietsch (2011), dass die konkrete Nutzung der Schulinspektion als Instrument der Schulentwicklung an Schulen nur zu einem sehr geringen Teil von inspektionsseitigen Faktoren abhängt und es vor allem von innerschulischen Voraussetzungen sowie außerschulischen Unterstützungsmaßnahmen abhängt, ob aus einer Inspektion auch tatsächlich Entwicklungen und Innovationen erwachsen. Wie komplex dieser Sachverhalt ist, zeigen dann auch Ehren und Visscher (2008) in einer Fallstudie an zehn niederländischen Schulen, die anhand ihres Innovationspotenziales ausgewählt wurden, auf. So zeigt diese Untersuchung, dass sowohl ein hohes Innovationspotenzial der Schule als auch ein inspektionsseitiges Rückmeldeformat, welches auf die Schwächen der inspizierten Schule fokussiert, wichtig sind, wenn Schulen sich infolge von Inspektionen weiterentwickeln sollen. Entsprechend kritisieren die Autoren die derzeit gängige Annahme, dass die Intervention durch Schulinspektionen alleine und direkt zu einer Qualitätssteigerung an Schulen führe, als naiv, da die gängigen Analysemodelle der Wirksamkeitsforschung und die damit einhergehenden Wirksamkeitsannahmen zu unterkomplex seien und innerschulisch sowie Kontextbedingungen und Interdependenzen nicht berücksichtigen würden. Konsequenterweise resümieren Pietsch et al. (2013, S. 162) in einer Zusammenschau zum Stand der Schulinspektionsforschung: „Die Forschung zur Effektivität von Schulinspektionsverfahren befindet sich (…) in etwa im Stadium der Schuleffektivitätsforschung der späten 1960er bis frühen 1970er Jahre, in der mittels ökonomischer Studien – die einem reinen Input-Output-Paradigma folgten und weder Prozess- noch Kontextvariablen berücksichtigten diagnostiziert wurde, dass „schools bring little influence to bear on a child's achievement"".

3 Kapazitäten organisationalen Lernens als Modell einer evidenzbasierten Schulentwicklung

Während für den Bereich der Kontextmerkmale in jüngster Zeit erste vielversprechende Ideen und Erklärungsmodelle eingebracht wurden und sich zeigt, dass insbesondere Entwicklungsdruck dazu beiträgt, dass sich Schulen infolge von Inspektionen entwickeln (vgl. Altrichter und Kemethofer 2015; Ehren et al. 2015) stehen vergleichbare Modellierungen für innerschulische Merkmale und Prozesse derzeit noch aus. Um diesem Problem zu begegnen, bietet es sich an auf Modelle zurückzugreifen, die die oben beschriebenen Lücken belastbar ausfüllen können. Ein solcher Ansatz, der im Rahmen der Schulinspektionsforschung besonders vielversprechend und auch einfach umsetzbar erscheint, ist derjenige der „Kapazitäten organisationalen Lernens" (vgl. Feldhoff 2011; Mark und Louis 1999). Dieser Ansatz unterscheidet sich von normativen Ansätzen der Schulentwicklung. Er beruht auf einer organisationstheoretischen Fundierung sowie empirischen und theoretischen Befunden der internationalen Forschung zum Organisationalen Lernen im schulischen Kontext (vgl. z. B. Leithwood und Louis 2000; Louis 2006) unter Bezugnahme allgemeintheoretischer Konzepte zum Organisationalen Lernen (vgl. z. B. Argyris und Schön 1978; Daft und Huber 1987; Duncan und Weiss 1979; March und Olsen 1976). Dabei unterscheidet der Ansatz sieben Dimensionen, die sich Großteils – da sich die Orientierungsrahmen der Schulinspektionen zumeist auf Annahmen der Schuleffektivitätsforschung beziehen (vgl. Ehren und Scheerens 2015) – mit den im Rahmen von Schulinspektionen zu evaluierenden Bereichen decken. Aufgrund dieser Tatsache ist es mithilfe von Inspektionsdaten möglich zu prüfen, inwieweit schulische Kapazitäten relevant für die Weiterentwicklung von Schule und Unterricht sind. Für die Analyse der schulischen Verarbeitung von Impulsen der Schulinspektion eignet sich dieser Ansatz besonders, da hier Interdependenzen sowie Erwerb, Austausch und Verarbeitung von Wissen als Ausgangspunkt und Produkt von Lernen innerhalb der Schule gesehen werden. Der Ansatz der Kapazitäten differenziert diesbezüglich sieben Dimensionen aus, die relevant sind, damit Schulentwicklung infolge von Evaluationen erfolgen kann:

(1) Organisationsstruktur

Die Ausgestaltung der Organisationsstruktur der Schule ist Teil der Kapazität Organisationalen Lernens. Nach Kruse et al. (1995) behindern traditionelle Schulstrukturen Organisationales Lernen; ist doch die zeitliche und räumliche Strukturierung des Schulalltages kaum auf Kooperation der Lehrkräfte ausgelegt: Organisationsstrukturen zeigen eine starke Fragmentierung (vgl. Rolff 1993); Bildungs- und Erziehungsziele sind in einzelne Teilziele für Jahrgangsstufen,

Fächer mit entsprechenden Zeitkontingenten, Fachlehrkräfte und Einteilung in Unterrichtseinheiten zergliedert (Arbeitsgruppe Bildungsforschung/Bildungsplanung 2004). Eine Zusammenführung dieser Teilbereiche findet allenfalls auf der curricularen Ebene statt. Die formalbürokratische Verwaltungsstruktur (vgl. Mintzberg 1992) erschwert das Lernen im Sinne der kooperativen Bearbeitung interdependenter Probleme. Jedoch wird gerade Aushandlungen für Innovationen eine hohe Bedeutung beigemessen. Die Aushandlung wird durch die traditionell hohe Autonomie der Lehrkräfte (vgl. Lortie 1972) erschwert.

Die Strukturen einer Organisation können Organisationales Lernen fördern, indem sie geeignete Rahmenbedingungen für Kooperation schaffen, um die Verbreitung und den Transfer von Wissen zu ermöglichen (vgl. Daft und Huber 1987; Duncan und Weiss 1979; Jones 2006). Diese Rahmenbedingungen können durch Änderungen der Zeitstruktur und den Aufbau von institutionalisierten Teams (wie z. B. Jahrgangs-, Klassen- oder Fachteams, sowie Steuer- und Arbeitsgruppen) erreicht werden (vgl. Kruse und Louis 2000). Innerhalb der Teams kann Wissen ausgetauscht und Interdependenzen bearbeitet werden. Während der Wissensaustausch vorrangig innerhalb der Teams erfolgt, besteht für Steuergruppen und Schulleitung zudem die Aufgabe, für einen schulweiten Austausch zwischen den Teams zu sorgen (vgl. Feldhoff 2011; Kruse und Louis 2000; Leithwood und Leonard et al. 2000). Analog zu Steuergruppen können auch informelle schulische Teams derartige Funktionen ausüben.

(2) Gemeinsame Ziel- und Wertvorstellungen und Kooperation im Kollegium

Für Organisationales Lernen ist auch die Bewertung und Entscheidung über die Relevanz von Informationen und wie diese in der Organisation genutzt werden, von großer Bedeutung (vgl. Duncan und Weiss 1979; Hedberg 1981; March und Olsen 1976). „Because a strong professional community is a vehicle for school wide knowledge processing, creating a professional community enhances a school's capacity for organizational learning" (Marks und Louis 1999, S. 713). Nach Weick und Roberts (1993) erfolgt in Teams ein Prozess der kollektiven Sinnkonstruktion („collective mind"). Dieser entsteht aus Mustern gemeinsamer Aktivitäten der Teammitglieder, d. h. aus Handlungsbeziehungen, in einem sozialen System (vgl. Zarcula 2006). „Collective mind is manifest when individuals construct mutually shared fields. The collective mind that emerges during the interrelating of an activity system is more developed and more capable of intelligent action the more heedfully that interrelating is done" (Weick und Roberts, 1993 S. 365). „Collective mind" ist ein Produkt sozialer Prozesse in der Interaktion von Organisationsmitgliedern. Solche Interaktionsprozesse haben einen großen Einfluss auf die organisationale

Nutzung und Weiterentwicklung von Wissen (vgl. Louis und Dentler 1988). Sie helfen den Lehrkräften, die Anschlussfähigkeit und Angemessenheit neuen Wissens in der Organisation zu testen und gegebenenfalls auch herzustellen. Eine fehlende oder nicht erkannte Anschlussfähigkeit kann vor allem bei neuen Wissensgebieten und Konzepten (z. B. Rezeption und Interpretation von Daten aus Reformimpulsen oder empirischen Studien, etc.) eine Barriere sein; die Anschlussfähigkeit der Wissensbestände kann durch Austausch über gemeinsame Ziel- und Wertvorstellung (wieder)hergestellt werden. Ohne gemeinsame Ziel- und Wertvorstellungen ist der Austausch von Wissen erschwert. Wie zu Beginn des Absatzes in dem Zitat von Marks und Louis (1999) skizziert, eignen sich für den Austausch und die Weiterentwicklung von Wissen Professionelle Lerngemeinschaften besonders. Durch folgende Merkmale unterscheiden sie sich von ihrem Anspruch her von anderen Teams:

- Durch einen reflektierten Dialog der Kolleginnen und Kollegen untereinander;
- einen offenen Austausch über die Unterrichtspraxis;
- die Schaffung einer gemeinsamen Wissensbasis zur Verbesserung des Unterrichts;
- eine Zusammenarbeit bei der Entwicklung neuer Materialien und Curricula;
- eine professionelle Kultur, bestehend aus gemeinsamen Normen der pädagogischen Praxis
- und einem Fokus auf das Lernen von Schülerinnen und Schülern (vgl. Louis und Marks 1998).

Durch Veränderungen und Wandel, vor allem wenn diese tiefergreifend und radikaler sind, wie das double-loop learning (vgl. Argyris und Schön 1978), entsteht in Organisationen ein Ungleichgewicht in Form von Diskontinuitäten und Unvorhersehbarkeiten. „Disequilibrium is a necessary part of any transformative process“ (Louis und Leithwood 2000, p. 277). Professionelle Lerngemeinschaften liefern durch ihre Merkmale eine Stabilität in Form dauerhafter Beziehungen der Mitglieder untereinander sowie beständiger Normen, Werte und Routinen. Diese Routinen und Normen sind selbst auf Veränderungen und die professionelle Entwicklung der einzelnen Lehrkräfte und der Schule als Ganzes ausgerichtet. Somit erzeugen sie eine Stabilität in der Veränderung.

(3) Wissen und Fertigkeiten

Die Verbreitung und Weiterentwicklung von Wissen und Fertigkeiten hat für die Kapazität Organisationalen Lernens eine hohe Bedeutung. Schulen stehen drei verschiedene prototypische Quellen des Wissens zur Verfügung (vgl. Huber 1991; Kruse und Louis 2000). Zunächst das individuelle Wissen, das jedes einzelne Organisationsmitglied aufgrund seiner Erfahrung und Ausbildung mit sich bringt.

Innerhalb des Kollegiums sind die Wissensbestände – hinsichtlich (Fach)Didaktik, Erziehung und Schulverständnis – in der Regel disparat. Oftmals ist Schulen nicht genügend bekannt, über welches Wissen die einzelnen Mitglieder verfügen und inwiefern es für die Organisation relevant sein könnte (vgl. Kruse und Louis 2000). Hanson (2001) verweist im Kontext des organisationalen Gedächtnisses (vgl. Hedberg 1981) auf die Bedeutung des individuellen Wissens der Organisationsmitglieder für das Organisationale Lernen. Die Qualität des organisationalen Gedächtnisses hängt seiner Meinung nach entscheidend von der Qualität des intellektuellen Kapitals ab, das aus dem kumulierten Wissen der Organisationsmitglieder besteht. Im Rahmen schulischer Personalentwicklung gilt es, erstens schulweit das Wissen der einzelnen Lehrkräfte möglichst umfassend zu erfassen und zweitens in Verbindung mit den gemeinsamen Ziel- und Wertvorstellungen eine Kultur des gegenseitigen Voneinander-Lernens zu schaffen. Meetz (2007) stellt jedoch in einer Studie fest, dass Personalentwicklung in Schulen bisher noch sehr gering ausgeprägt ist.

Die zweite Wissensquelle ist Wissen aus der schulischen Umwelt, sei es von Expertinnen und Experten, anderen Schulen oder Reformimpulsen. In Bezug auf externes Wissen verweisen Kruse und Louis (2000) auf den oft fehlenden direkten Zugang von Lehrkräften zu externen Wissensquellen. Die schulische Personalentwicklung beschränkt sich meist auf den Verweis auf extern angebotene Fortbildungsangebote für einzelne Lehrkräfte oder sporadische schulinterne Fortbildung (vgl. Meetz 2007). Dagegen existiert in Schulen selten eine systematische Fortbildungsplanung, die die Interessen der einzelnen Lehrkraft und der Schule als Ganzes in den Blick nimmt (vgl. Meetz 2007). Schulen mit solch einem Konzept entwickeln auch Strategien zur Dissemination der vermittelten Inhalte in das Kollegium. Kruse und Louis (2000) berichten von Schulen, in denen Lehrkräfte, die an spezifischen Fortbildungen teilnehmen dürfen, im Gegenzug als Multiplikatorinnen und Multiplikatoren die neuen Inhalte und Methoden an das Kollegium vermitteln.

Sind diese ersten beiden Quellen des Wissens Voraussetzung für die Weiterentwicklung des Wissens in Teams, so ist die dritte Quelle ein Ergebnis von Teamarbeit. Es handelt sich um Wissen, das sich Schulen aneignen, indem sie spezifische Probleme ihres Schulalltags bearbeiten und lösen. Dieses Wissen gilt es der Schule als Ganzes zur Verfügung zu stellen, um es in ähnlichen Situationen anwenden zu können. Ein solcher Austausch ist von der Durchlässigkeit des Wissens abhängig (vgl. Kruse und Louis 2000), d. h. von der Offenheit des Kollegiums für neues Wissen sowie die Bereitschaft, bestehendes Wissen zu hinterfragen.

(4) Führung und Management

Die Schulleitung kann Organisationales Lernen fördern, aber auch verhindern. Hierbei hat sie nach Marks und Louis (1999) auch das Lernen selbst im Blick. Um es zu fördern, kann die Schulleitung distributive, transformationale und unterrichtsbezogene Führungselemente kombinieren (vgl. Hallinger 2003; Hallinger und Heck 2010). Die distributive Führung beinhaltet neben einer Beteiligung des Kollegiums bei wichtigen Entscheidungen, die die Schule als Ganze betreffen, eine dezentrale Führung, die die Verantwortung bei anstehenden Reformprojekten auf einzelne Teams überträgt, diese unterstützt und motiviert (vgl. Leithwood et al. 1994; Murphy und Louis 1994). Hier wird eine enge Verbindung zum Qualitätsbereich „Partizipation" (siehe (7)) deutlich. Des Weiteren zeigen viele Studien, dass eine transformationale Führung der Schulleitung das Organisationale Lernen in den anderen Dimensionen fördern kann (vgl. Feldhoff 2011; Feldhoff und Rolff 2008; Larson-Knight 2000; Leithwood et al. 1994; Leithwood, Jantzi et al. 2000; Mulford und Silins 2003; Silins et al. 2000). Empirische Befunde zeigen, dass der Führungsstil „highly control oriented and narrowly focused on the core technology of curriculum and instruction" (Leithwood, Leonard et al. 2000, S. 122). Diese Befunde werden auch von Robinson, Lloyd und Rowe (2008) bestätigt, die in ihrer Studie den Einfluss von Führung auf das Lernen von Schülerinnen und Schülern untersucht haben. Dabei zeigen sie, dass eine Kombination von transformationaler und unterrichtsbezogener Führung den größten Effekt auf das Lernen der Schülerinnen und Schüler hat. Gerade die unterrichtsbezogene Führung scheint allerdings bei Schulleitungen an deutschen Schulen nicht so stark ausgeprägt zu sein (vgl. Feldhoff 2011; Feldhoff und Rolff 2008).

Neben der Schulleitung kann die Steuergruppe das Organisationale Lernen als Change Agent schulischer Entwicklungsprozesse fördern (vgl. Dalin und Rolff 1990; Feldhoff 2011; Feldhoff und Rolff 2008; Holtappels 2007). Sie kann auch die Kapazität Organisationalen Lernens in den anderen Dimensionen positiv beeinflussen (vgl. Feldhoff 2011; Feldhoff und Rolff 2008) und vermittelt über diese Dimensionen auf die Nachhaltigkeit von Schulentwicklungsprozessen und die Qualität von Unterricht wirken. Im Unterschied zur Schulleitung ist die Steuergruppe nicht in die formale Hierarchie der Schule eingebunden. Während die Schulleitung primär im Modus von Führung agiert, liegen die Aufgaben der Steuergruppe vornehmlich in Aushandlung, Partizipation, Beratung und Unterstützung (vgl. Feldhoff 2011). Ähnliche Funktionen können auch Formen erweiterter Schulleitung übernehmen. Aufgrund der Einbindung in die formale Hierarchie der Schule und mitunter geringerer Legitimierung durch die Schul- oder Lehrerkonferenz könnte ihre Akzeptanz im Kollegium jedoch geringer sein.

(5) Qualitätssicherung, Zielüberprüfung und Feedback

Qualitätssicherung, Zielüberprüfung und Feedback an den Schulen bilden die fünfte Dimension ihrer Kapazität Organisationalem Lernens. Eine sinnvolle Übernahme von Verantwortung setzt die Durchführung von Evaluation und den Einsatz von Qualitätsmanagement und -sicherungssystemen in Schulen voraus. Organisationen benötigen möglichst genaue und umfassende Informationen über ihre Leistungsfähigkeit und den Lernprozess (vgl. Argyris und Schön 1978; Daft und Huber 1987; Duncan und Weiss 1979; Hedberg 1981; Kruse und Louis 2000; March und Olsen 1976). Dafür benötigen sie zudem klare und eindeutige Indikatoren, an denen sie ihre eigene Leistung messen können (vgl. Stringfield 2000). Marks und Louis (1999) drücken es pointiert aus: ohne Feedback, Evaluation und klare Zielkriterien ist Organisationales Lernen defizitär. Es geht darum, wie Schulen dafür sorgen, dass sie erstens notwendige Informationen über die Ergebnisse ihres unterrichtlichen und schulischen Handelns erhalten und zweitens die Informationen interpretieren und für ihre professionelle Entwicklung im Sinne einer lernenden Organisation nutzen. Dafür benötigen sie einen Entfaltungs- und Gestaltungsraum (vgl. Feldhoff 2011; Feldhoff und Rolff 2008).

„Highly reliable organizations must constantly rely on the professional judgements of all their team members" (Stringfield 2000, S. 269). Nach Stringfield überprüfen Lehrpersonen an solchen Schulen regelmäßig Lernfortschritte der Schülerinnen und Schüler. Zudem spielt regelmäßiges Schulmonitoring nach Stringfield eine große Rolle. Marks und Louis (1999) zufolge ist es entscheidend, dass sowohl Leistungsindikatoren, als auch Anreize vom Kollegium mitgetragen werden (vgl. Kruse und Louis 2000).

Schulen, die über eine hohe Fähigkeit Organisationalen Lernens verfügen, entwickeln zusätzlich eigene Standards, während Schulen, die über eine geringe Kapazität des Organisationalen Lernens verfügen, sich stark an externen Standards orientieren (vgl. Newmann et al. 1997). Die Interpretation von Daten anhand schuleigener Standards dient wiederum als Grundlage für neue Entscheidungen und kann über entsprechende Transformationsprozesse Teil des organisationalen Wissens werden (vgl. Duncan und Weiss 1979; Hedberg 1981; Huber 1991). Dabei steht vor allem das Lernen der Schülerinnen und Schüler im Fokus. Durch interdisziplinäre Teams kann bei der Entwicklung solcher Standards auch die Auseinandersetzung über gemeinsame Normen und Ziele gefördert werden (vgl. Kruse und Louis 2000).

(6) Austausch mit der schulischen Umwelt

Eine weitere Komponente der Kapazität Organisationalen Lernens bildet der Austausch der Schule mit ihrer Umwelt. Durch verschiedene Formen Organisati-

onalen Lernens soll eine Passung zur sich wandelnden Umwelt hergestellt werden (vgl. Duncan und Weiss 1979; Hedberg 1981; Huber 1991; March und Olsen 1976). Das interne und externe Feedback kann wichtige Informationen liefern. Der Austausch mit der schulischen Umwelt geht über ein solches Feedback hinaus. Über einen gezielten Austausch mit Eltern, anderen Schulen oder Einrichtungen kann die Schule Informationen über sich erhalten, neues Wissen generieren, aber auch ihre Umwelt mitgestalten. Dies ist eng verbunden mit einem aktiven Scannen der Umwelt, um auf dortige Veränderungen adäquat reagieren zu können (vgl. Cousins 2000; Hedberg 1981; Huber 1991; Rait 1995). Nach Cousins (2000) sind Strategien der Umweltbeobachtung: Rezeption von Fachzeitschriften, Teilnahme an Fachtagungen, Verfolgung aktueller bildungspolitischer Debatten, Austausch mit anderen Schulen und eine Vernetzung auf regionaler Ebene. Die von Kruse und Louis (2000) beschriebenen Schulen (eine „Elementary" und eine „Middle school") nutzen derartige Maßnahmen: Die Grundschule kooperiert auf Distriktebene mit anderen Schulen im Bereich der Leseförderung. In der „Middle school" nehmen die Lehrkräfte regelmäßig an nationalen Treffen teil und sind über externe Netzwerke mit ihrer schulischen Umwelt verknüpft. Auch kann die schulische Umwelt in Gestalt der Kommune bzw. des Schulträgers oder ähnlicher Instanzen das Organisationale Lernen der Schule fördern (vgl. Leithwood, Jantzi et al. 2000; Silins et al., 2000). Gemeinsame Ziele und Visionen auf Ebene des Schulträgers können Schulen zum Lernen anregen (vgl. Leithwood, Jantzi et al. 2000), wenn die Schulen an deren Entwicklung beteiligt sind und sie sich mit diesen identifizieren können. Zudem kann ein regionales Informationssystem die Suche der Schulen nach relevanten Inhalten unterstützen. Durch den Austausch einer Schule mit ihrer Umwelt kann sich aber nicht nur die Schule selbst verändern, sie kann auch selbst proaktiv Einfluss auf ihre schulische Umwelt nehmen (vgl. Bormann 2001; Kruse und Louis 2000). Kruse und Louis verdeutlichen dies: „For example, a growing number of teachers have expanded their role to include teaching lower-income parents how to teach their own children" (Kruse und Louis 2000, S. 25). Zudem kann ein regionales Informationssystem die Suche der Schulen nach relevanten Inhalten unterstützen.

(7) Partizipation der Lehrkräfte

Die Partizipation der Lehrkräfte ist eine Schlüsseldimension der Kapazität Organisationalen Lernens, die in gewisser Weise Bedingung und Ergebnis des Lernens in den anderen Dimensionen ist. Da Organisationales Lernen nur über das Lernen der Organisationsmitglieder in Teams erfolgen kann (vgl. Wiegand 1998), kann dies auch nur gelingen, wenn die Mehrheit der Mitglieder an diesem Prozess beteiligt ist. Eine Beteiligung ist auch für die Akzeptanz von Veränderungen und die Übernahme von Verantwortung notwendig, wie die Forschung im Kontext der

Organisationsentwicklung und des Change Management zeigt (vgl. Lewin 1963; Schubert 2004). Ein Aspekt von Lehrerpartizipation bezieht sich auf die Mitwirkung der Lehrkräfte bei schulweiten Entscheidungen: Mitwirkung an Entscheidungen, die die Lehrkräfte selbst unmittelbar betreffen und Mitwirkung an Entscheidungen in Bezug auf Unterricht und das Lernen der Schülerinnen und Schüler; ein Aspekt, der sich mit Marks und Louis (1999) mit „Teacher empowerment" in Verbindung bringen lässt. Die Bedeutung von Partizipation heben Brown et al. (1999) in ihrer Studie mit 21 Sekundarschulen im Nordwesten von England und Wales hervor. Die Schulen wurden in Typen eingeteilt. Der Typ, in dem die Lehrkräfte von einer hohen Arbeitszufriedenheit und Motivation sprechen, ist neben anderen Faktoren des Organisationalen Lernens auch durch eine breite Partizipation der Lehrkräfte gekennzeichnet. Silins et al. (2000) kommen in ihrer Studie zu dem Ergebnis, dass eine Partizipation der Lehrkräfte gemessen am Grad der Mitwirkung und Entscheidung, einen, wenn auch indirekten, Einfluss auf Organisationales Lernen hat, vermittelt über Wertschätzung und aktives Engagement der Lehrkräfte sowie die Schulautonomie. Dieser Befund bestätigt auch die These von Marks und Louis (1999), dass ein Mehr an Mitbestimmungs- und Mitwirkungsmöglichkeiten der Lehrkräfte auch zu einem höheren Engagement der Lehrkräfte für schulweite Belange führt.

Weiter wird angenommen, dass eine Partizipation der Lehrkräfte auf kommunaler Ebene und Schulebene zu einer höheren Problemlösefähigkeit des Kollegiums führt und die Lehrkräfte stärker an den Bedürfnissen der Schule interessiert sind (vgl. Leithwood, Leonard et al. 2000). Marks et al.(2000) folgern mit Bezug auf Rait (1995), dass Lehrkräfte, die sich auf Schulebene für Weiterentwicklung und Reflexion schulischer Belange engagieren, dies auch auf ihren Unterricht übertragen.

Insgesamt hängen die genannten Kapazitätsdimensionen (KD) eng zusammen. So beziehen sich zum Beispiel Standards und Kriterien auf schulische Ziele und sind somit komplementär zu gemeinsamen Erwartungen und Zielen. Darüber hinaus wird durch Evaluation gezielt neues Wissen über die Leistungsfähigkeit der Organisation geschaffen, was eine enge Beziehung zum Qualitätsbereich „Wissen und Fertigkeiten" darstellt. Eine weitere Interdependenz besteht hinsichtlich der Verantwortung; sie ist Teil professioneller Gemeinschaften, ebenso Teil von Partizipation und Mitbestimmung. Zudem spiegelt sich eine gelebte schulische Partizipation auch in entsprechenden Strukturen (vgl. Leithwood, Jantzi et al. 2000) wieder. Weiter funktioniert schulische Mitwirkung von Lehrkräften nicht ohne eine Schulleitung, wobei distributive und transformationale Führung die Beteiligung der Lehrkräfte fördert (vgl. Larson-Knight 2000; Leithwood, Jantzi et al. 2000; Leithwood, Leonard et al. 2000; Silins et al. 2000). Da schulische Partizipation und Mitbestimmung kollektive Prozesse sind, berühren sie eng das gemeinsame Verständnis des Kol-

legiums von Schule und Kooperationen (vgl. Hanson 2001; Larson-Knight 2000; Leithwood, Leonard et al. 2000; Louis und Leithwood 2000). Es gibt viele weitere Zusammenhänge zwischen den KD, die jedoch an dieser Stelle nicht vertieft werden. Zusammengefasst fokussieren die Dimensionen auf verschiedene Bereiche, innerhalb derer Schulen aktiv werden, wenn es um Schulentwicklung geht.

4 Forschungsdesiderat und Fragestellung

Im Folgenden soll nun das Analysepotential des Kapazitäten-Ansatzes für die Verarbeitungsprozesse im Anschluss an die Schulinspektion verdeutlicht werden (vgl. hierzu auch Feldhoff et al. 2014). An dieser Stelle ist nur eine idealtypische Darstellung möglich (vgl. Weber 1972), die zwei Pole eines Kontinuums aufzeigt. Empirisch werden sich Schulen auf diesem Kontinuum zwischen den beiden Polen verorten. Für die idealtypische Darstellung der verschiedenen Reaktionen von Schulen wird ergänzend zu den Dimensionen der Kapazität Organisationalen Lernens, das Konzept der Handlungstheorie von Argyris und Schön (1978) genutzt. Ihre zwei Handlungsmodi eignen sich ideal, um die beiden idealtypischen Pole des Kontinuums zu beschreiben.

Nach Argyris und Schön verfügt jedes Organisationsmitglied über spezifische handlungsleitende „naive" Theorien „theory-in-use" oder auch Gebrauchstheorien. Die Gebrauchstheorien macht sich ein Individuum zur Bewältigung von Situationen als eine Art „naiver" Wissenschaftler zu Eigen. Hierbei formuliert es Hypothesen über Zusammenhänge von Handlungen und Faktoren und ordnet diesen dazu passende Handlungsmuster zu. Diese Hypothesen basieren auf grundlegende Normen und Werten. Die Gebrauchstheorien repräsentieren sein Bild von der Organisation und leitet sein Handeln in der Organisation (1978). Die Summe der einzelnen Gebrauchstheorien („theory-in-use") der Organisationsmitglieder bzw. deren gemeinsame Schnittmenge bildet die „organizational theory-in-use". Die „organizational theory-in-use" kann als eine Form „organisationalen Gedächtnisses" bezeichnet werden. Sie verändert sich ständig durch die Interaktion der Organisationsmitglieder und die individuelle Überprüfung und Modifizierung der eigenen Gebrauchstheorie („theorie-in-use"). Sie kann nur durch die Beobachtung der Handlung der Organisationsmitglieder rekursiv konstruiert werden. Diese Gebrauchstheorie („theory-in-use") ist Ausgangspunkt und Ergebnis von Organisationalem Lernen (ebd.). Sie dient dem Erwerb und der Anwendung von Wissen sowie der Herausbildung von Motivation (vgl. Geißler 1995). Neben der Gebrauchstheorie („theory-in-use") existiert immer ein „espoused theorie" oder auch

verlautbarte Theorie. Dies ist die Theorie über Zusammenhänge von Handlungen und Faktoren sowie die dazugehörigen eigenen Handlungsmuster, die ein Individuum nach Außen kommuniziert. Diese verlautbare Theorie muss nach Argyris und Schön nicht zwingend handlungsleitend sein. Sie ist dann handlungsleitend, wenn sie mit der Gebrauchstheorie („theory-in-use") kongruent ist. Weicht sie deutlich von dieser ab und ist somit nicht Handlungsleitend, wird Organisationales Lernen erschwert. Die Organisationsmitglieder selbst gehen i. d. R. davon aus, dass ihre verlautbarten Theorien mit ihren handlungsleitenden kongruent sind.

Um organisationale Lernprozesse und deren Hindernisse zu beschreiben differenzieren Argyris und Schön zwei Handlungsmodi (Modus I und II). Im Modus I sind das Handeln der Organisation und deren Mitglieder primär durch defensive Handlungsroutinen geprägt. In dem Handlungsmodus I kann Lernen nur innerhalb der grundlegende Normen und Werten (single-loop-learning) stattfinden. Das heißt ein Feedback von außen wird nur als sinnvoll erlebt, sofern es zu den grundlegenden Normen und Werten passt. Diese sind im Modus I vor allem auf Erhaltung des Status Quo ausgerichtet. Im Modus I gibt es bei Argyris und Schön eine starke Diskrepanz zwischen der verlautbarten Theorie („espoused theorie") und der Gebrauchstheorie („theorie in use"). Das heißt, das Handeln der Akteure weicht mehr oder weniger stark von dem ab, was nach „außen" verlautbart wird. Im Modus II ist diese Diskrepanz sehr gering und die Organisation ist zu höherwertigem Lernen in der Lage; der Reflektion der eigenen Normen und Werte (double-loop-learning) und zur Reflektion der eigenen Lernprozesse (deutero learning).

Eine idealtypische Darstellung, wie die folgende, darf natürlich nicht als normatives Ideal verstanden werden. Sicherlich sind in einigen Aspekten auch andere Reaktionsweisen von Schulen möglich, die in dieser idealtypischen Darstellung nicht zur Geltung kommen.

Bei Schulen im Handlungmodus II, die über eine hohe Kapazität verfügen, ist zunächst davon auszugehen, dass sie die Fähigkeit besitzen, ihre eigenes Handeln und Wissen sowie ihre Normen und Werte kritisch zu hinterfragen und neuen Sichtweisen und Wissen gegenüber offen sind (KD 3). Ihre „theorie in use" und „espoused theorie" sind relativ deckungsgleich. Solche Schulen diskutieren die Rückmeldung zunächst in den zuständigen Gremien, z. B. Steuergruppen oder erweiterte Schulleitung (KD 4). Die Rückmeldung wird zu dem Selbstbericht der Schule, den Ergebnissen der internen Evaluation (KD 5), den eigenen Normen und Werten, schulischen Zielen und Standards (KD 2) sowie den spezifischen Kontextbedingungen der Schule (KD 6) in Beziehung gesetzt. Dabei nutzen sie auch die Fähigkeit im Bereich Evaluation, die in den Rückmeldungen enthalten Befunde richtig zu lesen und angemessen zu interpretieren (KD 5). Anschließend werden die Ergebnisse schulweit im Kollegium und der Schulkonferenz kommuniziert,

rekontextualisiert und erneut diskutiert (KD 7). Danach wird eine Priorisierung vorgenommen und es beginnt ein kollektiver Prozess der Problemlösungssuche (KD 3). Zu diesem Zwecke werden ggf. spezielle Arbeitsgruppen eingesetzt (KD 1). Hierbei werden externe Quellen genutzt und ein systematisches Scannen der schulischen Umwelt in Gang gesetzt (KD 6). Anschließend werden die Ergebnisse der Problemlösungssuche im Kollegium präsentiert und entsprechende Maßnahmen verabschiedet (KD 7). Die Steuergruppe erhält ein Mandat einen entsprechenden Schulentwicklungsprozess in Gang zu setzen (KD 4). Die Schulleitung vereinbart mit der Schulaufsicht Zielvereinbarungen, zeigt die in der Schule beschlossenen Maßnahmen auf, und fordert ggf. Unterstützung in Form von Beratung oder Fortbildungen (KD 4, 3). In der Schule werden Arbeitsgruppen gebildet, die die Maßnahmen umsetzen und deren Arbeit von der Steuergruppe unterstützt und koordiniert wird (KD 1, 2, 4). Nach erfolgreicher Umsetzung werden die Maßnahmen dann schließlich evaluiert (KD 5).

Bei Schulen im Handlungsmodus I, die über eine geringe Kapazität verfügen, ist davon auszugehen, dass sie „lediglich" in der Lage sind ihr eigenes Handeln innerhalb ihrer Normen und Werte kritisch zu hinterfragen und neuen Sichtweisen und Wissen kritisch oder gar verschlossen gegenüber stehen (KD 2, 3). In diesen Schulen besteht eine größere Diskrepanz zwischen „theorie in use" und „espoused theorie". Die Schulen sind überwiegend von defensive Routinen, Normen und Werten geprägt, die auf den Erhalt den Status Quo und nicht auf eine professionelle Weiterentwicklung ausgerichtet sind. Zudem sind die Normen und Werte im Kollegium disparat und diffus. Die Schulen besitzen nur rudimentäre Kenntnisse die Befunde in den Rückmeldungen zu lesen und angemessen zu interpretieren (KD 5). Die Auseinandersetzung mit dem Inspektionsbericht kann sich auf eine externe Attribuierung oder der Isolierung einzelner Aussagen beschränken. Der Schule fällt es schwer die Befunde zu rekontextualisieren (KD 6); sie werden in der Schule kaum diskutiert, die Schulleitung als verantwortlich ausgeflaggt (KD 4). Sie versucht Zielvereinbarungen mit der Schulaufsicht vage zu halten oder eher oberflächliche Maßnahmen anzukündigen, mit deren Hilfe Aktivität demonstriert werden kann, ohne den Status Quo wesentlich zu verändern. Besteht in Teilen des Kollegiums der Wunsch zur Veränderung so fehlen oftmals entsprechende Strukturen (z. B. eine Steuergruppe, institutionalisierte Teams; KD 1, 2), das nötige spezifische Wissen sowie allgemeines Wissen über Schulentwicklungsprozesse mit Hilfe welche strategischer Maßnahmen (KD 3, 6) Defizite behoben werden können. Zudem fehlt die Unterstützung im Kollegium (KD 7) und/oder der Schulleitung (KD 4).

Auf Basis dieser idealtypischen Darstellung lässt sich folgenden Arbeitshypothese formulieren: Es ist davon auszugehen, dass Schulen die über geringe Kapazitäten Organisationalen Lernens verfügen, nur in sehr begrenztem Umfang in der Lage

sind die Rückmeldung so zu verarbeiten, dass daraus sichtbare Impulse für die Schul- und Unterrichtsentwicklung resultieren. Dagegen ist davon auszugehen, dass bei Schulen, die über hohe Kapazitäten Organisationalen Lernens verfügen, die Rückmeldung gut für die Schul- und Unterrichtsentwicklung ist. Bisher ungeklärt ist, wie hoch die Fähigkeit der Schule in einzelnen Kapazitätsbereichen sein muss, damit diese Produktiv genutzt werden und ob Schulen mit besseren innerschulischen Voraussetzungen ggf. mehr und/ oder andere Maßnahmen ergreifen als Schulen mit weniger guten Voraussetzungen für die Schulentwicklung.

5 Evaluation von Inspektionswirkungen am Beispiel der Schulinspektion Hamburg

Überprüft wird diese Annahme im Folgenden anhand der Schulinspektion Hamburg, die seit dem Jahr 2007 jährlich bis zu 80 Schulen evaluiert, 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 wurden im ersten Zyklus der Inspektion (2007-2013) nicht veröffentlicht und nur der Schulöffentlichkeit zur Verfügung gestellt, im zweiten Zyklus der Inspektion (seit 2014) werden der Öffentlichkeit Kurzberichte zum Download im Internet bereit 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 (vgl. Diedrich 2015a).

Im Folgenden wird zunächst das Rückmeldeverfahren der Hamburger Schulinspektion erläutert (5.1). Dem schließt sich ein Vergleich der Qualitätsbereiche des Hamburger Orientierungsrahmens als Grundlage für die Bewertung der Schulinspektion und den Dimensionen der Kapazitäten an. Es soll gezeigt werden, in welchen Bereichen und welchem Ausmaß sich die Kapazitäten in den Qualitätsbereichen widerspiegeln und ob die Einschätzung der Schulinspektion in den Qualitätsbereichen als Indikator für die Kapazitäten dienen kann (5.2). Sodann werden die Untersuchung selbst, das methodische Vorgehen und die Stichprobe erläutert (5.3). Zum Abschluss werden die Befunde präsentiert (5.4).

5.1 Das Rückmeldeverfahren der Schulinspektion Hamburg

Das Rückmeldeverfahren der Schulinspektion Hamburg bestand im ersten Inspektionszyklus aus sechs Elementen (vgl. Diedrich 2015a; Pietsch et al. 2014): (a) einem Feedbackgespräch zwischen Inspektionsteam und Schulleitung am letzten Tag des Schulbesuches, (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. Die Schulaufsicht führte spätestens zwölf Wochen nach der Ergebnisrückmeldung ein sogenanntes Responsegespräch mit der Schulleitung. Hier tauschte sie sich mit der Schulleitung über ihre Einschätzungen und Bewertungen der Inspektionsergebnisse aus und schloss verbindliche Vereinbarungen über die abzuleitenden Entwicklungsziele und -maßnahmen. Nach Möglichkeit sollten diese in eine neue Ziel- und Leistungsvereinbarung münden. Darüber hinaus führten die Schulaufsichten ein Jahr nach Abschluss der Schulinspektion mit der Schule ein sogenanntes Bilanzgespräch, in dem die wichtigsten Entwicklungen und weiterer Veränderungsbedarf evaluiert und reflektiert wurden.

5.2 Bestimmung schulischer Kapazitäten

Analysegegenstand der nachfolgenden Untersuchung ist die Schulinspektion im ersten Zyklus. Der damalige Bericht, welchen die Schule nach Abschluss der Inspektion erhielt, basiert auf der durch das Inspektionsteam vorgenommenen Bewertung von 14 Qualitätsmerkmalen (vgl. Tab. 1). Diese 14 Qualitätsbereiche waren im ersten Inspektionszyklus aller Hamburger Schulen inhaltlich im bis zum Jahr 2012 gültigen Orientierungsrahmen für Hamburger Schulen (vgl. Behörde für Schule und Sport 2006) verankert. Der Orientierungsrahmen Schulqualität basiert dabei auf verschiedenen wissenschaftlichen Studien über die Effektivität von Schule und Unterricht (vgl. z. B. Ehren und Scheerens 2015).

Tab. 1 Qualitätsdimensionen und -bereiche des Hamburger Orientierungsrahmens Schulqualität

Qualitätsbereich		
1. Führung und Management	**2. Bildung und Erziehung**	**3. Wirkungen und Ergebnisse**
1.1 Führung wahrnehmen	2.1 Das schuleigene Curriculum gestalten	3.1 Zufriedenes Personal, zufriedene Schüler/innen, Eltern und Betriebe
1.2 Personal entwickeln	2.2 Unterrichten, Lernen, Erziehen	3.2 Bildungslaufbahnen und Kompetenzen
1.3 Finanz- und Sachmittel gezielt einsetzen	2.3 Organisatorische Rahmenbedingungen sichern	
1.4 Profil entwickeln und Rechenschaft ablegen	2.4 Leistungen beurteilen	
	2.5 Prozesse und Ergebnisse evaluieren	
	2.6 Förderkonzepte entwickeln	
	2.7 Beratungskonzepte gestalten	
	2.8 Die Schulgemeinschaft beteiligen	

Die Merkmale und Indikatoren einiger Qualitätsbereiche des Orientierungsrahmens lassen sich verschiedenen Aspekten einzelner Kapazitäten-Dimensionen (KD) zuordnen. Um den Rahmen des Beitrags nicht zu sprengen, werden die Parallelen nur auf der Ebene von Qualitätsbereichen und -dimensionen aufgezeigt. Die Mehrzahl der Indikatoren des Qualitätsbereichs „Führung wahrnehmen" beziehen sich auf das Schulleitungshandeln (KD 4). Einige Indikatoren lassen sich darüber hinaus den Aspekten der Normen, Wertvorstellungen und Zielen (KD 4), der Personalentwicklung (KD 3), der internen Evaluation und Qualitätssicherung (KD 5) sowie der Vernetzung mit der schulischen Umwelt (KD 6) zuordnen. Die Indikatoren des Qualitätsbereichs „Personal entwickeln" beziehen sich auf den Bereich der Nutzung und Weiterentwicklung des Wissens in Form von Personalentwicklung (KD 3). Jedoch lassen sich manche Indikatoren auch organisationalen Regelungen und Strukturen (KD 1) sowie der Kooperation (KD 2) zuordnen. Die Indikatoren des Qualitätsbereichs „Finanz- und Sachmittel gezielt einsetzen" sind im Kapazitätenansatz beim Schulleitungshandeln (KD 4) angesiedelt. Die Indi-

katoren des Qualitätsbereichs „Profil entwickeln und Rechenschaft" lassen sich den Aspekten Ziele, Normen und Werte (KD 2), Austausch mit der schulischen Umwelt (KD 6), interne Evaluation als Instrument der Rechenschaftslegung (KD 5) zuordnen. Die Indikatoren des Qualitätsbereichs „Organisatorische Rahmenbedingungen sichern" sind im Bereich organisationale Strukturen und Regelungen (KD 1) angesiedelt. Indikatoren des Qualitätsbereichs „Prozesse und Ergebnisse evaluieren" beziehen sich auf die Nutzung interner und externer Evaluation für die Unterrichtsentwicklung (KD 5). Zusammenfassend kann festgehalten werden, dass sich der Kern der verschiedenen Kapazitätsdimensionen im Qualitätsbereich „Führung und Management" sowie den Bereichen „Organisatorische Rahmenbedingungen sichern" und „Prozesse und Ergebnisse evaluieren" wieder finden lassen. Somit kann die Einschätzung der Schulen in diesen Bereichen durch die Schulinspektion als geeigneter Indikator für die Kapazität des Organisationalen Lernens der Schulen dienen.

Bei den anderen Qualitätsbereichen werden teilweise Interdependenzen deutlich zum Beispiel bei „Unterrichten, Lernen, Erziehen" zur Aufgabe Professioneller Lerngemeinschaften (KD 2). Jedoch beziehen sich die Indikatoren auf Aspekte, die über den Kapazitätenansatz hinausgehen und sich primär auf die Unterrichtsprozesse, die Entwicklung und Unterstützung (QD 2.1, QD 2.2, QD 2.4, QD 2.6) sowie deren Wirkungen und Ergebnisse (3.) beziehen.

5.3 Anlage und Durchführung der Untersuchung

5.3.1 Datengrundlage

Schulische Kapazitäten

Wie bereits oben erläutert, bieten Informationen aus Inspektionen – sofern diese sich auf dieselben theoretischen und empirischen Grundlagen beziehen wie der Ansatz der Kapazitäten – die Möglichkeit, als relativ robuster Indikator für Kapazitäten Organisationalen Lernens zu fungieren. Daher liegen den nachfolgenden Analysen die Qualitätsberichte von Schulen aus dem ersten Zyklus zugrunde. Für diese Berichte wurden einzelne Indikatoren, die sich auf den Hamburger Orientierungsrahmen Schulqualität beziehen, auf einer vierstufigen Bewertungsskala (1 = deutlich mehr Schwächen als Stärken, 2 = eher mehr Schwächen als Stärken, 3 = eher mehr Stärken als Schwächen, 4 = deutlich mehr Stärken als Schwächen) bewertet, um anschließend durch Aggregation zu einer Bewertung der jeweiligen Qualitätsbereiche zu gelangen. Die Indikatoren wiederum wurden mit empirischen Daten unterfüttert, die mithilfe von Onlinebefragungen und teilstandardisierten

Interviews, Dokumentenanalysen sowie systematischen Unterrichtsbeobachtungen erhoben wurden.

Schulentwicklungsberichte

Die Schulinspektion Hamburg befindet sich seit dem Jahr 2013 in ihrem zweiten Erhebungszyklus (vgl. Diedrich 2015b). Im Rahmen einer so genannten Vorerhebung wird in diesem Zyklus seitens der Schulinspektion ein Entwicklungsbericht der inspizierten Schulen angefordert. In ihm legen Schulleiterinnen und Schulleiter dar, welche Veränderungen es seit der ersten Inspektion gegeben hat und welche Entwicklungsschritte bzw. Maßnahmen die Schulen eingeleitet haben sowie ob und wie diese umgesetzt wurden. Die Berichte sollen auf den zwischen Schule und Schulaufsichten getroffenen Ziel-Leistungs-Vereinbarungen basieren. Der Schulaufsicht obliegt es auch die Umsetzung der Maßnahmen zu kontrollieren. Die Schulinspektion erhält mit dem Bericht bereits vor ihrem Schulbesuch einen vertieften Einblick in die Entwicklung der Schule in den vergangenen Jahren aus Sicht der Schulleitung. Diese Entwicklungsberichte beinhalten in tabellarischer Form die angestrebten beziehungsweise durchgeführten Maßnahmen seit der letzten Inspektion, diesbezügliche Verantwortlichkeiten sowie konkrete Umsetzungstermine.

5.3.2 Methode

Um den Zusammenhang zwischen schulischen Kapazitäten und Entwicklungen mithilfe der oben dargestellten Daten zu analysieren bedarf es eines zweischrittigen Verfahrens. In einem ersten Schritt wurde eine quantitative Inhaltsanalyse der Schulentwicklungsberichte in Bezug auf die von ihr umgesetzten Entwicklungsmaßnahmen durchgeführt. Bei der quantitativen Inhaltsanalyse handelt es sich um ein Instrument der quantitativen Datenauswertung, die das „Zählen" bestimmter Aspekte zum Zweck der genaueren Analyse durchführt. Ziel der quantitativen Inhaltsanalyse ist es, „Wortmaterial hinsichtlich bestimmter Aspekte (stilistische, grammatische, inhaltliche, pragmatische Merkmale) zu quantifizieren" (Bortz und Döring, 1995, S. 138). Dabei werden einzelne Teile eines Textes größeren Einheiten oder Kategoriensystemen zugeordnet. Bei der Zuordnung der Textteile oder Worte gibt es zwei mögliche Alternativen. Entweder erfolgt die Zuordnung deduktiv zu bereits vorhandenen Kategoriensystemen oder es wird ein neues Kategoriensystem erarbeitet, was einem induktiven Vorgehen entspricht. In der hier durchgeführten Studie wurde deduktiv vorgegangen und die in den Entwicklungsberichten genannten Maßnahmen den Qualitätsbereichen des Hamburger Orientierungsrahmen Schulqualität zugeordnet. In einem zweiten Schritt wurden mithilfe dieser Daten

und der Informationen aus den Inspektionsberichten des ersten Zyklus Wahrscheinlichkeits- sowie Korrelationsstudien durchgeführt.

5.3.3 Stichprobe

Grundlage der nachfolgenden Analysen sind Informationen von 49 Schulen, die bis zum Frühjahr 2014 zum zweiten Mal durch die Schulinspektion Hamburg extern evaluiert wurden, wobei die ersten Inspektionen der Schulen in den Jahren 2007 und 2010 stattfanden. Die Stichprobe setzt sich aus 30 Grundschulen, 12 Gymnasien, sechs Stadtteilschulen und einer Sonderschule zusammen. Einige Grundschulen wurden im ersten Zyklus als Grund- Haupt und Realschulen geführt, sind jedoch infolge von Reformmaßnahmen in reine Grundschulen überführt worden, mussten also auch mit strukturellen Veränderungen umgehen.

5.4 Befunde

5.4.1 Stärken und Schwächen der Schulen

Zunächst wurde anhand der durch die Schulinspektion identifizierten Stärken und Schwächen der Schulen in den Qualitätsbereichen, die Kapazität des Organisationalen Lernens der Schulen identifiziert (vgl. Tab. 2). Die grau markierten Bereiche stellen diejenigen Qualitätsbereiche dar, die als Indikator für die Kapazität des Organisationalen Lernens dienen.

Dabei fällt auf, dass keine der inspizierten Schulen durchgehend in allen Qualitätsbereichen als schwach eingestuft wird. In den Bereichen „Personal entwickeln", „das schuleigene Curriculum gestalten", „Beratungsangebote gestalten" erhält nur jeweils eine Schule das Prädikat „schwach". In dem Bereich „Prozesse und Ergebnisse evaluieren" sind es fünf Schulen, im Bereich „Leistung beurteilen" vier Schulen.

Tab. 2 Von der Schulinspektion identifizierte Stärken und Schwächen in den 49 Schulen

Einschätzung durch die Schulinspektion	schwach	eher schwach	eher stark	stark
Qualitätsbereich				
1. Führung wahrnehmen	-	8	30	11
2. Personal entwickeln	1	22	22	4
3. Finanz- und Sachmittel gezielt einsetzen	-	11	27	11
4. Profil entwickeln und Rechenschaft ablegen	-	8	30	11
5. Das schuleigene Curriculum gestalten	1	30	18	-
6. Unterrichten, Lernen, Erziehen	-	13	35	1
7. Organisatorische Rahmenbedingungen sichern	-	3	36	10
8. Leistung beurteilen	4	31	14	-
9. Prozesse und Ergebnisse evaluieren	5	32	12	-
10. Förderkonzepte entwickeln	-	16	32	1
11. Beratungsangebote gestalten	1	8	34	6
12. Die Schulgemeinschaft beteiligen	-	3	33	6
13. Zufriedenes Personal, Schüler, Eltern, Betriebe	-	2	36	11

Acht der 49 Schulen, erhalten in 10 der 13 Bereiche das Prädikat „eher schwach". Betrachtet man die Verteilung der Schulen in Bezug auf die Bewertung durch die Schulinspektion in den einzelnen Bereichen, so erhalten mehr als die Hälfte der Schulen (32) im Bereich „Prozesse und Ergebnisse evaluieren", sowie etwas weniger als die Hälfte der Schulen (22) im Bereich „Personal entwickeln" das Prädikat „eher schwach". Ein ähnlicher Befund, findet sich auch in anderen Studien zur Kapazität Organisationalen Lernens (vgl. Feldhoff et al. 2008; Feldhoff, 2011). Aus inhaltlicher Perspektive ist der Befund nicht unproblematisch. Da die Rezeption der Inspektionsdaten in Form des Lesens und angemessen Interpretierens (Qualitätsbereich „Prozesse und Ergebnisse evaluieren") zentrale Voraussetzung für die

weiteren Verarbeitungsprozesse ist, um entsprechende Schulentwicklungsprozesse einzuleiten. Auch der Personalentwicklung kommt im Sinne der Informationsbeschaffung und zielgerichteten Nutzung von Fortbildungen eine Schlüsselrolle zu. Ebenfalls mehr als die Hälfte der Schulen ist „eher schwach" bei der Gestaltung des Schulcurriculums und der Leistungsbeurteilung. Diese Aspekte sind für die Qualität des Unterrichts sehr bedeutsam. In Bezug auf die Frage welche Voraussetzungen für schulische Verarbeitungsprozesse im Anschluss an die Inspektion notwendig sind, spielen sie jedoch keine Rolle. Sie sind nicht Teil dieser Prozesse, sondern die Veränderung dieser Aspekte soll stattdessen Ergebnisse erfolgreicher Verarbeitungsprozesse in Form von Schul- und Unterrichtsentwicklung sein.

Die Bewertungskategorie „eher stark" erhält etwas mehr als die Hälfte der Schulen in 10 von 13 Qualitätsbereichen. In den Bereichen „Prozesse und Ergebnisse evaluieren", „Das schuleigene Curriculum gestalten" und „Leistung beurteilen " erhalten nur rund ein Viertel der Schule ein solches Prädikat. Keine der Schulen erhält in allen Qualitätsbereichen das Prädikat „stark". 10 Schulen immerhin in fünf Bereichen.

Somit ist davon auszugehen, dass alle Schulen ein Mindestmaß an Kapazitäten besitzen, um die Rückmeldungen für Schulentwicklung zu nutzen. Bei der Hälfte der Schulen sind zumindest Probleme und/oder Hindernisse bei den Verarbeitungsprozessen zu erwarten. Gerade der Bereich der Evaluation scheint den Schulen insgesamt Schwierigkeiten zu bereiten.

5.4.2 Anzahl und Verteilung von Schulentwicklungsmaßnahmen im Anschluss an eine Inspektion

Nach der Identifikation der schulischen Kapazitäten wurden als nächstes die von den Schulen im Anschluss an die Inspektion initiierten Maßnahmen beleuchtet. Insgesamt setzten die 49 untersuchten Schulen im Zeitraum zwischen der ersten und der zweiten Inspektion 1066 Maßnahmen um. Dabei wurden im Mittel rund 22 Maßnahmen pro Schule umgesetzt, wobei die Spannbreite von minimal Null bis maximal 53 Schulentwicklungsmaßnahmen reicht. Die Maßnahmen bezogen sich vor allem auf die Bereiche „Förderkonzepte entwickeln", „das schuleigene Curriculum gestalten", „Unterrichten, Lernen, Erziehen" sowie „organisatorische Rahmenbedingungen sichern" (vgl. Tab. 3).

Tab. 3 Anzahl der Schulentwicklungsmaßnahmen an den inspizierten Schulen

Qualitätsbereich	N	min	max.	Summe	M	SD	%
1. Führung wahrnehmen	49	0	4	63	1,29	1,10	5,91
2. Personal entwickeln	49	0	5	89	1,82	1,56	8,35
3. Finanz- und Sachmittel gezielt einsetzen	49	0	6	96	1,96	1,55	9,01
4. Profil entwickeln und Rechenschaft ablegen	49	0	6	82	1,67	1,25	7,69
5. Das schuleigene Curriculum gestalten	49	0	9	142	2,90	1,65	13,32
6. Unterrichten, Lernen, Erziehen	49	0	7	119	2,43	1,89	11,16
7. Organisatorische Rahmenbedingungen sichern	49	0	7	114	2,33	1,89	10,69
8. Leistung beurteilen	49	0	7	69	1,41	1,40	6,47
9. Prozesse und Ergebnisse evaluieren	49	0	2	18	0,37	0,60	1,69
10. Förderkonzepte entwickeln	49	0	8	149	3,04	2,13	13,98
11. Beratungsangebote gestalten	49	0	3	19	0,39	0,67	1,78
12. Die Schulgemeinschaft beteiligen	49	0	6	61	1,24	1,30	5,72
13. Zufriedenes Personal, Schüler, Eltern, Betriebe	49	0	4	45	0,92	0,93	4,22
Gesamt				1066			100

5.4.3 Ausbau kapazitiver Stärken und Abbau kapazitiver Schwächen durch Schulentwicklung

Als nächstes wurde überprüft, ob es einen Zusammenhang zwischen den schulischen Kapazitäten und der Umsetzung von Maßnahmen infolge einer Inspektion gibt. Hierfür wurde getestet, ob Schulen mit positiver (stark oder sehr stark) Bewertung in einem Bereich eher Maßnahmen als Schulen mit negativer (schwach oder eher schwach) Bewertung umsetzen. Wie Tabelle 4 zeigt, lassen sich nur für die Bereiche „Führung wahrnehmen", „Personal entwickeln", „Prozesse evaluieren" und „Die Schulgemeinschaft beteiligen" statistisch signifikante Zusammenhänge feststellen. Infolge einer Inspektion verbessern vor allem Schulen, die bereits Stärken im Bereich

der Führung haben, diese noch einmal. Die Chance, dass Schulinspektion einen solchen Effekt nach sich zieht, ist an Schulen mit bereits ausgeprägten Kapazitäten im Bereich Führung 33-mal höher als an Schulen, die in diesem Bereich eher schwach aufgestellt sind. Schwächen werden hingegen vor allem im Bereich der Personalentwicklung abgestellt. An Schulen, die in diesem Bereich eher geringe Kapazitäten aufweisen, ist die Chance, dass Maßnahmen im Bereich der Personalentwicklung ergriffen werden, rund 7-mal höher als an Schulen, die in diesem Bereich bereits über ausgeprägte Kapazitäten verfügen.

Tab. 4 Chancen (odds ratios), dass eine Schule infolge einer Inspektion Maßnahmen in einem Bereich mit kapazitiven Stärken umsetzt

Qualitätsbereich	Odds Ratio	p
1. Führung wahrnehmen	33.000	<0.001
2. Personal entwickeln	0.152	<0.050
3. Finanz- und Sachmittel gezielt einsetzen	-	n.s.
4. Profil entwickeln und Rechenschaft ablegen	-	n.s.
5. Das schuleigene Curriculum gestalten	-	n.s.
6. Unterrichten, Lernen, Erziehen	-	n.s.
7. Organisatorische Rahmenbedingungen sichern	-	n.s.
8. Leistung beurteilen	-	n.s.
9. Prozesse und Ergebnisse evaluieren	5.075	<0.050
10. Förderkonzepte entwickeln	-	n.s.
11. Beratungsangebote gestalten	-	n.s.
12. Die Schulgemeinschaft beteiligen	12.310	<0.100
13. Zufriedenes Personal, Schüler, Eltern, Betriebe	-	n.s.

5.4.4 Anzahl von Schulentwicklungsmaßnahmen infolge einer Inspektion

Zu guter Letzt wurde untersucht, ob Schulen in Abhängigkeit ihrer kapazitiven Stärken und Schwächen unterschiedlich viele Maßnahmen ergreifen. Diesbezüglich zeigt sich, dass an Schulen insbesondere in denjenigen Bereichen Maßnahmen geplant und umgesetzt wurden, in denen ihre kapazitiven Stärken liegen. So wurde in Bereichen mit Schwächen (Bewertung „deutlich mehr Schwächen als Stärken“ und „eher mehr Schwächen als Stärken“) 367 Maßnahmen umgesetzt. In denjenigen Bereichen hingegen, die mit 3 („eher mehr Stärken als Schwächen“) und 4 („deutlich mehr Stärken als Schwächen“) bewertet wurden, wurden hingegen 699 Maßnahmen umgesetzt. Dieser Unterschied ist statistisch signifikant ($p<0.001$) und praktisch

bedeutsam (Cohen's d=0.45, r=0.21). Schulen, die über ausgeprägtere Kapazitäten in einem Bereich verfügen, setzen somit infolge einer Inspektion mutmaßlich deutlich mehr Maßnahmen in diesem Bereich um als Schulen mit geringen Kapazitäten.

Wie Tabelle 5 verdeutlicht, sind diese Zusammenhänge jedoch statistisch selten nachweisbar und vor allem artifiziell, was damit zusammenhängt, dass Schulen im Rahmen der Schulinspektion überdurchschnittlich häufig stärker als schwächer bewertet werden. Betrachtet man die vorhandenen Informationen hingegen mithilfe von Korrelationsanalysen, zeigt sich, dass sich mit Blick auf die Anzahl von Maßnahmen, die infolge einer Inspektion ergriffen werden, fast keine nachweisbaren Zusammenhänge feststellen lassen. Einzig in den Bereichen „Finanz- und Sachmittel gezielt einsetzen" sowie „Zufriedenes Personal, Schüler, Eltern, Betriebe" setzen Schulen mit gering ausgeprägten Kapazitäten in diesen Bereichen mehr Schulentwicklungsmaßnahmen um als Schulen mit ausgeprägten Stärken in diesen Bereichen. Auffällig ist weiterhin, dass ausgeprägte Stärken im Evaluationsbereich infolge einer Inspektion weiter ausgebaut werden.

Tab. 5 Zusammenhang innerschulischer Kapazitäten und der Anzahl umgesetzter Maßnahmen (Korrelationskoeffizienten)

Qualitätsbereich	Kendall-Tau-b	Kendall-Tau-c	Pearson-R	p
1. Führung wahrnehmen		-		n.s
2. Personal entwickeln		-		n.s.
3. Finanz- und Sachmittel gezielt einsetzen	-0,300	-0,314	-0,310	<0.001
4. Profil entwickeln und Rechenschaft ablegen		-		n.s.
5. Das schuleigene Curriculum gestalten		-		n.s.
6. Unterrichten, Lernen, Erziehen		-		n.s.
7. Organisatorische Rahmenbedingungen sichern		-		n.s.
8. Leistung beurteilen		-		n.s.
9. Prozesse und Ergebnisse evaluieren	0,290	0,210	0,266	<0.050
10. Förderkonzepte entwickeln		-		n.s.
11. Beratungsangebote gestalten		-		n.s.
12. Die Schulgemeinschaft beteiligen		-		n.s.
13. Zufriedenes Personal, Schüler, Eltern, Betriebe	-0,290	-0,227	-0,330	<0.001

6 Zusammenfassung, Diskussion und Ausblick

Ziel des Beitrags war es, zu untersuchen, inwieweit es einen Zusammenhang zwischen den schulischen Kapazitäten Organisationalen Lernens als Voraussetzung für innerschulische Verarbeitungsprozesse und den von Schulen im Anschluss an die Inspektion ergriffenen Maßnahmen der Schul- und Unterrichtsentwicklung gibt. Die Ausgangshypothese war, dass Schulen mit geringeren Kapazitäten weniger in der Lage sind Maßnahmen zu ergreifen als Schulen mit höheren Kapazitäten. Darüber hinaus stellte sich die Frage, ob Schulen mit unterschiedlichen Entwicklungskapazitäten sich in der Anzahl an Maßnahmen in den einzelnen Bereichen unterscheiden.

Die Befunde zeigen, dass sich die Ausgangshypothese nicht bestätigen lässt. Schulen mit geringeren Kapazitäten unterscheiden sich nicht systematisch von Schulen mit höheren Kapazitäten darin, wie wahrscheinlich es ist, dass sie Maßnahmen im Anschluss an die Inspektion ergreifen und umsetzen. Nur in drei Qualitätsbereichen „Führung wahrnehmen", „Prozesse und Ergebnisse evaluieren" und „Schulgemeinschaft beteiligen" ist die Wahrscheinlichkeit Maßnahmen zu ergreifen in Schulen mit eher stark oder stark ausgeprägten Kapazitäten höher. Beim Bereich „Personal entwickeln" zeigt sich hingegen der gegenteilige Effekt, dass Schulen mit geringeren Kapazitäten mit höherer Wahrscheinlichkeit Maßnahmen ergreifen.

In Bezug auf die Quantität von Maßnahmen zeigen sich ebenfalls kaum Unterschiede. Entsprechend decken sich die hier berichteten Befunde mit denjenigen aus anderen Untersuchungen, die versuchen, die berichtete Anzahl umgesetzter Veränderungen in Relation zur Inspektion zu setzen, und die ebenfalls zu dem Schluss kommen, dass diesbezüglich kein Zusammenhang feststellbar ist (vgl. Ehren und Visscher 2008; Gärtner et al. 2009). Die hier berichteten Befundmuster lassen auf Basis des theoretischen Modells keine eindeutigen Interpretationen zu.

Jedoch ist grundsätzlich auffällig, dass Maßnahmen an Schulen mit ausgeprägten Kapazitäten vor allem dort ergriffen werden, wo sie vonseiten der Schulleitung direkt steuerbar sind. Maßnahmen, die eine breite Beteiligung von Lehrerkollegien und damit vergleichsweise komplexe innerschulische Abstimmungsprozesse erfordern, werden infolge einer Schulinspektion hingegen eher nicht ergriffen. Auch scheint es so, dass bei geringen Kapazitäten vor allem auf eine Vielzahl symbolischer Entwicklungsmaßnahmen gesetzt wird; es werden dann möglichst viele Maßnahmen mit Blick auf (a) die Außenwirkung umgesetzt, die dokumentieren, dass Mittel gezielt eingesetzt werden und (b) die Zufriedenheit der schulischen Akteure durchgeführt, die zeigen sollen, dass diese seitens der Schule Beachtung finden. Praktisch werden in diesen Fällen z. B. Klassen mit Smartboards ausgestattet, neue Gebäude gebaut oder aber auch Zufriedenheitsbefragungen durchgeführt.

Sichtbar wird insofern, dass Schulinspektion ihre Wirkung primär auf Ebene der Schulleitung entfaltet, wobei dies, wie Preuß et al. (2015, S. 127f.) betonen, sowohl Vor- als auch Nachteile für die Schulentwicklung mit sich bringt: „Die Schulleitungen sind die zentralen Ansprechpartner der Inspektion. Sie tragen gegenüber Lehrkräften, Eltern, SchülerInnen und Schulaufsicht die Verantwortung für die Umsetzung bildungspolitischer Reformen in ihren Schulen. (…) Durch die Inspektion werden Schulleitungen – in Relation zu den Lehrkräften – strukturell aufgewertet, da sie die Ansprechpartner der Inspektion sind. Dies verstärkt ihren Einfluss auf Schulentwicklungsprozesse; die Handlungsmöglichkeiten der weiteren Akteure in der Schule bleiben von der Qualität der Führung abhängig."

Mit Blick auf den weiteren Forschungsbedarf macht die vorliegende Studie darüber hinaus deutlich, dass die einfach-linearen Annahmen zur Wirkungsweise und zur Wirksamkeit von Schulinspektionen auf Schulentwicklung, die aktuell im Rahmen der Schulinspektionsforschung häufig anzutreffen sind, deutlich unterkomplex sind und selbst durch das hier genutzte Vorgehen nur im Ansatz beschrieben und evaluiert werden können. Im Folgenden werden daher zwei Erklärungsansätze diskutiert, die es in weiteren Studien zu überprüfen gilt:

1. Ein Grund für die nicht bestätigte Hypothese, dass Schulen mit geringeren Kapazitäten weniger in der Lage sind Maßnahmen zu ergreifen als Schulen mit höheren Kapazitäten, könnte daran liegen, dass keine der Schulen von der Schulinspektion in allen Qualitätsbereichen, die der Kapazität zugerechnet werden können, als „schwach" eingeschätzt wird und somit alle Schulen über ein Minimum an Kapazitäten verfügen, um Entwicklungsprozesse in Gang zu bringen.
2. Ein weiterer Grund könnte darin liegen, dass sich die Schulen zwar nicht systematisch in der Quantität der Maßnahmen, sondern in der Qualität (der Art und Weise wie tiefgreifend, umfangreich und wirksam sie sind) unterscheiden. So werden im Bereich „Lernen, Unterrichten, Erziehen" beispielsweise auf der einen Seite Maßnahmen ergriffen, die auf eine Individualisierung der Unterrichtspraktiken abzielen, aber auf der anderen Seite eben auch solche, die eine stärkere Medienorientierung im Unterricht fokussieren. Auch im Bereich „Leistungen beurteilen" zeigt sich ein ähnlich breites Bild: So erarbeiten einige Schulen infolge einer Inspektion Maßstäbe zur Bewertung von Schülerleistungen, während andere sich um die Modalitäten der Zeugnisausgabe bemühen. Dabei muss mit Blick auf die vorgelegte Studie auch in Rechnung gestellt werden, dass es sich bei den für die Analysen genutzten Daten um Selbstberichte der Schulen handelt, die aufgrund der vorherigen Inspektionsergebnisse unter einem Legitimationsdruck gegenüber Schulaufsicht und Inspektion stehen. Dies gilt insbesondere

für Schulen, die im ersten Inspektionszyklus das Prädikat „eher schwach" erhalten haben. Somit ist davon auszugehen, dass diese Schulen eventuell noch mehr darauf bedacht sind, Schulentwicklungsaktivitäten zu demonstrieren, als Schulen die weniger unter Legitimationsdruck stehen. Dies gilt insbesondere in Qualitätsbereichen, von denen die Schulen bzw. Schulleitungen ausgehen, dass diesen von Seiten der Administration und ggf. auch Eltern eine höhere Priorität eingeräumt wird und in denen gleichzeitig die Schulinspektion einen erhöhten Bedarf bescheinigt hat. Hierfür könnte die Anzahl der umgesetzten Maßnahmen pro Schule ein Hinweis sein. Die drei Bereiche mit der höchsten durchschnittlichen Anzahl an Maßnahmen pro Schule betreffen alle das Kerngeschäft der Schule, den Unterricht („Förderkonzepte entwickeln", „Das schuleigene Curriculum gestalten" und „Unterrichten, Lernen, Erziehen"). Offen bleibt neben der Qualität, inwieweit sich zumindest ein Teil der Maßnahmen lediglich dem „Als-Ob-Handeln" im Sinne der „espoused theorie" von Argyris und Schön (1978) zuzuschreiben lassen. Solche Maßnahmen dienen dann nur der Legitimation bei gleichzeitigem Erhalt des Status Quo. Somit haben sie keinen Einfluss auf die pädagogische Arbeit in Schule und Unterricht.

Die unklaren Befundmuster weisen auf die Notwendigkeit hin, die Verarbeitungsprozesse selbst in den Blick zu nehmen, um zu verstehen, welche Impulse von der Schulinspektion ausgehen. Die Berichte der Schulen könnten ein guter Ausgangspunkt für eine qualitative Studie an den Schulen sein. Ziel einer solchen Studie könnte es sein, nähere Informationen über die Rezeption und Nutzung der Inspektionsdaten und die Ableitung der Maßnahmen, deren Umsetzung und Bedeutung für die pädagogische Arbeit zu gewinnen. Die Kapazitäten des Organisationalen Lernens könnten hier als Analyseheuristik dienen, um die beteiligten Prozesse und deren Beitrag in den Blick zu nehmen. Im Sinne von Argyris und Schön ginge es auch darum zu rekonstruieren, inwieweit unterschiedliche Handlungsmodi in Schulen erkennbar sind und ob Schulen, die eine hohe Diskrepanz zwischen „espoused theorie" und „theorie in use" haben, zu einem höheren als „Als-Ob-Handeln" tendieren und weniger tiefergreifende Maßnahmen umsetzen. Hierbei wäre es auch gewinnbringend, die schulischen Strategien zu analysieren, die mit dem Bericht verfolgt werden. Der Nachteil einer solchen Vorgehensweise ist, dass die Prozesse nur im Nachhinein rekonstruiert werden können. Um dies zu vermeiden, wäre es sinnvoll, Schulen im Längsschnitt direkt im Anschluss an die Inspektion oder gar schon im Vorfeld des Inspektionsbesuchs bis zum nächsten Inspektionszyklus wissenschaftlich zu begleiten.

Literatur

Argyris, C. & Schön, D. (1978). *Organizational Learning – A Theory of Action Perspective*. Reading, Mass: Addison Wesley.

Arbeitsgruppe Bildungsforschung/Bildungsplanung. (2004). *Das deutsche Schulsystem. Entstehung. Struktur. Steuerung*. Retrieved from www.uni-essen.de/bfp/lhe/skripte.php

Van Ackeren, I., & Klemm, K. (2009). Entstehung, Struktur und Steuerung des deutschen Schulsystems. Wiesbaden: VS.

Altrichter, H., & Kemethofer, D. (2015). Does Accountability Pressure through School Inspections Promote School Improvement? *School Effectiveness and School Improvement*, *26*(1), 32–56.

Altrichter, H., & Maag Merki, K. (Hrsg.) (2010). Educational Governance: Bd. 7. *Handbuch neue Steuerung im Schulsystem* (1. Aufl). Wiesbaden: VS, Verl. für Sozialwiss.

Behörde für Bildung und Sport (2006). Orientierungsrahmen Schulqualität– Qualitätsentwicklung an Hamburger Schulen. Hamburg: Behörde für Bildung und Sport.

Böhm-Kasper, O., & Selders, O. (2013). „Schulinspektionen sollten regelmäßig durchgeführt werden"?: Ländervergleichende Analyse der Wahrnehmung und Akzeptanz von Schulinspektionsverfahren. In I. van Ackeren, M. Heinrich, & F. Thiel (Hrsg.), *Evidenzbasierte Steuerung im Bildungssystem* (S. 121–153).

Bormann, I. (2001). Organisationsentwicklung und organisationales Lernen von Schulen. Eine empirische Untersuchung am Beispiel des Umweltmanagements. Opladen: Leske + Budrich.

Bortz, J. & Döring, N. (1995). Forschungsmethoden und Evaluation für Sozialwissenschaftler. Berlin: Springer.

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.

Brown, M., Boyle, B., & Boyle, T. (1999). Commonalities between Perception and Practice in Models of School Decision Making Systems in Secondary Schools in England and Wales. *Annual Meeting of the American Educational Research Association*. Montreal.

Cousins, J. B. (2000). Intellectual Roots of Organizational Learning. In J. Cibulka (Hrsg.), Advances in Research and Theories of School Management and Educational Policy: Vol. 4. Organizational learning in educational policy systems (9th Aufl., S. 219–235). Stamford, Connecticut.

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

Daft, R. L., & Huber, G. P. (1987). How Organizations Learn – A Communication Framework. *Research in the Sociology of Organizations, 5,* 1–36.

Dalin, P., & Rolff, H.-G. (1990). Institutionelles Schulentwicklungsprogramm: Eine neue Perspektive für Schulleiter, Kollegium und Schulaufsicht. Soest: Soester Verl.-Kontor.

Dedering, K., Fritsch, N., & Weyer, C. (2012). Die Ankündigung von Schulinspektionen und deren innerschulische Effekte – hektisches Treiben oder genügsame Gelassenheit? In S. Hornberg & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 205–222). Münster: Waxmann.

Diedrich, M. (2015a). Aufbau und Rolle der Schulinspektion Hamburg. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), *Schulinspektion in Hamburg: Der erste Zyklus 2007-2013 – Grundlagen, Befunde, Perspektiven* (S. 57-76). Münster: Waxmann.

Diedrich, M. (2015b). Der zweite Zyklus der Schulinspektion Hamburg – Ein Ausblick. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), *Schulinspektion in Hamburg: Der erste Zyklus 2007-2013 – Grundlagen, Befunde, Perspektiven* (S. 419-436). Münster: Waxmann.

Duncan, R., & Weiss, A. (1979). Organizational Learning: Implications for Organizational Design. *Research in Organizational Behavior,* 75–123.

Ehren, M. C. M & Scheerens, J. (2015). Evidenzbasierte Referenzrahmen zur Schulqualität als Grundlage von Schulinspektion. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), *Schulinspektion in Hamburg: Der erste Zyklus 2007-2013 – Grundlagen, Befunde, Perspektiven* (S. 233-272). Münster: Waxmann.

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.

Ehren, M. C. M., & Visscher, A. J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies*, *56*(2), 205–227.

Ehren, M. C. M., Gustafsson, J.-E., Altrichter, H., Skedsmo, G., Kemethofer, D. & Huber, S. G. (2015). *Comparing effects and side effects of different school inspection systems across Europe. Comparative Education*. DOI: 10.1080/03050068.2015.1045769

Feldhoff, T. (2011). Schule organisieren. Der Beitrag von Steuergruppen und Organisationalem Lernen zur Schulentwicklung. Wiesbaden: VS Verlag.

Feldhoff, T., Gromala, L., & Brüsemeister, T. (2014). Organisationales Lernen von Schulen im Kontext datenbasierter Steuerung. In H. G. Holtappels (Ed.), *Schulentwicklung und Schulwirksamkeit als Forschungsfeld*. Münster: Waxmann. (S. 241-258).

Feldhoff, T., Kanders, M. & Rolff, H.-G. (2008). Verortung und empirische Operationalisierung erweiterter Selbstständigkeit. In H. G. Holtappels, K. Klemm & H.-G. Rolff (Hrsg.), *Schulentwicklung durch Gestaltungsautonomie. Ergebnisse der Begleitforschung zum Modellvorhaben ‚Selbstständige Schule' in Nordrhein- Westfalen.* (S. 47–62). Münster: Waxmann.Feldhoff, T., & Rolff, H.-G. (2008). Einfluss von Schulleitungs- und Steuergruppenhandeln. In H.-G. Holtappels, K. Klemm, & H.-G. Rolff (Hrsg.), *Schulentwicklung durch Gestaltungsautonomie. Ergebnisse der Begleitforschung zum Modellvorhaben ‚Selbstständige Schule' in Nordrhein-Westfalen* (S. 293–302). Münster: Waxmann.

Gärtner, H., Hüsemann, D., & Pant, H. A. (2009). Wirkungen von Schulinspektion aus Sicht betroffener Schulleitungen. Die Brandenburger Schulleiterbefragung. *Empirische Pädagogik*, *23*(1), 1–18.

Geißler, H. (1995). Grundlagen des Organisationslernens.. Weinheim: Deutscher Studien Verlag.

Hallinger, P. (2003). Leading Educational Change: reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329–352.

Hallinger, P., & Heck, R. H. (2010). Leadership for learning: does collaborative leadership make a difference in school improvement? *Educational Management Administration & Leadership, 38*(6), 654–678.

Hanson, M. (2001). Institutional Theory and Educational Change. *Educational Administration Quarterly, 5*(37), 637–661.

Hedberg, B. (1981). How organizations learn and unlearn. In P. C. Nystrom & W. H. Starbuck (Hrsg.), *Handbook of Organizational Design. Adapting organizations to their environments*. Oxford: Oxford University Press.

Heinrich, M., Lambrecht, M., Böhm-Kasper, O., Brüsemeister, T., & Wissinger, J. (2014). Funktionen von Schulinspektion?: Zum Governance-Programm der Vergewisserung und der Weiterentwicklung der Qualität schulischer Arbeit. In C. Fischer (Hrsg.), *Münstersche Gespräche zur Pädagogik: Vol. 30. Damit Unterricht gelingt. Von der Qualitätsanalyse zur Qualitätsentwicklung* (S. 19–53). Münster, Westf: Waxmann.

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

Holtappels, H.-G. (2007). Schulentwicklungsprozesse und Change Management. Innovationstheoretische Reflexionen und Forschungsbefunde über Steuergruppen. In N. Berkemeyer & H.-G. Holtappels (Hrsg.), *Schulische Steuergruppen und Change Management*. (S. 5–34). Weinheim und München: Juventa.

Huber, G. P. (1991). Organizational learning: The contributing processes and the literature. *Organization Science, 2/2,* 88–115.

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.

Jones, O. (2006). Developing Absorptive Capacity in Mature Organizations – The Change Agent's Role. *Management Learning, 37 (3),* 355–376.

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.

Kruse, S. D., & Louis, K. S. (2000). Creating Community in Reform: Images of Organizational Learning in Inner- City Schools. In J. Cibulka (Hrsg.), *Advances in Research and Theories of School Management and Educational Policy: Vol. 4. Organizational learning in educational policy systems* (9th Aufl., S. 17–46). Stamford, Connecticut.

Kruse, S. D., Louis, K. S., & Bryk, A. S. (1995). An Emerging Framework for Analyzing School-Based Professional Community. In K. S. Louis & S. D. Kruse (Hrsg.), *Professionalism and Community: Perspectives on Reforming Urban Schools* (S. 23–44). Thousand Oaks: Sage Publications.

Landwehr, N. (2011). Thesen zur Wirkung und Wirksamkeit der externen Schulevaluation. In C. Quesel, V. Husfeldt, N. Landwehr & P. Steiner (Hrsg.), *Wirkungen und Wirksamkeit der externen Schulevaluation* (S. 35–70). Bern: hep.

Larson-Knight, B. (Hrsg.) (2000). Advances in Research and Theories of School Management and Educational Policy. Leadership, Culture and Organizational Learning (4th Aufl.). Stamford, Connecticut.

Leithwood, K. A., Jantzi, D., & Fernandez, A. (1994). Transformational leadership and teachers commitment to change. In Murphy, J. F. & Louis, K. S. (Hrsg.), *Reshaping the principal ship: Insights from transformational reform efforts* (S. 77–98). Thousand Oaks.

Leithwood, K. A., Jantzi, D., & Steinbach, R. (2000). Leadership and other Conditions which foster Organizational Learning in Schools. In J. Cibulka (Hrsg.), *Advances in Research and Theories of School Management and Educational Policy: Vol. 4. Organizational learning in educational policy systems* (9th Aufl., S. 67–89). Stamford, Connecticut.

Leithwood, K. A., Leonard, L., & Sharratt, L. (2000). Conditions Fostering organizational Learning in Schools. In K. A. Leithwood & K. S. Louis (Hrsg.), *Advances in Research and Theories of School Management and Educational Policy. Understanding Schools as intelligent Systems* (S. 99-123).. Stamford, Connecticut.

Lewin, K. (1963). *Feldtheorie in den Sozialwissenschaften*. Bern: Hans Huber Verlag.

Lortie, D. C. (1972). Team Teaching. Versuch der Beschreibung einer zukünftigen Schule. In H.-W. Dechert (Hrsg.), *Team Teaching in der Schule*. (S. 37–76). München.

Louis, K. S. (Hrsg.). (2006). *Organization for School change*. Milton Park, Abingdon.

Louis, K. S., & Dentler, R. A. (1988). Knowledge use and school improvement. *Curriculum Inquiry, 18*(1), 32–62.

Louis, K. S., & Leithwood, K. A. (2000). From Organizational Learning to Professional Learning. In J. Cibulka (Hrsg.), Advances in Research and Theories of School Management and Educational Policy: Vol. 4. Organizational learning in educational policy systems (9th Aufl., S. 275–285). Stamford, Connecticut.

Louis, K. S., & Marks, H. M. (1998). Does professional community affect the classroom?: Teacher work and student work in restructuring schools. *American Journal of Education, 106 (4),* 532–575.

March, J. G., & Olsen, J. P. (1976). *Ambiguity and choice in organizations*. Bergen.

Marks, H. M., & Louis, K. S. (1999). Teacher Empowerment and the Capacity for Organizational Learning. *Educational Administration Quarterly, 35*(5), 707–750.

Marks, H. M., Louis, K. S., & Printy, S. M. (2000). the Capacity for Organizational Learning – Implications for pedagogical quality and student achievement. In K. A. Leithwood (Hrsg.), *Understanding schools as intelligent systems, Advances in Research and Theories of School Management and Educational Policy* (4th Aufl., S. 239–265). Stamford, Connecticut.

Meetz, F. (2007). Personalentwicklung als Element der Schulentwicklung: Bestandsaufnahme und Perspektiven/ Frank Meetz. Klinkhardt Forschung. Bad Heilbrunn: Klinkhardt.

Mintzberg, H. (1992). *Die Mintzberg-Struktur: Organisation effektiver gestalten*. Landsberg/ Lech: Verl. Moderne Industrie.

Mulford, B., & Silins, H. C. (2003). Leadership for Organisational Learning and Improved Student Outcomes—What Do We Know? *Cambridge Journal of Education, 33*(2), 175–195.

Murphy, J. F., & Louis, K. S. (1994). Reshaping the Principalship: Insights from transformational reform efforts. Thousand Oaks.

Newmann, F. M., King, M., & Ridgon, M. (1997). Accountability and School Performance. Implications from restructuring Schools. *Harvard Educational Review, 67*(1), 41–74.

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

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, M. & N. Berkemeyer (Hrsg.), *Schul- und Bildungsforschung. Diskussionen, Befunde und Perspektiven* (S. 167-185). Münster: Waxmann.

Pietsch, M., Janke, N., & Mohr, I. (2014). Führt Schulinspektion zu besseren Schülerleistungen? Difference-in-Differences-Studien zu Effekten der Schulinspektion Hamburg auf Lernzuwächse und Leistungstrends. *Zeitschrift für Pädagogik*, *60*(3), 446-470.

Preuß, B., Brüsemeister, T., & Wissinger, J. (2015). Einführung der Schulinspektion: Struktur und Wandel regionaler Governance im Schulsystem. In H. J. Abs, Brüsemeister, T., M.

Schemmann, & T. Wissiger (Hrsg.), *Governance im Bildungssystem: Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 117-142). Wiesbaden: VS.

Rait, E. (Hrsg.) (1995). Images of School. Against the Current – Organizational Learning in schools. California.

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

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The Impact of Leadership on Student Outcomes: An Analysis of the Differential Effects of Leadership Types. *Educational Administration Quarterly, 44*(5), 635–674. doi:10.1177/0013161X08321509

Rolff, H.-G. (1993). Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule. Weinheim und München: Juventa.

Schubert, H.-J. (2004). *Management von Veränderungen* (Studienbrief PE1110). Kaiserslautern.

Silins, H. C., Mulford, B., Zarins, S., & Bishop, P. (2000). Leadership for Organizational Leadership in Australian Secondary Schools. In K. A. Leithwood (Hrsg.), *Understanding schools as intelligent systems, Advances in Research and Theories of School Management and Educational Policy* (4th Aufl., S. 267–291). Stamford, Connecticut.

Sowada, M. G. (2015). Expertenurteile – Achillesferse oder Trumpf der Schulinspektion. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), *Schulinspektion in Hamburg: Der erste Zyklus 2007-2013 – Grundlagen, Befunde, Perspektiven* (S. 137-156). Münster: Waxmann.

Stringfield, S. (2000). Organizational Learning and Current Reform Efforts: From Exploitation to Exploration. In J. Cibulka (Hrsg.), *Advances in Research and Theories of School Management and Educational Policy: Vol. 4. Organizational learning in educational policy systems* (9th Aufl., S. 261–274). Stamford, Connecticut.

Visscher, A. J., & Coe, R. (2003). School Performance Feedback Systems: Conceptualisation, Analysis and Reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349.

Weber, M. (1972). *Wirtschaft und Gesellschaft.* Grundriss der verstehenden Soziologie (5. Aufl.). Tübingen.

Weick, K. E., & Roberts, K. H. (1993). Collective Mind in Organizations: Heedful Interrelating on Flight Decks. *Administrative Science Quarterly, 38(3),* 357–381.

Wiegand, M. (1998). Prozesse organisationalen Lernens. Wiesbaden.

Zarcula, A.-M. (2006). *Wissensmanagement in Forschungseinrichtungen Konzeption und Praxis*. Retrieved from http://deposit.ddb.de/cgi-bin/dokserv?idn=981150381&dok_var=d1&dok_ext=pdf&filename=981150381.pdf (10:21, 11.10.2007)

Empirische Pädagogik
2017, 31. Jahrgang, Heft 2, S. 202-220

Von der Schulinspektion zur Unterrichtsentwicklung: Welche Rolle spielt die Schulleitung?

Marcus Pietsch und Ingmar Hosenfeld

In diesem Beitrag wird untersucht, welchen Einfluss unterschiedliche Facetten des Führungshandelns (Transformationale, Instruktionale, Kollaborative Führung) von Schulleitungen in Relation zur Innovationskapazität, der Qualität der Kooperation und der Nutzung von Schulinpsektionsrückmeldungen auf die Weiterentwicklung der Unterrichtsqualität besitzen. Dazu wurden die Daten von n = 933 Lehrkräften aus 82 niedersächsischen Schulen mit Hilfe von Strukturgleichungsmodellen analog zu den logischen Modellen von Leithwood et al. (2008; 2006; 2002) analysiert. Das empirische Modell weist einen sehr guten Modellfit auf. Es zeigt sich, dass transformationales Führungshandeln insgesamt den größten Einfluss auf die Verbesserung der Unterrichtsqualität besitzt, wenngleich es keinen signifikanten direkten Zusammenhang mit der Kriteriumsvariable gibt. Die Instruktionale Führung wirkt insgesamt nur wenig schwächer auf die AV, hierbei ist jedoch auch der direkte Einfluss auf das Kriterium signifikant. Schließlich prädiziert auch das Ausmaß der Nutzung der Inspektionsdaten die berichtete Verbesserung des Unterrichts. Unterwartet ist der negative direkte Einfluss der Kooperation auf die Verbesserung des Unterrichts.

Schlagwörter: Instruktionale Führung – Schulentwicklung – Schulinspektion – Transformationale Führung – Wirksamkeit

1 Einleitung

Als Instrument der Neuen Steuerung im Bildungssystem sollen Schulinspektionen zu einer positiven und zielgerichteten Entwicklung von Schule und Unterricht beitragen, die letztlich in verbesserten Schülerleistungen mündet. Anders als z. B. Vergleichsarbeiten richten sich Schulinspektionen in erster Linie an Schulleitungen. Denn aus Perspektive der Bildungsadministration zeichnen Schulleitungen einerseits verantwortlich für die Umsetzung bildungspolitischer Innovationen und Reformen in der Einzelschule sowie andererseits für die Einhaltung von bildungsrelevanten (Mindest-)Standards.

Durch die Etablierung der Inspektionsverfahren in den Ländern seit der Jahrtausendwende, wurden Schulleitungen, im Vergleich zu Lehrkräften und anderem schulischen Personal, strukturell aufgewertet und es wurde ihnen eine herausgehobene und besonders führungsorientierte und steuernde Rolle zugewiesen, mit der Folge, dass die Handlungsmöglichkeiten der weiteren Schulbeteiligten und damit auch deren Möglichkeiten, Schule und Unterricht weiter zu entwickeln, zunehmend von der Qualität der Führung einer Schule abhängen (Preuß, Wissinger & Brüsemeister, 2015).

Damit eine datengestützte Schul- und Unterrichtsentwicklung gelingt, gilt eine Vielzahl externer aber auch innerschulischer Faktoren als relevant. So spielen diesbezüglich, neben der Anlage und Ausgestaltung von Datenrückmeldungen (Visscher & Coe, 2003) und externen Bedingungen, wie z. B. Wettbewerbsdruck (Levin & Belfield, 2003), auch innerschulische Voraussetzungen zum produktiven und zielgerichteten Umgang mit Daten zur Weiterentwicklung von Schule und Unterricht eine wichtige Rolle (Anderson, Leithwood & Strauss, 2010). Solche innerschulischen Kapazitäten organisationalen Lernens (Feldhoff, Gromala & Brüsemeister, 2014) umfassen dann beispielsweise Organisationsstrukturen, personale Eigenschaften von Lehrkräften, gemeinsame Zielvorstellungen der Schulbeteiligten und Merkmale der Schulkultur.

Einen besonders hohen Stellenwert nimmt dabei wiederum das Führungshandeln von Schulleitungen ein, das als Katalysator für die datengestützte Schul- und Unterrichtsentwicklung gilt (Leithwood, Aitken & Jantzi, 2006). So zeigt eine Vielzahl von Untersuchungen, dass eine solche vor allem dann gelingt (Marsh & Farrell, 2015; Schildkamp & Lai, 2013), wenn Schulleitungen förderliche Rahmenbedingungen schaffen und z. B. ausreichend Zeit bereit stellen, damit sich Lehrkräfte mit Evaluationsergebnissen intensiv auseinandersetzen können und Schulleitungen Kooperationsstrukturen zum Austausch sowie zur Abstimmung aufbauen. Des Weiteren können Schulleitungen Einfluss darauf nehmen, wie Lehrerinnen und Lehrer mit Informationen umgehen, indem sie deren Einstellungen und Haltungen mit Blick auf die datengestützte Schul- und Unterrichtsentwicklung beeinflussen (Levin & Datnow, 2012). Auch sind Schulleitungen maßgeblich für die Datennutzungskultur einer Schule verantwortlich und können diese positiv beeinflussen, indem sie dafür sorgen, dass verbindliche und transparente Regeln zum Umgang mit Evaluationsdaten aufgestellt und durch das Kollegium mitgetragen werden (Marsh, 2012). Mit Blick auf die konkrete Unterrichtsentwicklung ist es wiederum hilfreich, wenn Schulleitungen sich als Experten für den Unterricht verstehen und entsprechend koordinierend und steuernd in das Lernen und Lehren an einer Schule eingreifen (Halverson, Grigg, Prichett & Thomas, 2007; van Geel, Visscher & Teunis, 2017).

Mit Blick auf die Wirksamkeit von Schulinspektion sind daher die Einstellungen und Verhaltensweisen von Schulleitungen grundsätzlich gewichtige Faktoren, die die Fähigkeit von Lehrkräften bzw. einer Schule zur Erneuerung und Weiterentwicklung maßgeblich beeinflussen und in der Folge sowohl Schul- als auch Unterrichtsentwicklung fördern aber auch behindern können (Dedering & Müller, 2008).

Studien, die den Einfluss des Schulleitungshandelns auf die Nutzung von Inspektionsdaten in Schulen untersuchen und prüfen, in wie weit das Führungshandeln die angenommene Wirksamkeit von Schulinspektionen moderiert, liegen für den deutschsprachigen Raum bislang jedoch nicht vor. Daher wird im Rahmen dieses

Beitrages untersucht, wie unterschiedliche Facetten des Führungshandelns von Schulleitungen die inspektionsbasierte Weiterentwicklung der Unterrichtsqualität sowie die Innovationskapazität, die Qualität der Kooperation im Kollegium sowie die Nutzung von Schulinspektionsrückmeldungen beeinflussen.

2 Führung an Schulen und wie sie auf die Unterrichtsentwicklung wirkt

2.1 Führung durch Schulleitungen

Um Schul- und Unterrichtsentwicklung voran zu treiben, können Schulleitungen transformationale und unterrichtsbezogene Führungsfacetten miteinander kombinieren, deren Wirksamkeit unter Umständen durch ergänzende kollaborative bzw. distributive Führungsmaßnahmen verstärkt werden können (Hallinger, 2003; Heck & Hallinger, 2010).

Transformationale Führung an Schulen bezeichnet dabei einen Führungsstil, bei dem es darum geht, Mitarbeiterinnen und Mitarbeiter durch die Formulierung attraktiver, sinnstiftender Zukunftsvisionen für Schule und Unterricht zu inspirieren und zu motivieren, ihre Kreativität und Fähigkeit zur eigenständigen Problemlösung zu fördern, auf ihre individuellen Bedürfnisse, Talente und Potenziale einzugehen sowie als Vorbild zu agieren, um es den Schulbeteiligten auf diesem Wege zu ermöglichen, sich mit der Schule und ihren Zielen zu identifizieren (Bass & Avolio, 1994). Ziel transformationaler Führung ist es, Schulen in ihrer Gesamtheit aktiv und nachhaltig weiter zu entwickeln, sie in lernende Organisationen zu verwandeln und es dem schulischen Personal zu ermöglichen, kritisch-konstruktiv mit Innovationen und Veränderungen umzugehen (Leithwood & Jantzi, 2005).

Schulleitungen, die einen solchen Führungsstil häufig nutzen, können durch ihr Handeln den Schülerinnen und Schülern an den durch sie geleiteten Schulen einen deutlichen Lernvorsprung ermöglichen (vgl. Robinson, Lloyd & Rowe, 2008). Darüber hinaus lassen sich verschiedene positive Auswirkungen auf lehr- und lernrelevante Voraussetzungen nachweisen, die ihrerseits wiederum einen Einfluss auf die Schul- und Unterrichtsentwicklung, und vermittelt hierüber, auf einen möglichen Lernerfolg von Schülerinnen und Schülern haben können. So erhöht ein transformationales Schulleitungshandeln u. a. die Fähigkeit von Lehrerinnen und Lehrern, mit Innovationen und Veränderungen, z. B. Schulreformen oder neuartigen inhaltlichen und / oder pädagogischen Anforderungen an sie und ihre Arbeit, konstruktiv umzugehen – die so genannte Innovationskapazität – um das Sechs- bis Achtfache und steigert darüber hinaus die Chance auf eine gelingende Zusammenarbeit im Kollegium deutlich (Leithwood & Jantzi, 2006). Auch wird der Umgang mit Daten aus Evaluationen positiv beeinflusst (Stump, Zlatkin-Troitschanskaia & Mater, 2016),

was wiederum mit einer um bis zu rund 20 Prozent gesteigerten Weiterbildungsaktivität von Lehrkräften (Thoonen, Sleegers, Oort, Peetsma & Geijsel, 2011) einhergeht, die sich letztendlich in einer nachhaltigen Verbesserung des Unterrichts auswirkt (Pietsch, Lücken, Thonke, Klitsche & Musekamp, 2016). Häufig transformational geführte Lehrkräfte wenden entsprechend deutlich häufiger neuartige, zeitgemäße und für den Lernerfolg von Schülerinnen und Schülern optimierte Unterrichtsmethoden und -strategien an als ihre Kolleginnen und Kollegen, die nur selten in einem solchen Stil durch ihre Schulleitung geführt werden (Leithwood & Jantzi, 2006; Thoonen et al., 2011).

Instruktionale Führung wiederum bezeichnet einen Führungsstil, bei dem es darum geht, lern- und lehrrelevante Bedingungen und Prozesse an einer Schule zu optimieren (Hallinger & Murphy, 1985). Ziel instruktionalen Handelns von Schulleitungen ist es, Schülerinnen und Schülern ein bestmögliches Lernen und einen optimalen Kompetenzerwerb zu ermöglichen. Schulleitungen, die diesem Modell folgen, prägen und gestalten Unterricht und Erziehung an den durch sie geleiteten Schulen aktiv. Sie stellen hohe Ansprüche an Lern- und Entwicklungsziele, stimmen curriculare Ziele mit tatsächlichen Lehr- und Lernpraktiken an der Schule ab, richten individuelle Fortbildungsmaßnahmen an diesen Zielen aus, supervidieren Unterricht und Lernfortschritte, geben evidenzbasierte Rückmeldungen zum Lernen und Lehren und sind kompetente Ansprechpartner, wenn Probleme in Unterrichtsfragen auftreten (Hallinger, 2011; Hallinger, Leithwood & Heck, 2010).

Schulleitungen, die derartige Führungspraktiken häufig nutzen, tragen zum Lernerfolg der Schülerinnen und Schülern an den durch sie geleiteten Schulen in besonders hohem Maße bei (Robinson et al., 2008). Insbesondere dann, wenn sie ihr Handeln stark auf den Unterricht fokussieren, sich selber also als Expertinnen und Experten für Unterrichtsfragen verstehen und sich entsprechend aktiv in das Lernen und Lehren an der Schule einmischen, kann die Chance, dass die Schülerinnen und Schüler an den durch sie geleiteten Schulen überdurchschnittliche Kompetenzzuwächse erzielen, um mehr als das Vierfache gesteigert werden (Marzano, McNulty & Waters, 2005; Seashore Louis, Leithwood, Wahlstrom & Anderson, 2010). Eine wichtige Rolle spielen dabei wiederum Daten aus Evaluationen, da diese, wenn vorhanden, dazu genutzt werden, um den Unterricht an einer Schule wissensbasiert und zielgerichtet zu verbessern (Halverson et al., 2007). Marks und Printy (2003) untersuchen das Zusammenspiel von Transformationaler und Instruktionaler Führung und zeigen, dass vor allem die Kombination zu hoher Instruktionsqualität und Schülerleistung führt.

Kollaborative Führung beinhaltet neben einer Beteiligung des Kollegiums bei wichtigen Entscheidungen, die die Schule als Ganze betreffen, eine dezentrale Führung,

die die Verantwortung bei anstehenden Reformprojekten auf einzelne Teams überträgt, diese unterstützt und motiviert (Denis, Langley & Sergi, 2012). Diese Führungsfacette ist darauf ausgerichtet, Mitarbeiterinnen und Mitarbeiter in Entscheidungsprozesse einzubeziehen und Führungsverantwortung auf verschiede-ne Beteiligte innerhalb einer Organisation zu verteilen. Ziel eines solchen Führungshandelns ist es, eine Kultur zu schaffen, in deren Rahmen sich die Mitarbeiterinnen und Mitarbeiter einer Schule aktiv in die Ausgestaltung der Führung einbringen können. Auf diesem Wege soll innerschulische Kohärenz hergestellt, die Bereitschaft zur Verantwortungsübernahme von Einzelpersonen gesteigert und letztlich die Effektivität des Steuerungshandelns erhöht werden. Ein solch partizipatives und teilendes Führungsverhalten soll es konsequenterweise ermöglichen, Führung auf allen Ebenen der Organisation wirksam werden zu lassen und flächendeckend komplexe Veränderungen und Innovationen zu ermöglichen, die sich letztendlich auf die Schul- und Unterrichtsentwicklung niederschlagen.

Kollaborative Führung wirkt dabei als eine Art „Kitt" (vgl. Harris, 2004), die die vielfältigen Expertisen, Entscheidungen und Zielsetzungen innerhalb einer Schule orchestriert und auf ein gemeinsames Ganzes hin ausrichtet. Dabei wirkt sich eine solche Führungspraxis vor allem auf die Einstellungen und Verhaltensweisen von Mitarbeiterinnen und Mitarbeitern aus (Wang, Waldman & Zhang, 2014), etwa hinsichtlich der Identifikation von Lehrerinnen und Lehrern mit ihren Schulen (Hulpia, Devos & van Keer, 2009). Hierüber wirkt sie auf die Schul- und Unterrichtsentwicklung (Harris, 2008) und letztlich auf die Lernentwicklung von Schülerinnen und Schülern (Leithwood & Mascall, 2008).

2.2 Wie wirken Schulleitungen auf die Unterrichtsentwicklung?

Schulleitungen wirken sowohl direkt als auch indirekt auf Unterricht und Schülerleistungen (Scheerens, 2012). Einerseits können Schulleitungen direkt Einfluss auf den Unterricht nehmen, indem sie instruktional und somit unterrichtsbezogen führen (Pietsch et al., 2016). Andererseits zeigen Leithwood et al. (2008; 2006; 2002), dass Effekte Transformationaler Führung auf die Veränderung des Unterrichts in der Regel vermittelt erfolgen. Dabei können Schulleitungen sowohl die Arbeitsbedingungen als auch emotionale und motivationale Faktoren aufseiten von Lehrkräften relativ stark und direkt beeinflussen, wohingegen sie nur wenig Einfluss auf deren Innovationskapazität haben.

Von besonderer Relevanz für eine positive Gestaltung und Entwicklung des Unterrichts ist es dabei, ob Lehrerinnen und Lehrer daran glauben, dass sie in der Lage sind, mit Innovationen und Veränderungen umzugehen und befähigt werden, den eigenen Unterricht auch tatsächlich weiter entwickeln zu können, also ihre unter-

richtsbezogenen Selbstkonzepte, Selbstwirksamkeitserwartungen und das Vertrauen in die eigenen Fähig- und Möglichkeiten – die so genannte Innovationskapazität (Leithwood et al., 2002).

Dabei kann das Schulleitungshandeln in der Regel etwa 20 bis 30 Prozent der Variation in der Unterrichtsgestaltung durch Lehrkräfte bzw. der Unterrichtsentwicklung aufklären (Leithwood & Jantzi, 2006; Pietsch et al., 2016; Thoonen et al., 2011).

3 Präzisierung der Fragestellung

Es ist bislang unklar, in wie weit die Nutzung extern gewonnener Daten den Einfluss des Führungshandelns auf die Unterrichtsentwicklung moderiert. Zwar zeigen verschiedene Untersuchungen, dass z. B. die Führung einer Schulleitung einen empirisch nachweisbaren direkten Effekt auf die Datennutzung von Lehrkräften haben kann (z. B. Stump et al., 2016) und dass organisationale Faktoren wie die Kooperation im Kollegium oder die Nutzung von Daten sowohl positiv als auch negativ beeinflussen und Bemühungen seitens der Schulleitung hinsichtlich einer datengestützten Schul- und Unterrichtsentwicklung insofern einerseits fördern andererseits aber auch konterkarieren können (z. B. Levin & Datnow, 2012). Gleichwohl fehlen bislang Studien, die dies mit Blick auf die inspektionsbasierte Unterrichtsentwicklung untersuchen und in der darüber hinaus die Effekte verschiedener Führungsstile simultan, also in einem integrativen Modell, geprüft werden.

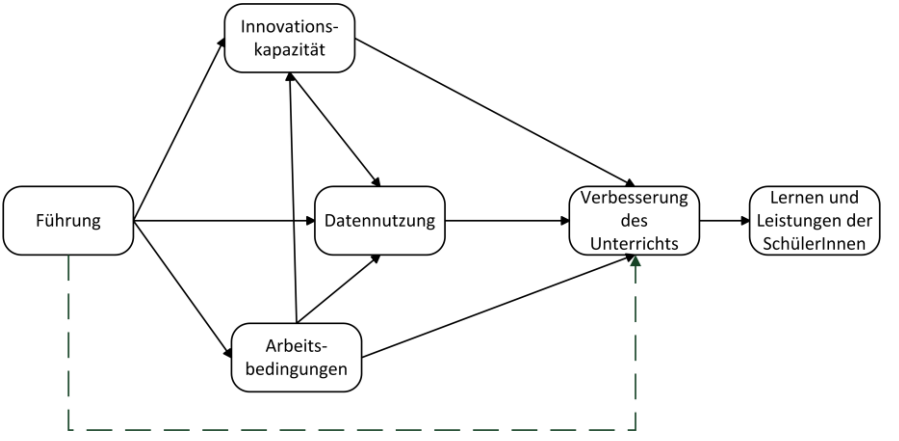


Abbildung 1: Theoretisches Modell zur Beschreibung der Zusammenhänge von Schulleitungshandeln, Organisationsmerkmalen, Merkmalen von Lehrkräften, Datennutzung und Unterrichtsentwicklung

Daher wird im Folgenden das vorgestellte Modell von Leithwood und Kollegen (s. Abb. 1), analog zum Vorgehen von Thoonen et al. (2011), um die Variable „Datennutzung" erweitert, um zu prüfen, ob, und falls ja, welche Rolle die Nutzung von Daten aus der Schulinspektion durch Lehrkräfte bei der schulleitungsgesteuerten Unterrichtsentwicklung spielt.

In diesem Modell wird einerseits angenommen, dass das Schulleitungshandeln bzw. Instruktionale, Transformationale und Kollaborative Führung durch Schulleitung direkt sowohl auf die Nutzung von Inspektionsdaten als auch die Unterrichtsentwicklung wirkt. Andererseits wird erwartet, dass auch vermittelte Effekte sichtbar werden, wobei diese sowohl durch die Innovationskapazität der Lehrkräfte als auch die Rahmenbedingungen, unter denen Datennutzung und Unterrichtsentwicklung erfolgen, moderiert werden.

4 Anlage und Durchführung der Untersuchung

4.1 Datengrundlage

Die Untersuchung nutzt Daten, die im Rahmen des Projektes „Evaluation der Impulswirkung von Schulinspektionen und Vergleichsarbeiten auf die Qualitätsentwicklung an Schulen" (EISVQS) erhoben wurden. Im Rahmen des Projektes untersuchte das Zentrum für Empirische Pädagogische Forschung (zepf) im Auftrag des niedersächsischen Kultusministeriums u. a. inwiefern die Schulinspektion und die Rückmeldungen aus Vergleichsarbeiten tatsächlich intendierte Veränderungen an Schulen evozieren. Das Projekt umfasste zwei online durchgeführte Befragungen an insgesamt 100 niedersächsischen Realschulen, Gymnasien und Integrierten Gesamtschulen der Sekundarstufe I, bei denen zwischen November 2014 und Januar 2016 eine Schulinspektion erfolgte. Zur freiwilligen Befragung waren die Leitungen sowie alle Lehrkräfte der jeweiligen Schulen eingeladen. Diese Untersuchung nutzt nur Daten der n = 933 Lehrkräfte (an 82 Schulen), die an der zweiten Befragung (März bis November 2016) teilgenommen haben. Die Daten stammen von 271 Lehrkräften an 27 Gymnasien, 460 Lehrkräften an 29 Realschulen und 202 Lehrkräften an 26 Integrierten Gesamtschulen. Im Mittel unterrichten die Lehrkräfte zum Zeitpunkt der Befragung seit 13.18 Jahren (SD = 10.55).

4.2 Variablen

Insgesamt werden 14 Items des Fragenbogens für die nachfolgende Modellierung genutzt, wobei mit Blick auf das Führungsverhalten der Schulleitungen Instruktionale und Transformationale Führung unterschieden werden. Mit Blick auf die Arbeitsbedingungen der Lehrerinnen und Lehrer werden Skalen zur Kooperation im

Kollegium sowie zur Kollaborativen Führung bzw. zu Steuerungsentscheidungen genutzt. Alle weiteren Dimensionen wurden durch einzelne Skalen oder Einzelitems erfasst.

Die *Instruktionale Führung* wurde dabei mithilfe dreier selbstkonstruierter Items erfasst, die, in Anlehnung an die Principal Instructional Management Rating Scale (PIRMS, Hallinger, 1994), die Stimulation, Supervision und das Monitoring von Lehr- und Lernprozessen (Managing the Instructional Programm, Beispielitem: „An unserer Schule ist die Schulleitung stetig damit beschäftigt, die pädagogische Arbeit voranzutreiben.") erfassen. McDonalds Omega der Skala liegt bei .87.

Die *Transformationale Führung* wurde mithilfe zweier Items erfasst, die der deutschsprachigen Fassung des Multifactor Leadership Questionnaire (MLQ, Bass & Avolio, 2000) entlehnt wurden und die Subdimensionen Inspirational Motivation und Individualized Consideration indizieren (Beispielitem: „Die Schulleitung formuliert eine überzeugende Zukunftsvision."). Erfasst wird somit, ob eine Schulleitung motivierende Leitbilder vorgibt und ob es ihr gelingt, die Entwicklungsbedarfe der Lehrkräfte zu erkennen und deren Potentiale systematisch zu fördern. McDonalds Omega der Skala liegt bei .76.

Die *Kooperation im Kollegium* wurde anhand von vier selbstkonstruierten Items erfasst. Diese fokussieren die wahrgenommene Qualität der Zusammenarbeit und ähneln damit dem Konzept der *Team Psychological Safety* (Edmondson, 1999), das als eine Voraussetzung für tiefere Formen der Kooperation angesehen werden kann (Gräsel, Fußangel & Pröbstel, 2006). (Beispielitem: „Spannungen und Konflikte unter den Lehrkräften werden gut gelöst.") McDonalds Omega der Skala liegt bei .88.

Bei der Skala *Kollaborative Führung* handelt es sich um eine Eigenkonstruktion, die in Anlehnung an ein Instrument von Wahlstrom und Louis (2008) die Teilhabe von Lehrkräften an der Steuerung der Schule, im Sinne eines *Shared Leadership*, mithilfe zweier Items erfasst (Beispielitem: „Die Schulleitung sorgt für eine umfassende Beteiligung, wenn Entscheidungen zur Schulentwicklung anstehen."). Die Skala misst somit, ob eine Schulleitung bestrebt ist, ein Arbeitsumfeld zu schaffen, in dem die Mitarbeiterinnen und Mitarbeiter durch Beteiligung am Steuerungshandeln Initiative zeigen und Verantwortung übernehmen können. McDonalds Omega der Skala liegt bei .82.

Mithilfe von Einzelitems wurden die Variablen *Innovationskapazität der Lehrkraft* (Item: „Ich bin mir sicher, dass ich die Qualität meines Unterrichts weiter verbessern kann."), *Nutzung von Inspektionsdaten durch die Lehrkraft* (Item: „Ich habe mich umfassend mit den Ergebnissen der Schulinspektion beschäftigt.") und, als abhängige Variable, die erlebte und berichtete *Verbesserung der Unterrichtsqualität* infolge

der Schulinspektion (Item: „Wenn Sie auf die Zeit nach der Schulinspektion zurückblicken, in welchen Bereichen sehen Sie Verbesserungen an Ihrer Schule, die Sie auf die Schulinspektion zurückführen können? … Unterrichtsqualität") erhoben.

Die hier genutzten Variablen konnten durch die befragten Lehrkräfte auf einer vierstufigen Skala von „trifft überhaupt nicht zu" bis „trifft völlig zu" beurteilt werden. Darüber hinaus gab es die Möglichkeit zu vermerken, wenn eine Beurteilung nicht möglich war (für die Analysen wurden diese Fälle als Missings kodiert).

4.3 Analysenmethoden

Die Berechnung des Strukturgleichungsmodells erfolgte mithilfe der Software MPLUS 7.3 (Muthén & Muthén, 2014). Zum Umgang mit fehlenden Werten wurde ein Maximum Likelihood-Schätzer genutzt. Da die befragten Lehrkräfte in den jeweiligen Schulen hierarchisch geschachtelt resp. genestet sind, wurden robuste Standardfehler mithilfe der Prozedur *Complex* ermittelt. Zur Bestimmung der Modellparameter wurde ein robuster Maximimum Likelihood-Schätzer (MLR; Yuan & Bentler, 2000) eingesetzt. Um den Modellfit bewertend einordnen zu können, werden klassische Fit-Maße herangezogen. Demnach kann von einer guten Modellanpassung in etwa bei CFI > .90, RMSEA < .06 und SRMR < .08 ausgegangen werden (Hu & Bentler, 1999; vgl. Marsh, Hau & Wen, 2004).

5 Ergebnisse

5.1 Deskriptiva

Tabelle 1 gibt die deskriptiven Kennwerte der Skalen bzw. Items wieder. Es wird deutlich, dass für alle unabhängigen und vermittelnden Variablen die Mittelwerte über dem jeweiligen theoretischen Mittelpunkt der Skala (2.5) liegen, d. h. dass die Lehrkräfte den entsprechenden Aussagen eher zustimmen als sie ablehnen. Dies gilt jedoch nicht für die abhängige Variable, Verbesserung der Unterrichtsqualität, denn der Mittelwert liegt unter dem theoretischen Skalenmittel. Zugleich weist dieses Item die größte Streuung auf, d. h. dass sich die Qualität des Unterrichts in Folge der Schulinspektion verbessert hätte, wird von den Lehrkräften mehrheitlich eher verneint, aber dies ist keine von allen geteilte Einschätzung und variiert vergleichsweise stark.

Die in Tabelle 2 dargestellten Korrelationen verdeutlichen darüber hinaus, dass die drei erfassten Facetten von Führung (Instruktionale, Transformationale, Kollaborative Führung) enger miteinander zusammenhängen (r > .6), als die restlichen Konstrukte (r < .45). Der stärkste bivariate Zusammenhang mit dem Kriterium „Verbesserung der Unterrichtsqualität" ergibt sich für die Kollaborative Führung (r = .306), der niedrigste überraschenderweise für die Kooperation (r = .013). Der

Befund passt nicht gut zum aktuellen Trend, Kooperation als eine zentrale Gelingensbedingung im Kontext der Schulentwicklung aufzufassen (Richter & Pant, 2016; Vangrieken, Dochy, Raes & Kyndt, 2015). Auffällig ist zudem, dass Transformationale und Kollaborative Führung nur wenig spezifische Varianz aufweisen. Angesichts einer latenten Korrelation von r > .9 mag es fraglich erscheinen, ob die Separation der Konstrukte gerechtfertigt ist, aber mit Blick auf die möglichst differenzierte Betrachtung der verschiedenen Facetten des Führungshandelns und der zumindest in der Literatur möglichen klaren Abgrenzung der Konstrukte, sehen wir von einer Zusammenlegung der beiden Facetten in diesem Beitrag ab.

Tabelle 1: Mittelwerte (M), Standardabweichungen (SD) und Anzahl der berücksichtigten Fälle (n) für die eingesetzten Skalen bzw. Items

	M	SD	n
Instruktionale Führung	3.13	0.69	811
Transformationale Führung	2.87	0.77	808
Kooperation	3.29	0.51	839
Kollaborative Führung	2.93	0.76	814
Innovationaskapazität	3.26	0.58	826
Nutzung von Inspektionsdaten	2.87	0.86	920
Verbesserung der Unterrichtsqualität	2.15	0.89	674

Tabelle 2: Latente Korrelationen zwischen den untersuchten Konstrukten

	Konstrukt	2	3	4	5	6	7
1	Instruktionale Führung	.653	.407	.608	.161	.108	.256
2	Transformationale Führung		.474	.903	.141	.147	.272
3	Kooperation			.448	.183	.074	.013
4	Kollaborative Führung				.137	.143	.306
5	Innovationskapazität					.119	.089
6	Nutzung von Inspektionsdaten						.235
7	Verbesserung der Unterrichtsqualität						

5.2 Direkte Effekte des Führungshandelns auf die inspektionsinduzierte Unterrichtsentwicklung

Um zu evaluieren, ob, und falls ja welche Effekte das Schulleitungshandeln auf die Unterrichtsentwicklung infolge einer Schulinspektion hat, wurde in einem ersten Schritt das oben dargestellte logische Wirkmodell in Form eines Strukturgleichungsmodells geprüft (s. Abb. 2). Hierfür wurde in Anlehnung an Thoonen et al. (2011) davon ausgegangen, dass neben dem Führungsstil der Schulleitung auch die Arbeitsbedingungen der Lehrkräfte sowie deren Innovationskapazität einen Einfluss auf den Umgang mit Schulinspektionsergebnissen sowie auf die Unterrichtsentwicklung haben können.

Die Gütekriterien für das Modell (χ^2 = 93.605, df: 59, RMSEA: .025, SRMR: .023, CFI: .990) weisen auf eine insgesamt sehr gute Passung hin. Dabei lassen sich rund 17 Prozent der Varianz in der inspektionsbasierten Unterrichtsentwicklung mithilfe der Modellvariablen erklären, jedoch nur drei Prozent der Unterschiede im Umgang mit Inspektionsergebnissen durch Lehrkräfte.

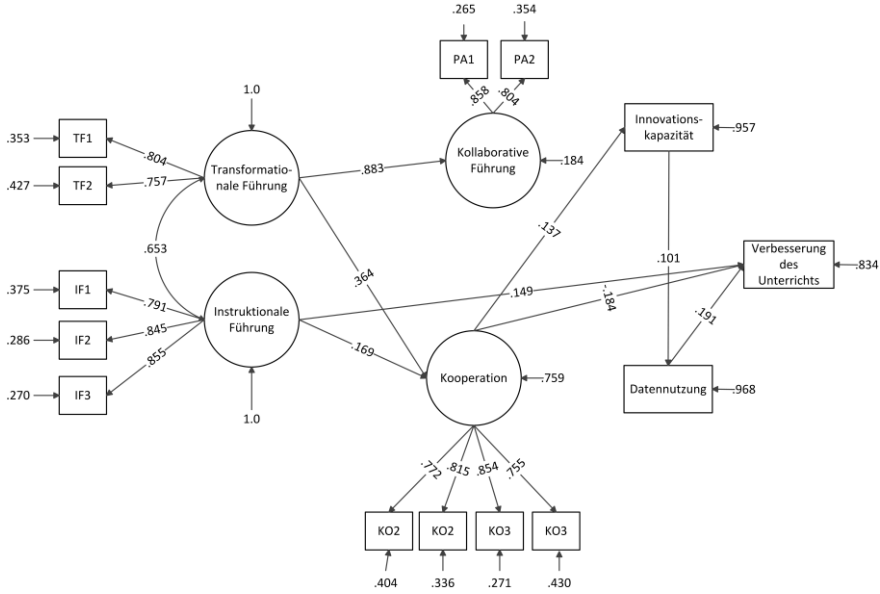


Abbildung 2: Strukturgleichungsmodell zur Beschreibung des Einflusses von Schulleitungshandeln auf den Umgang mit Inspektionsergebnissen von Lehrkräften sowie die inspektionsinduzierte Unterrichtsentwicklung

Einen direkten und empirisch nachweisbaren Einfluss darauf, ob Lehrerinnen und Lehrer sich intensiv mit den Ergebnissen aus einer Inspektion auseinandersetzen, hat ausschließlich deren Innovationskapazität (β = 0.101), also die Überzeugung, dass sie in der Lage sind, den eigenen Unterricht aus eigener Kraft weiter zu entwickeln.

Den größten direkten Einfluss darauf, ob infolge einer Schulinspektion der Unterricht an einer Schule weiter entwickelt wird, hat die Auseinandersetzung der Lehrkräfte mit den Inspektionsergebnissen (β = 0.191). Auch die Instruktionale Führung der Schulleitung hat einen nachweisbar positiv direkten Einfluss auf die Unterrichtsentwicklung (β = 0.149), während sich für die Kooperation im Kollegium ein negativer direkter Effekt (β = -0.184) findet. Hinsichtlich der Entwicklung des Unterrichts lassen sich für alle anderen Modellvariablen keine direkten Effekte nachweisen, die gegen den Zufall abgesichert sind.

5.3 Gesamteffekte des Führungshandelns auf die inspektionsinduzierte Unterrichtsentwicklung

Gleichwohl wird eine ausschließlich direkte Modellierung dem Untersuchungsgegenstand nur unzureichend gerecht, da davon ausgegangen werden muss, dass Schulleitungen im komplexen Gefüge von Schule und Unterricht auch indirekt Wirkung entfalten. Daher wurden in einem zweiten Schritt Gesamteffekte (total effects, die Summe der direkten und der indirekten Effekte) der endogenen sowie der vermittelnden Variablen mithilfe des MPLUS-Kommandos *MODEL INDIRECT* geschätzt. Dies hat den Vorteil, dass einerseits auch indirekte Effekte geprüft und nachgewiesen werden und dass andererseits auch kleine Effekte berücksichtigt werden können, die für sich alleine genommen nicht nachweisbar, in der Gesamtheit jedoch statistisch signifikant sein können.

Die in Tabelle 3 dargestellten Befunde machen deutlich, dass es, theoriekonform, für die inspektionsgestützte Unterrichtsentwicklung eines Zusammenspiels von Instruktionaler (β = 0.136) und Transformationaler (β = 0.183) Führung bedarf, wobei letztere ausschließlich indirekt zum Tragen kommt. Der negative Effekt der Kooperation im Kollegium auf die Unterrichtsentwicklung bleibt im Vergleich zum direkten Effekt (s. a. Abb. 2) mit einem β von -0.178 jedoch nahezu stabil. Auffällig ist weiterhin, dass die Auseinandersetzung der Lehrkräfte mit Inspektionsergebnissen von zwei Faktoren abhängt: Einen positiven Einfluss darauf hat einerseits, ob die Schulleitung im Sinne der Transformationalen Führung motivierende Leitbilder vorgibt und ob es ihr gelingt, die Entwicklungsbedarfe der Lehrkräfte zu erkennen und deren Potentiale systematisch zu fördern (β = 0.134). Andererseits spielt es auch eine Rolle, ob Lehrkräfte sich befähigt fühlen, mit Veränderung und Innovationen umzugehen bzw. diese in den eigenen Unterricht zu implementieren, wenn es

darum geht, dass sie sich intensiv mit den Inspektionsergebnissen auseinandersetzen (β = 0.101; vgl. Abb. 2).

Tabelle 3: Direkte, indirekte und Gesamteffekte (total effects) der Modellvariablen auf die Verbesserung der Unterrichtsqualität und die Nutzung der Inspektionsdaten

	Direkter Effekt		Indirekte Effekte		Gesamteffekt	
	β	p	β	p	β	p
Verbesserung der Unterrichtsqualität						
Instruktionale Führung	0.149	(.007)	-0.012	(.531)	0.136	(.014)
Transformationale Führung	-0.063	(.827)	0.246	(.315)	0.183	(.008)
Kooperation	-0.184	(.001)	0.006	(.553)	-0.178	(.001)
Kollaborative Führung	0.323	(.231)	0.011	(.791)	0.334	(.256)
Innovationskapazität	0.041	(.230)	0.019	(.013)	0.060	(.069)
Nutzung der Inspektionsdaten						
Instruktionale Führung	0.009	(.894)	0.012	(.332)	0.021	(.748)
Transformationale Führung	0.088	(.702)	0.046	(.812)	0.134	(.046)
Kooperation	-0.012	(.789)	0.014	(.055)	0.002	(.958)
Kollaborative Führung	0.050	(.823)	0.003	(.853)	0.053	(.813)

6 Diskussion und Fazit

Schulinspektionen richten sich in erster Linie an Schulleitungen. Diese sollen Rückmeldungen aus Inspektionen für eine datengestützte Entwicklung von Schule und Unterricht nutzen. Da Schulleitungen im Zuge der Einführung von Schulinspektionen, im Vergleich zu Lehrkräften und anderem schulischen Personal, strukturell massiv aufgewertet wurden, hängt die Wirkung bzw. der Effekt einer Inspektion in besonderem Maße von der Qualität der Führung einer Schule ab.

In der vorliegenden Studie wurde daher anhand einer Stichprobe von 82 Schulen aus Niedersachsen untersucht, welche Rolle Transformationale, Instruktionale und Kollaborative Führungpraktiken von Schulleitungen mit Blick auf die inspektionsinduzierte Unterrichtsentwicklung spielen.

Die Ergebnisse zeigen, dass die Entwicklung des Unterrichts infolge einer Schulinspektion vor allem dann zu erwarten ist, wenn Schulleitungen sowohl motivierende

Leitbilder vorgeben sowie die Entwicklungsbedarfe von Lehrkräften erkennen und deren Potentiale systematisch fördern als auch die Stimulation, Supervision und das Monitoring von Lehr- und Lernprozessen forcieren.

Während Instruktionale Führung mit Blick auf die Unterrichtsentwicklung vor allem direkt wirkt, wirkt Transformationale Führung in erster Linie vermittelt über die Kooperation im Kollegium, die Innovationskapazität der Lehrkräfte und die Bereitschaft der Lehrkräfte, sich mit Daten aus der Inspektion ausführlich auseinander zu setzen. Für die Teilhabe des Kollegiums an Steuerungsentscheidungen, die Kollaborative Führung, hingegen ließen sich keine Effekte nachweisen.

Auffällig ist darüber hinaus, dass Lehrerinnen und Lehrer sich vor allem dann mit Rückmeldungen aus Schulinspektionen auseinandersetzen, wenn eine Schulleitung motivierende Leitbilder für Schule und Unterricht vorgibt und die Lehrkräfte selber davon überzeugt sind, dass sie in der Lage sind, den eigenen Unterricht auch tatsächlich weiter entwickeln zu können.

Besonders augenfällig ist weiterhin die Rolle, die die Kooperation im Kollegium bei der inspektionsgestützten Unterrichtsentwicklung spielt. So zeigen sich diesbezüglich einerseits stark negative direkte Zusammenhänge – je stärker die Kooperation im Kollegium desto weniger wird der Unterricht infolge einer Inspektion weiterentwickelt. Andererseits hängt die Innovationskapazität der Lehrkräfte wiederum in besonderem Maße davon ab, wie gut in einer Schule zusammengearbeitet wird. Dass Kooperation im Kollegium nicht nur positive Effekt haben kann (Vangrieken et al., 2015), weil z. B. Lehrkräfte sich in ihrer Autonomie beschnitten fühlen, mehr Konkurrenz zwischen Kollegen wahrnehmen oder den Eindruck haben, dass abweichende Meinungen weniger toleriert werden, ist zwar bekannt (vgl. Richter & Pant, 2016), bezieht sich aber meist auf das Ausmaß oder die Art der Kooperation und nicht auf die wahrgenommene Qualität der Kooperation, wie sie in dieser Studie erfasst wurde. Eine mögliche Erklärung für den hier berichteten negativen Zusammenhang könnte sein, dass Lehrkräfte, die ausgeprägt positive Einschätzungen der Kooperation besitzen, der Meinung sind, dass an ihrer Schule eine Verbesserung der Unterrichtsqualität gar nicht notwendig sei und dazu passend berichten, dass keine Verbesserungen stattgefunden haben. Noch plausibler wird eine solche (wahrscheinliche) Überschätzung, wenn man annimmt, dass hohe berichtete Werte der Kooperation auch als Indikator für niedrige Ansprüche an die Qualität von Kooperationen verstanden werden können, weil solche niedrigen Ansprüche auch in Bezug auf die Unterrichtsqualität gelten könnten. Auch der sehr hohe Skalenmittelwert (vgl. Tab. 1) kann als Hinweis auf die Existenz selbstwertdienlicher Verzerrungen aufgefasst werden.

Interessant ist auch, dass die Kollaborative Führung keine nachweisbar signifikanten Effekte auf die Entwicklung des Unterrichts infolge einer Schulinspektion nach sich zieht, obwohl der empirisch nachweisbare Einfluss, gemessen an den Regressionsgewichten, durchaus groß ist. Dies könnte vor allem daran liegen, dass sich die Konstrukte des Shared und Transformational Leadership, wie sie im Rahmen der vorliegenden Studie genutzt werden, recht ähnlich sind. So konstatiert Avolio (2011, S. 51) "transformational leadership involves the process whereby leaders develop followers into leaders". Entsprechend geht eine ausgeprägte Transformationale Führung in der Regel mit einer hohen Beteiligung an Steuerungsentscheidungen einher (Wang et al., 2014), worauf auch die hohen Korrelationen der beiden Faktoren in den EISVQS-Daten hindeuten. Dass sich keine gegen den Zufall absicherbaren Effekte der Kollaborativen Führung finden, dürfte daher der geringen Trennschärfe der Konstrukte bzw. dem besonders hohen Anteil der gemeinsam geteilten Varianz geschuldet sein.

Gleichwohl spielen die organisationalen Voraussetzungen auf Ebene der Einzelschule den Befunden zufolge keine große Rolle dabei, wie intensiv sich Lehrkräfte mit Inspektionsergebnissen auseinandersetzen. Zwar geben die befragten Lehrerinnen und Lehrer an, dass sie sich durchaus mit den Ergebnissen der Schulinspektion beschäftigt haben – warum sie dies jedoch taten bzw. was dafür maßgeblich verantwortlich ist, konnte die vorliegende Untersuchung nicht hinreichend aufzeigen. Einzig ob eine Schulleitung motivierende Leitbilder vorgibt und ob es ihr gelingt, die Entwicklungsbedarfe der Lehrkräfte zu erkennen und deren Potentiale systematisch zu fördern, spielt diesbezüglich eine Rolle auf der Ebene der schulischen Rahmenbedingungen. Auf Individualebene wiederum sind ihre Überzeugung davon, ob sie in der Lage sind, den eigenen Unterricht aus eigener Kraft weiter zu entwickeln, relevant.

Die vorgelegten Befunde machen deutlich, dass Schulinspektionen auf komplexe innerschulische Bedingungsgefüge treffen und eine gelingende Unterrichtsentwicklung infolge einer Inspektion nicht selbstverständlich ist. Dabei ist es vor allem vom Führungshandeln einer Schulleitung abhängig, ob Daten aus einer Inspektion durch Lehrkräfte überhaupt rezipiert werden und in welchem Maße der Unterricht weiterentwickelt wird. Wichtig ist es hierbei, dass Schulleitungen einerseits aktiv steuernd in das Unterrichtsgeschehen an ihrer Schule eingreifen, andererseits aber auch den Lehrkräften eine attraktive Vision und klare Orientierung geben. Dazu müssen sie kommunizieren, was in Blick auf den Unterricht besser sein wird als bislang und wen sie diesbezüglich wie individuell fordern und fördern, kurz: die Schulleitungen müssen die Lehrerinnen und Lehrer trainieren und coachen und ihnen dadurch bei der Entwicklung und Entfaltung ihrer Potenziale und ihrer individuellen Stärken helfen.

Auch wenn weite Teile der hier berichteten Ergebnisse bereits gut dokumentierte Beziehungen betreffen, so unterliegt diese Untersuchung doch einer Reihe von Beschränkungen, die vor allem bei der Interpretation der unerwarteten Befunden berücksichtigt werden sollten. Dies betrifft z. B. die Qualität der Daten. Alle präsentierten Daten sind Selbstauskünfte der Lehrkräfte, unterliegen also potenziell einer ganzen Reihe von Verzerrungen. Darüber hinaus variieren die Rücklaufquoten zwischen den Schulen und liegen absolut betrachtet auf niedrigem Niveau. Dabei ist nicht davon auszugehen, dass die Selbstselektion der Studienteilnehmer completely at random erfolgt.

Die in dieser Untersuchung gefundenen, erwartungswidrigen Zusammenhänge zwischen Kooperation und Verbesserung des Unterrichts stellen einen interessanten Ausgangspunkt für weitere Analysen dar, in denen möglicherweise die oben angedeutete Hypothese geprüft werden könnte. Zusätzlich besteht die Möglichkeit, das hier untersuchte Wirkgefüge durch Informationen der zugehörigen Schulleitungen zu verfeinern und möglicherweise weitere Varianz der Kriteriumsvariable zu erklären. Darüber hinaus konnte insbesondere die Datennutzung durch Lehrkräfte im Rahmen der vorliegenden Studie nur unzureichend erklärt werden. Dies kann auch daran liegen, dass, im Gegensatz zu den Modellen von Leithwood et al., keine Motivationsvariablen mit in das Strukturgleichungsmodell aufgenommen werden konnten. In weiteren Studien wäre daher zu klären, welche Rolle Variablen wie affektives Commitment, Arbeitszufriedenheit und Motivation bei der Auseinandersetzung von Lehrkräften mit Inspektionsrückmeldungen spielen und welche weiteren Faktoren einen Einfluss darauf haben, dass Lehrerinnen und Lehrer Inspektionsergebnisse nutzen, um den Unterricht an ihrer Schule wissensbasiert weiter zu entwickeln.

Literatur

Anderson, S., Leithwood, K. & Strauss, T. (2010). Leading data use in schools: organizational conditions and practices at the school and district levels. Leadership and Policy in Schools, 9 (3), 292-327.

Avolio, B. J. (2011). Full range leadership development (2nd ed.). Thousand Oaks, Calif.: SAGE Publications.

Bass, B. M. & Avolio, B. J. (Eds.). (1994). Improving organizational effectiveness through transformational leadership. Thousand Oaks: Sage.

Bass, B. M. & Avolio, B. J. (2000). Mlq multifactor leadership questionnaire. Redwood City: Mind Garden.

Dedering, K. & Müller, S. (2008). Schulinspektion in Deutschland - Forschungsbereiche und -desiderata. In W. Böttcher, W. Bos, H. Döbert & H. G. Holtappels (Hrsg.), Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive: Dokumentation zur Herbsttagung der Kommission Bildungsorganisation, -planung, -recht (KBBB) (S. 241-252). Münster: Waxmann.

Denis, J.-L., Langley, A. & Sergi, V. (2012). Leadership in the plural. The Academy of Management Annals, 6 (1), 211-283.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44 (2), 350-383.

Feldhoff, T., Gromala, L. & Brüsemeister, T. (2014). Organisationales Lernen von Schulen im Kontext datenbasierter Steuerung. In H. G. Holtappels (Hrsg.), Schulentwicklung und Schulwirksamkeit als Forschungsfeld: Theorieansätze und Forschungserkenntnisse zum schulischen Wandel (S. 241-258). Münster: Waxmann.

Gräsel, C., Fußangel, K. & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen – eine Aufgabe für Sisyphos? Zeitschrift für Pädagogik, 52 (2), 205-219.

Hallinger, P. (1994). A resource manual for the principal instructional management rating scale (pirms manual 2.2). Nashville: Center for the Advanced Study of Educational Leadership.

Hallinger, P. (2003). Leading educational change: reflections on the practice of instructional and transformational leadership. Cambridge Journal of Education, 33 (3), 329-352.

Hallinger, P. (2011). A review of three decades of doctoral studies using the principal instructional management rating scale: a lens on methodological progress in educational leadership. Educational Administration Quarterly, 47 (2), 271-306.

Hallinger, P., Leithwood, K. & Heck, R. H. (2010). Leadership: instructional. In P. Peterson, E. Baker & B. McGaw (Eds.), International encyclopedia of education (pp. 18-25). Amsterdam: Elsevier.

Hallinger, P. & Murphy, J. (1985). Assessing the instructional management behavior of principals. The Elementary School Journal, 86 (2), 217-247.

Halverson, R., Grigg, J., Prichett, R. & Thomas, C. (2007). The new instructional leadership: creating data-driven instructional systems in schools. Journal of School Leadership, 17 (2), 159-194.

Harris, A. (2004). Distributed leadership and school improvement: leading or misleading? Educational Management Administration & Leadership, 32 (1), 11-24.

Harris, A. (2008). Distributed leadership: according to the evidence. Journal of Educational Administration, 46 (2), 172-188.

Heck, R. H. & Hallinger, P. (2010). Testing a longitudinal model of distributed leadership effects on school improvement. The Leadership Quarterly, 21 (5), 867-885.

Hu, L.-t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6 (1), 1-55.

Hulpia, H., Devos, G. & van Keer, H. (2009). The influence of distributed leadership on teachers' organizational commitment: a multilevel approach. The Journal of Educational Research, 103 (1), 40-52.

Leithwood, K., Aitken, R. & Jantzi, D. (2006). Making schools smarter: a system for monitoring school and district progress (3rd ed.). Thousand Oaks, California: Corwin.

Leithwood, K., Harris, A. & Hopkins, D. (2008). Seven strong claims about successful school leadership. School Leadership & Management, 28 (1), 27-42.

Leithwood, K. & Jantzi, D. (2005). A review of transformational school leadership research 1996-2005. Leadership and Policy in Schools, 4 (3), 177-199.

Leithwood, K. & Jantzi, D. (2006). Transformational school leadership for large-scale reform: effects on students, teachers, and their classroom practices. School Effectiveness and School Improvement, 17 (2), 201-227.

Leithwood, K., Jantzi, D. & Mascall, B. (2002). A framework for research on large-scale reform. Journal of Educational Change, 3, 7-33.

Leithwood, K. & Mascall, B. (2008). Collective leadership effects on student achievement. Educational Administration Quarterly, 44 (4), 529-561.

Levin, H. M. & Belfield, C. R. (2003). The marketplace in education. Review of Research in Education, 27 (1), 183–219.

Levin, J. A. & Datnow, A. (2012). The principal role in data-driven decision making: Using case-study data to develop multi-mediator models of educational reform. School Effectiveness and School Improvement, 23 (2), 179-201.

Marks, H. M. & Printy, S. M. (2003). Principal leadership and school performance: an integration of transformational and instructional leadership. Educational Administration Quarterly, 39 (3), 370-397.

Marsh, H. W., Hau, K.-T. & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. Structural Equation Modeling: A Multidisciplinary Journal, 11 (3), 320-341.

Marsh, J. A. (2012). Interventions promoting educators use of data: research insights and gaps. Teachers College Record, 114 (11), 1-47.

Marsh, J. A. & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making. Educational Management Administration & Leadership, 43 (2), 269-289.

Marzano, R. J., McNulty, B. A. & Waters, T. (2005). School leadership that works: from research to results. Alexandria, Va: Association for Supervision and Curriculum Development.

Muthén, L. K. & Muthén, B. O. (2014). Mplus [Computer software]. Los Angeles: Muthèn & Muthén.

Pietsch, M., Lücken, M., Thonke, F., Klitsche, S. & Musekamp, F. (2016). Der Zusammenhang von Schulleitungshandeln, Unterrichtsgestaltung und Lernerfolg. Zeitschrift für Erziehungswissenschaft, 19 (3), 527-555.

Preuß, B., Wissinger, J. & Brüsemeister, T. (2015). Einführung der Schulinspektion: Struktur und Wandel regionaler Governance im Schulsystem. In H. J. Abs, T. Brüsemeister, M. Schemmann & J. Wissinger (Hrsg.), Governance im Bildungssystem (S. 117-142). Wiesbaden: Springer Fachmedien Wiesbaden.

Richter, D. & Pant, H. A. (2016). Lehrerkooperation in Deutschland: Eine Studie zu kooperativen Arbeitsbeziehungen bei Lehrkräften der Sekundarstufe I. Verfügbar unter: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/lehrerkooperation-in-deutschland/ [12.07.2017].

Robinson, V. M. J., Lloyd, C. A. & Rowe, K. J. (2008). The impact of leadership on student outcomes: an analysis of the differential effects of leadership types. Educational Administration Quarterly, 44 (5), 635-674.

Scheerens, J. (Ed.). (2012). School leadership effects revisited: review and meta-analysis of empirical studies. Dordrecht: Springer Netherlands.

Schildkamp, K. & Lai, M. K. (2013). Conclusions and a data use framework. In K. Schildkamp, M. K. Lai & L. Earl (Eds.), Data-based decision making in education: challenges and opportunities (pp. 177-191). Dordrecht: Springer.

Seashore Louis, K., Leithwood, K., Wahlstrom, K. L. & Anderson, S. E. (2010). Investigating the links to improved student learning: learning from leadership project. Final report of research Findings. University of Minnesota; University of Toronto.

Stump, M., Zlatkin-Troitschanskaia, O. & Mater, O. (2016). The effects of transformational leadership on teachers' data use. Journal for Educational Research Online, 8 (3), 80-99.

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D. & Geijsel, F. P. (2011). How to improve teaching practices. Educational Administration Quarterly, 47 (3), 496-536.

Van Geel, M., Visscher, A. J. & Teunis, B. (2017). School characteristics influencing the implementation of a data-based decision making intervention. School Effectiveness and School Improvement, 118 (9), 1-20.

Vangrieken, K., Dochy, F., Raes, E. & Kyndt, E. (2015). Teacher collaboration: a systematic review. Educational Research Review, 15, 17-40.

Visscher, A. J. & Coe, R. (2003). School performance feedback systems: conceptualisation, analysis, and reflection. School Effectiveness and School Improvement, 14 (3), 321-349.

Wahlstrom, K. L. & Louis, K. S. (2008). How teachers experience principal leadership: the roles of professional community, trust, efficacy, and shared responsibility. Educational Administration Quarterly, 44 (4), 458-495.

Wang, D., Waldman, D. A. & Zhang, Z. (2014). A meta-analysis of shared leadership and team effectiveness. The Journal of applied psychology, 99 (2), 181-198.

Yuan, K.-H. & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology, 30 (1), 165-200.

From school inspection to improvement of teaching: Principal's influence

This article investigates the effects of different facets of school leadership (transformational, instructional and collaborative leadership) on the improvement of teaching following an inspection with respect to the innovation capacity of teachers, the quality of within school collaboration and use of feedback from school inspection. Analyses are based on data from n=933 teachers within 83 Lower-Saxony schools by applying a structural equation which has previously been designed and employed by Leithwood et al. (2006; 2002; 2008). The empirical model fits very well. The results show that transformational leadership has the biggest total effect on the improvement of teaching following an inspection, although no direct effects on the dependent variable are detectable. The effects of instructional leadership are somewhat smaller, but direct effects are statistically significant. Furthermore, the extent in which data from an inspection is used by teachers predicts the improvement of teaching after an inspection. The negative effect of teacher collaboration on the improvement of teaching is unexpected.

Keywords: effectiveness – instructional leadership – school improvement – school inspection – transformational leadership

Autoren

Dr. Marcus Pietsch, Leuphana Universität Lüneburg,
Prof. Dr. Ingmar Hosenfeld, Universität Koblenz-Landau, Zentrum für Empirische Pädagogische Forschung (zepf), Landau.
Korrespondenz an: pietsch@leuphana.de

Z Erziehungswiss
DOI 10.1007/s11618-016-0692-4

SCHWERPUNKT

ZfE

Der Zusammenhang von Schulleitungshandeln, Unterrichtsgestaltung und Lernerfolg

Eine argumentbasierte Validierung zur Interpretier- und Nutzbarkeit von Schulinspektionsergebnissen im Bereich Führung von Schulen

Marcus Pietsch · Markus Lücken · Franziska Thonke · Stefan Klitsche · Frank Musekamp

Zusammenfassung Qualitätsindikatoren und Messinstrumente, die bei Schulinspektionen Verwendung finden, werden in der Regel aus Qualitätsrahmen oder -tableaus abgeleitet. Die dort beschriebenen Aspekte von Schulqualität rekurrieren meist auf Befunde aus der Forschung zur Effektivität von Schule und Unterricht. Es wird angenommen, dass ein Zusammenhang zwischen den gemessenen Prozessmerkmalen und Schülerleistungen besteht. Diese Annahme ist allerdings im Rahmen von Schulinspektionen wenig untersucht worden, empirische Belege stehen aus. Im Beitrag wird anhand einer Stichprobe von n = 37 Schulen, n = 1663 Lehrkräften und der Leistungsentwicklung von n = 23.943 Schülerinnen und Schülern geprüft, wie valide Schulinspektionsergebnisse im Bereich *Führung von Schulen* interpretiert und als Grundlage für eine mögliche Schulentwicklung genutzt werden können. Zum Einsatz kommt dabei der Ansatz der argumentbasierten Validierung nach Kane. Die Befunde zeigen, dass Schulleitungen an wiederholt leistungsstarken Schulen häufig

Dr. M. Pietsch (✉)
Institut für Bildungswissenschaft, Leuphana Universität Lüneburg, Scharnhorststraße 1, 21335 Lüneburg, Deutschland
E-Mail: pietsch@leuphana.de

Dr. M. Lücken · F. Thonke · S. Klitsche · Dr. F. Musekamp
Institut für Bildungsmonitoring und Qualitätsentwicklung, Beltgens Garten 25, 20537 Hamburg, Deutschland

Dr. M. Lücken
E-Mail: markus.luecken@ifbq.hamburg.de

F. Thonke
E-Mail: franziska.thonke@ifbq.hamburg.de

S. Klitsche
E-Mail: stefan.klitsche@ifbq.hamburg.de

Dr. F. Musekamp
E-Mail: frank.musekamp@ifbq.hamburg.de

Published online: 30 August 2016

instruktional führen. Deutlich wird jedoch auch, dass sie dabei auch grundsätzlich anders führen als Schulleitungen an Schulen mit geringerer Leistungsperformanz, denn Schulleitungshandeln erfolgt vor allem situiert. Die Validierung zeigt, dass Bewertungen und Verallgemeinerungen im Bereich Führung von Schulen im Rahmen von Schulinspektionen durchaus möglich sind. Die Extrapolation dieser Bewertungen, also eine Verallgemeinerung und Übertragung auf Realsituationen, sowie die Ableitung zielgerichteter Entscheidungen zur Schulentwicklung jedoch eher nicht. Die Analysen verdeutlichen dabei auch, dass es im Rahmen von Schulinspektionen nicht hinreichend ist, ausschließlich Stärken und Schwächen an Schulen zurück zu melden, um Schulentwicklung zu stimulieren.

Schlüsselwörter Schulinspektion · Schulleitung · Schülerleistungen · Unterrichtsqualität · Validierung

The Relation of School Leadership, Instructional Quality and Student Achievement

An argument based validation study on the interpretations and uses of school inspection results regarding school leadership

Abstract Indicators and instruments for determining and measuring school quality within inspection systems are usually based on frameworks for inspection. These frameworks rely heavily upon school and teacher effectiveness research. Thus, a central assumption is that the effectiveness and improvement-oriented school conditions, as measured within an inspection, are related to student achievement. It is unclear if this assumption really holds true, as empirical evidence is still lacking. This study uses data from a random sampling of schools (n = 37) and teachers (n = 1663) and the achievement data from students (n = 23,943) to validate the interpretations and uses of school inspection results regarding the factor *school leadership*. The study follows Kane's argument-based approach for validation. Results reveal that principals of schools with recurrent high student achievement very often demonstrate instructional leadership. It is evident that these principals are also leading in a fundamentally different way from principals in schools with lower achievement in that they lead with the specific school context in mind. The study demonstrates that it is possible to make inferences from scores provided by school inspections and to generalize from them. However, this generalizability does not extend to making extrapolations or decisions based on these scores. The analyses make it clear that providing feedback solely on the strengths and weakness of a school is insufficient when it comes to stimulating school improvement through inspection.

Keywords Instructional Quality · School Inspection · School Leadership · Student Achievement · Validation

1 Einleitung

Effektive Schulinspektionen stimulieren Schulentwicklung und steigern Schülerleistungen. Dies soll gelingen, indem sie Schulen durch die Rückmeldung beobachteter Stärken und Schwächen Impulse für die Weiterentwicklung von Schule und Unterricht liefern. Hierfür inspizieren Schulinspektoren Schulen nach externen Vorgaben. Es wird erwartet, dass Schulverantwortliche die zurückgemeldeten Informationen für eine wissensbasierte Schul- und Unterrichtsentwicklung nutzen und dies sich wiederum in verbesserte Schülerleistungen niederschlägt.

Grundlage für die externe Einzelschulevaluation bilden dabei Qualitätsrahmen oder Qualitätstableaus, die Anforderungen an Schulqualität definieren und auf Grundannahmen der Schul- und Unterrichtseffektivitätsforschung rekurrieren (vgl. Ehren und Scheerens 2015). Basierend auf diesen Qualitätskatalogen wurden in allen Inspektionen Leistungsindikatoren definiert, die die Qualität einzelner Prozesse möglichst differenziert erfassen sollen (vgl. Böttcher und Kotthoff 2010). Die Datenerhebung selbst erfolgt über ein umfangreiches Repertoire an Methoden. Fragebögen und Interviews, die sich an verschiedene Schulbeteiligte richten, Schulbegehungen und die Begutachtung schulinterner Dokumente und Statistiken gehören derzeit ebenso wie die Beobachtung von Unterrichtseinheiten zum länderübergreifenden Methodenstandard der Schulinspektion (vgl. Döbert et al. 2008). Abschließend werden diese Informationen (statistisch oder heuristisch) zusammengefasst und im Sinne eines Stärken-Schwächen-Profils an Schulen zurückgemeldet (vgl. Gärtner und Pant 2011).

Gleichwohl ist derzeit nahezu unbekannt, ob die im Rahmen von Schulinspektionen eingesetzten Instrumente und Verfahren mit Blick auf ihre theoretischen Annahmen haltbar, hinsichtlich der Mess- und Bewertungsmodelle belastbar und aus der zugrunde liegenden Schuleffektivitätsforschung bekannte Zusammenhänge mit verschiedenen Außenkriterien in der Inspektionspraxis tatsächlich empirisch nachweisbar sind (vgl. Gärtner und Pant 2011; Ehren und Pietsch 2016). Vollkommen unklar ist darüber hinaus bislang, inwiefern das standardisierte Instrument der Schulinspektion überhaupt dazu geeignet ist, die Qualität von Schule und Unterricht in deren jeweilig spezifischen Kontexten zu bestimmen und zu bewerten (vgl. Gilroy und Wilcox 1997), und inwieweit die ebenfalls standardisierten Rückmeldungen von Inspektionsergebnissen eine „intentionskonforme Rekontextualisierung“ (Fend 2008, S. 28) und damit letztlich auch das intendierte Ziel einer zielgerichteten, evidenzbasierten Schulentwicklung durch die innerschulischen Akteure ermöglichen (vgl. Preuß et al. 2015).

Insofern schwingt bei Fragen nach der Wirksamkeit von Inspektionsverfahren stets auch die Frage nach deren Validität mit und dabei wird vor allem die Frage, wie valide die Interpretation und Nutzung von Informationen aus Schulinspektionen als Grundlage für das intendierte Ziel der Entwicklung von Schule und Unterricht überhaupt ist, in den Fokus gerückt. Woher weiß man also, ob Schulinspektionen und die Instrumente, die sie nutzen, überhaupt ein angemessenes und akkurates Abbild der mutmaßlich erhobenen Schulqualität wiedergeben? Erfassen die Standards, Indikatoren und Messinstrumente wirklich das, was gemessen werden soll und auch relevant ist (z. B. Schulqualität oder Unterrichtsqualität)? Und hat der Bericht der

ermittelten Stärken und Schwächen überhaupt das Potenzial, Handlungen auf Ebene der Einzelschule abzuleiten, die in der Konsequenz zu den erhofften Wirkungen führen?

Wie Gärtner und Pant (2011) betonen, ist die Bedeutung von validen Inspektionen insbesondere auch daher sehr hoch, da die Qualitätsentwicklung der Einzelschule zunehmend durch die Ergebnisse externer Evaluationen bestimmt wird. Nicht valide Inspektionen führen zu fehlerhaften Bewertungen und Urteilen, die wiederum zu fehlerhaften administrativen und/oder politischen Entscheidungen führen, die dann keine oder im schlimmsten Fall gar negative Effekte nach sich ziehen (vgl. Lane und Stone 2002; Ehren und Pietsch 2016).

Dabei meint Validität, den Standards für pädagogisches und psychologisches Testen von (1999 und 2014, S. 9) zufolge, „the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests". Wie Newton und Shaw (2014) hervorheben, kann der Begriff Test dabei im Sinne des Messvorgangs (*measurement procedure*) verstanden werden und umfasst sowohl dessen einzelne Elemente (*items*) sowie den Gesamttest als auch den Entscheidungsvorgang (*test use*) und die Testphilosophie (*testing policy*). Bei einer Validierung geht es demnach nicht nur darum, die Charakteristika von Instrumenten zu validieren, sondern insbesondere auch darum, die Angemessenheit sowie die Genauigkeit möglicher Schlussfolgerungen zu betrachten, die auf Basis der Ergebnisse gezogen werden können (vgl. Sireci und Sukin 2013). Dabei gilt es dann stets die folgenden beiden Fragen zu beantworten (Newton und Shaw 2014, S. 131):

1. Is the test any good as a measure of the characteristic that it purports to assess?
2. Should the test be used for its present purpose?

Mit Blick auf die Validierung von Schulinspektionen bedeutet dies letztlich (vgl. Pietsch et al. 2015), Argumente für und gegen die geplante Interpretation von Inspektionsergebnissen zu konstruieren und zu evaluieren. Dabei ist Validität keine Eigenschaft des Tests oder der Untersuchung, sondern bezieht sich vor allem auf deren Interpretationen. Entsprechend fassen Ehren und Pietsch (2016, S. 50) zusammen:

> *Validity in the context of school inspections entails the extent to which inspection frameworks, guidelines, protocols used in the assessment of schools are a good measure of the characteristic that it purports to assess (e. g. school quality), and whether these frameworks, guidelines and protocols should be used for its present purpose (e. g. control, improvement, liaison).*

Praktisch lässt sich Validität dann wiederum daran bemessen, in welchem Maße Evidenz und Theorie die Interpretation von Ergebnissen im Rahmen der Nutzungsabsicht unterstützen (vgl. Sireci und Sukin 2013). Neuere Validierungsansätze nutzen daher ein übergreifendes Konzept von Validität, das eine systematische Validierung mithilfe eines argumentationsbasierten Ansatzes praktikabel ermöglicht (vgl. Kane 2013). Zu prüfen sind bei jeder Validierung zumindest die folgenden vier Fragen:

1. Lässt sich beobachtetes Verhalten bewerten, um daraus Ergebnisse abzuleiten?
2. Lassen sich aus den beobachteten Ergebnissen Verallgemeinerungen ableiten?

3. Lässt sich aus den Verallgemeinerungen das Verhalten und/oder die Ergebnisse in anderen, nicht testimmanenten Situationen extrapolieren bzw. hochrechnen?
4. Lassen sich aus der Extrapolation weiterführende, zielgerichtete Entscheidungen ableiten?

Ist eine dieser Inferenzen nicht erfüllt, ist die gesamte Argumentationskette nicht haltbar. Auch Stärken in einem Bereich können die Schwächen in einem anderen Bereich nicht kompensieren. In einem solchen Fall ist dann davon auszugehen, dass die Annahme, die im Rahmen der Validierung geprüft werden soll, nicht haltbar ist.

Im Folgenden wird dieser Ansatz genutzt, um mithilfe von Daten der Schulinspektion Hamburg sowie Daten der Hamburger KERMIT (Kompetenzen ERMITteln)-Erhebungen zu prüfen, inwieweit Informationen, die im Rahmen von Schulinspektionen erhoben werden, geeignet sind, Schulentwicklung zu stimulieren. Fokussiert wird dabei auf den Aspekt des Schulleitungshandelns, da die Arbeit von Schulleitungen bei Schulinspektionen einerseits ein zentraler Aspekt der Evaluation ist (vgl. Döbert et al. 2008), Schulleitungen andererseits aber auch zentrale Adressaten von Schulinspektionsinformationen sind und daher auch maßgeblich die inspektionsbasierte Schulentwicklung steuern und beeinflussen (vgl. Preuß et al. 2015).

Entsprechend werden zuerst Annahmen und Befunde zur Wirkung sowie zur Wirkungsweise von Schulleitungen berichtet, die im Rahmen der Schuleffektivitätsforschung entwickelt wurden und entsprechend auch Eingang in die Arbeit von Schulinspektionen gehalten haben. Anschließend wird dann mithilfe von Strukturgleichungen ein theoretisches Wirkungsmodell geprüft, das den Einfluss von Schulleitungen auf die Unterrichtsgestaltung durch Lehrkräfte beschreibt. Danach wird überprüft, ob sich das Schulleitungshandeln an Schulen, an denen häufig hohe Lernzuwächse erreicht werden, von anderen Schulen unterscheidet. Abschließend werden die Ergebnisse mit Blick auf die Validität sowie die Effektivität von Schulinspektionsverfahren diskutiert.

2 Annahmen zur Wirkung von Schulleitungen auf Schülerleistungen

2.1 Führungsstile von Schulleitungen aus Sicht der Schuleffektivitätsforschung

Wie dargestellt, beziehen sich Schulinspektionen mit Blick auf die von ihnen evaluierten Prozessmerkmale von Schule und Unterricht in der Regel auf Annahmen und Befunde zur Effektivität von Schule und Unterricht (vgl. Ehren und Scheerens 2015). Ziel der Inspektionen ist es, Schulentwicklungsprozesse zu stimulieren, die wiederum in besseren Schülerleistungen münden. Entsprechend spielt die Erhebung und Darstellung von Ergebnissen zur Qualität der Arbeit von Schulleitungen im Rahmen von Schulinspektionsverfahren eine wichtige Rolle (vgl. z. B. Döbert et al. 2008; Diedrich 2015), gilt doch Schulleitungshandeln auf institutioneller Ebene gleich nach den ebenfalls bedeutsamen Merkmalen von Unterricht und Lehrkraft als zweitwichtigste Einflussgröße für erfolgreiches Lernen (vgl. Leithwood et al. 2008).

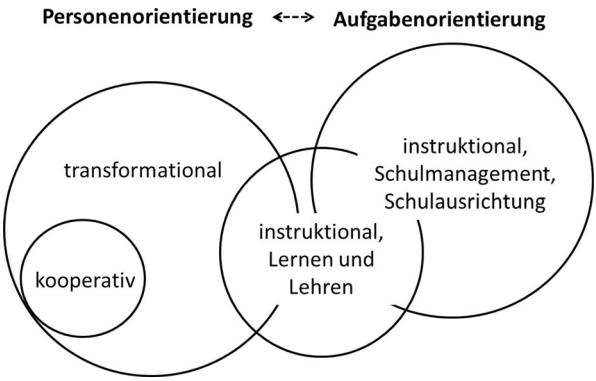


Abb. 1 Führungsstile in der Schuleffektivitätsforschung. (Nach Scheerens 2012)

Schulleitungen, die zum Bildungserfolg der Schülerinnen und Schüler an ihren Schulen beitragen, legen die Schwerpunkte ihrer Arbeit dabei vor allem darauf, 1) den Schulbeteiligten Wege und Ziele vorzugeben, 2) Mitarbeiterinnen und Mitarbeiter (weiter) zu entwickeln, 3) die Schule (neu) zu gestalten und 4) das Lernen und Lehren an der Schule aktiv zu steuern (vgl. Leithwood und Jantzi 2006; Scheerens 2012).

Erfolgreiche Schulleitungen sind dabei sowohl Führungsperson als auch Manager. Dabei versuchen sie durch Management, eine produktorientierte Ordnung und Berechenbarkeit im komplexen System Schule zu schaffen, und durch Führung, organisationale Veränderungen herbeizuführen und auf diesem Wege den Umgang mit besonderen und/oder neuen Herausforderungen zu ermöglichen (vgl. Kotter 1990).

Die Schuleffektivitätsforschung unterscheidet entsprechend zwei Konzepte voneinander: pädagogische bzw. instruktionale und transformationale Führung (vgl. Abb. 1; Hallinger 2003; Scheerens 2012). Instruktionale Führung umfasst vor allem Managementaspekte; sie erfolgt primär aufgaben- und produktorientiert, zielt auf die Optimierung vorhandener Strukturen und Prozesse ab und führt im Idealfall zu einer Verbesserung bereits vorhandener Prozesse und Mechanismen. Die Schulleitung kontrolliert und koordiniert entsprechend gezielt Aspekte des Schul- und Unterrichtsgeschehens, die den Lernfortschritt der Schülerinnen und Schüler betreffen, und nimmt direkten Einfluss auf den Unterricht und das Curriculum, z. B. durch die aktive Anleitung von Lehrkräften mittels Zielvorgaben, abgestimmten Fortbildungsmaßnahmen und der Evaluation von Schülerleistungen.

Transformationale Führung hingegen beinhaltet in der Regel Führungsaspekte, erfolgt meist mitarbeiterorientiert und zielt auf die nachhaltige Veränderung der schulischen Lern- und Arbeitskultur ab. Dieser Führungsstil soll daher primär zu innerschulischen Innovationen und Veränderungen führen und ist maßgeblich dadurch geprägt, dass die Schulleitung eine sinnstiftende Zukunftsvision für die Schule entwickelt, Lehrkräfte inspiriert und motiviert, einzelne Lehrerinnen und Lehrer gezielt unterstützt und fördert sowie ihnen intellektuelle Herausforderungen bietet.

2.2 Effekte von Schulleitungen auf Schülerleistungen

Betrachtet man diejenigen empirischen Studien, die seit der Jahrtausendwende einzelne Untersuchungen systematisch in Metaanalysen, systematischen Reviews oder einer Synthese zusammengefasst haben (vgl. Witziers et al. 2003; Marzano et al. 2005; Creemers und Kyriakides 2008; Hattie 2009; Scheerens 2012; Ehren und Scheerens 2015), wird deutlich, dass nur ein kleiner Teil der schülerseitigen Leistungsunterschiede durch den Einfluss von Schulleitungen aufgeklärt werden kann (vgl. zur Übersicht Pietsch 2014b).

Im Mittel liegt der Effekt, den Schulleitungen auf Schülerleistungen haben, demnach bei einer Stärke von Cohen's d = 0,20, was in etwa einer Varianzaufklärung von rund einem Prozent entspricht. Dabei variieren die Befunde zwischen den einzelnen Metastudien jedoch recht stark: Zwischen null (Witziers et al. 2003, Cohen's d = 0,04) und sechs (Marzano et al. 2005, Cohen's d = 0,56) Prozent der Unterschiede in den individuellen Schülerleistungen lassen sich den Metastudien zufolge durch Schulleitungshandeln erklären. Hattie (2009) kommt in seiner Synthese von Metaanalysen zu dem Ergebnis, dass sich rund drei Prozent (Cohen's d = 0,36) der Leistungsunterschiede aufseiten von Schülerinnen und Schülern auf wirksames Schulleitungshandeln zurückführen lassen.

Die Befundlage dazu, welcher Führungsstil eher Wirkung auf Schülerleistungen entfaltet, ist relativ eindeutig: So verdeutlichen verschiedene Studien zwar, dass transformationale Führung zu einer hohen Arbeitszufriedenheit, niedrigeren Krankenständen und einer Identifikation von Lehrkräften mit ihrer Schule führt (vgl. z. B. Judge und Piccolo 2004). Nichtsdestotrotz scheinen diese Faktoren mit dem Blick auf Schülerleistungen keine hinreichende Bedingung zu sein. Entsprechend zeigt eine Meta-Analyse von Robinson et al. (2008) ebenso wie die Synthese von Hattie (2009), dass instruktionale Führung die Chance auf überdurchschnittliche Lernergebnisse um rund 118 % (r = 0,21, d = 0,43, Odds Ratio = 2,18) steigert, wohingegen transformationale Führung diese Chance nur um gut 24 % (r = 0,06, d = 0,12, Odds Ratio = 1,24) erhöht.

Von herausragender Bedeutung für wirksames Schulleitungshandeln ist es entsprechend, dass Schulleitungen sich als Experten für Unterrichtsfragen verstehen. Geben sie erstens den Lehrkräften an der Schule ein elaboriertes Feedback zur Qualität des Unterrichts, stehen sie zweitens als Berater und Diskussionspartner für Unterrichtsfragen zur Verfügung und leiten sie drittens Lehrerinnen und Lehrer mit Blick auf die Gestaltung effektiven Unterrichtens gezielt an, erhöht dies die Chance von Schülerinnen und Schülern auf einen überdurchschnittlichen Lernzuwachs um ein Vielfaches (vgl. Marzano et al. 2005; Robinson et al. 2008; Shatzer et al. 2013).

2.3 Wirkungsweise von Schulleitungen auf Schülerleistungen

Lange Zeit wurde in der Schulforschung angenommen, dass Schulleitungen direkt und unvermittelt Einfluss auf die Lernzuwächse von Schülerinnen und Schülern nehmen – eine Annahme, die jedoch empirisch nicht haltbar ist (vgl. De Maeyer et al. 2007). Vielmehr wirken Schulleitungen auch indirekt, indem sie effizientes und effektives Lernen und Lehren an der Schule ermöglichen (vgl. Hallinger und Heck

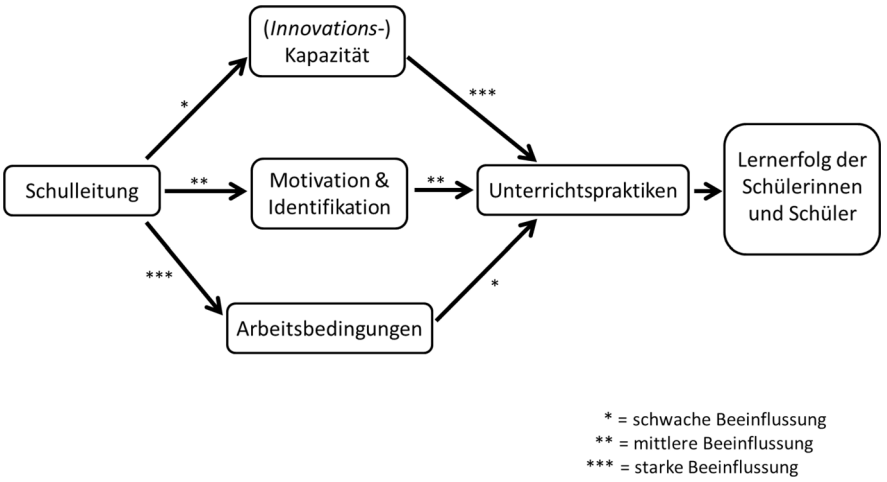


Abb. 2 Der Einfluss von Schulleitungen auf die Unterrichtsgestaltung. (Nach Leithwood et al. 2008)

1998). Relevant für den Lernerfolg von Schülerinnen und Schülern ist es dabei vor allem, dass seitens der Schulleitung einerseits Ziele und Visionen vorgegeben sowie innerschulische Strukturen und Prozesse optimiert werden. Und dass andererseits Lehrkräfte dazu befähigt werden, bestmöglich zu lehren, indem sie in die Lage versetzt werden, reflektiert, effizient sowie inhaltlich und methodisch auf der Höhe der Zeit zu unterrichten (vgl. Hallinger 2011).

Wirksames Schulleitungshandeln hat im Verständnis der Schuleffektivitätsforschung dabei in erster Linie das Ziel, den Unterricht an der Schule zu optimieren und zu verbessern. Und dies geschieht vor allem, indem Schulleitungen die Motivation und Identifikation von Lehrkräften mit ihrer Schule, deren (subjektiv erlebte) Arbeitsbedingungen sowie deren Kapazitäten (im Englischen *capacity beliefs*) beeinflussen, d. h. insbesondere ihre Fähigkeiten und ihr Wissen um wirksames Lernen und Lehren (vgl. Abb. 2, Leithwood et al. 2008). Dies sind allesamt Faktoren, die ihrerseits wiederum relevant für eine wirksame Unterrichtsführung sind.

Leithwood et al. (2002, 2008) zeigen, dass Schulleitungen in besonderem Maße die Arbeitsbedingungen von Lehrkräften beeinflussen können, sie also großen Einfluss darauf haben, ob und wie kooperiert wird, ob und wie Lehrkräfte an Steuerungsentscheidungen beteiligt werden. Auch haben Schulleitungen demnach einen großen Einfluss auf emotionale und motivationale Faktoren der Lehrkräfte, wie z. B. das affektive Commitment und die Arbeitszufriedenheit. Diese Faktoren sind wichtig, damit sie sich professionell weiterentwickeln und auch mit Veränderungen und Innovationen umgehen können. Nur bedingt Einfluss nehmen können Schulleitungen diesen Befunden zufolge hingegen auf die Innovationskapazitäten der Lehrkräfte, ihre Kompetenzen und Wissen sowie ihre Selbstwirksamkeitserwartungen. Interessant ist dann wiederum, dass sich die Effekte dieser vermittelnden Variablen auf die Unterrichtsgestaltung dem Literaturreview zufolge jedoch genau umgekehrt proportional verhalten: Arbeitsbedingungen haben einen geringen Einfluss auf die Unterrichtsgestaltung und Innovationskapazität hat wiederum einen großen Einfluss.

Dabei kann das Schulleitungshandeln bis etwa 30 % der Variation im Unterrichtshandeln aufklären (vgl. Leithwood und Jantzi 2006), wobei jedoch durchaus negative Effekte im Zusammenhang mit transformationaler Führung beobachtbar sein können (vgl. Thoonen et al. 2011). Zusammenfassend lässt sich feststellen, dass vor allem eine ausgewogene Mischung aus instruktionaler und transformationaler Führung (*integrated leadership*) besonders erfolgversprechend ist (vgl. Marks und Printy 2003) und darüber hinaus pluralisierende Führungsformen (z. B. *collaborative, distributed* oder *shared leadership*) die Effekte des Schulleitungshandelns, sowohl positiv als auch negativ, verstärken können (vgl. zur Übersicht Denis et al. 2012).

3 Klärung und Präzisierung der Forschungsfrage

Wie eingangs dargestellt, geht es bei der Validierung von Schulinspektionsverfahren stets darum, zu klären, ob es mithilfe des Instruments oder Teilen davon gelingt, a) diejenigen Aspekte zu messen, die intendiert und relevant sind und b) zu evaluieren, ob diese Befunde für das angestrebte Ziel überhaupt geeignet sind. Die zugrunde liegende Annahme lautet in diesem Fall: „*Schulinspektion führt, vermittelt über Schulentwicklung, zu besseren Schülerleistungen*“. Um diese Annahme mit Blick auf das Schulleitungshandeln zu stützen, müssen nun folgende Argumente geprüft werden: a – *Bewertung und Verallgemeinerung*) die oben dargestellten Modelle und Befunde lassen sich mithilfe von Inspektionsdaten empirisch replizieren, b – *Extrapolation*) ein Zusammenhang zwischen vermittelnden Variablen sowie Schülerleistungen und Schulleitungshandeln ist nachweisbar und c – *Entscheidungsfindung*) es lassen sich Unterschiede im Schulleitungshandeln zwischen Schulen mit unterschiedlichem Leistungsniveau finden, die darauf hindeuten, dass die Veränderung des Schulleitungshandels infolge einer Inspektionsrückmeldung sowohl mit einer veränderten Unterrichtspraxis als auch höheren Schülerleistungen einhergehen kann. Praktisch-konkret gilt es daher im Sinne Kanes (s. Abschn. 1) dabei argumentativ folgendes zu prüfen:

1. Lässt sich das im Rahmen der Schulinspektion erfasste Schulleitungshandeln mithilfe eines Vergleichs (z. B. sozial, kriterial, ipsativ) bewerten?
2. Lassen sich aus den beobachteten Ergebnissen zum Schulleitungshandeln Verallgemeinerungen ableiten? Lässt sich also ein Repräsentationsschluss ziehen?
3. Lässt sich aus den Verallgemeinerungen zum Schulleitungshandeln, wie es im Rahmen der Schulinspektion bestimmt wird, auf andere Kontexte wie z. B. Realsituationen schließen?
4. Lassen sich mithilfe der Informationen, die im Rahmen der Schulinspektion zum Schulleitungshandeln erhoben und berichtet werden, Entscheidungen treffen, die es ermöglichen die Schule und Unterricht zielgerichtet weiter zu entwickeln?

4 Methodisches Vorgehen

4.1 Datengrundlage

Die kommende Studie beruht auf Daten der Schulinspektion Hamburg sowie Daten der Hamburger KERMIT-Erhebungen. Die Schulinspektion Hamburg führt im Rahmen der Inspektion umfangreiche Datenerhebungen durch und befragt mithilfe von Onlinefragebögen an jeder Schule alle Schülerinnen und Schüler ab der dritten Jahrgangsstufe, alle Eltern sowie alle Lehrkräfte (vgl. Diedrich 2015). Sie orientiert sich dabei nicht nur inhaltlich an den Annahmen und Befunden der Schuleffektivitätsforschung (vgl. Behörde für Schule und Berufsbildung 2012), sondern nutzt im Rahmen der Befragung von Lehrkräften auch ausdrücklich das oben dargestellte Modell nach Leithwood et al. (2008) sowie erprobte Instrumente (vgl. Pietsch 2014a). Die Grundlage für die vorliegende Studie bildet eine Stichprobe von n = 37 Hamburger Sekundarschulen (davon 22 Gymnasien und 15 Stadtteilschulen), die von 2012 bis 2015 durch die Schulinspektion extern evaluiert wurden. Die weiteren Details werden im Folgenden dargestellt.

4.1.1 Prozessdaten: Operationalisierung des Referenzrahmens Schulqualität

Die oben vorgestellten Annahmen und Befunde zur Wirksamkeit von Schulleitungen haben Eingang in den Hamburger Orientierungsrahmen Schulqualität (vgl. Behörde für Schule und Berufsbildung 2012) gefunden. Der Hamburger Orientierungsrahmen beschreibt in 22 Qualitätsbereichen, die wiederum den drei Qualitätsdimensionen „Führung und Management“, „Bildung und Erziehung“ sowie „Wirkungen und Ergebnisse“ zugeordnet sind, das Idealbild guter Schule und damit die Ziele gelingender, erfolgreicher Schul- und Unterrichtsentwicklung. Dabei unterscheidet der Orientierungsrahmen zwischen a) Voraussetzungen, b) Merkmalen und Prozessen in Schulen und c) den Ergebnissen auf Seiten von Schülerinnen und Schülern. Ob eine Schule erfolgreich in ihrer Arbeit ist, bemisst sich dabei vor allem an einer Frage: Erwerben die Schülerinnen und Schüler diejenigen Kompetenzen, die sie zur Teilhabe an der Gesellschaft und zu einem erfolgreichen Berufsleben benötigen?

Skalen zur Erhebung von Führung und Management Wie eine Schulleitung führt, wird im Rahmen der Schulinspektion mithilfe von Mitarbeiterbefragungen eingeschätzt. Die Schulinspektion Hamburg greift hierfür auf erprobte Fragebögen zur Befragung von Lehrkräften zurück. Einerseits wird eine Kurzform des *Multifactor Leadership Questionnaire* (MLQ, vgl. Bass und Avolio 1995; Felfe 2006; Harazd und van Ophuysen 2011) zur Messung der transformationalen Führung und andererseits werden Skalen zur Messung der instruktionalen Führung aus dem *Teaching and Learning International Survey* (TALIS, vgl. Schmich und Schreiner 2008; OECD 2009, 2010) eingesetzt.

Das MLQ differenziert dabei die transformationale Führung in drei Stufen aus: transformational, transaktional und passiv-vermeidend resp. laissez-faire. Im Modell wird davon ausgegangen, dass diese einzelnen Stufen bestimmten Aktivitätsniveaus der Schulleitungen auf einem Kontinuum entsprechen, wobei mit zunehmender Ak-

tivität auch eine steigende Effektivität des Handelns erwartet wird. Diese Dreiteilung soll es ermöglichen, die gesamte Breite personenbezogener Führung (*full range of leadership*) von quasi nicht vorhanden (laissez-faire) über sachlich, rational und distanziert (transaktional) bis hin zu charismatisch, visionär und zugewandt (transformational) beschreib- und darstellbar zu machen.

Erfasst werden diese Subdimensionen mithilfe von insgesamt 21 Items, die auf einer vierstufigen Likert-Skala, die von 1 (nie) bis 4 ((fast) immer) reicht, beantwortet werden konnten. Dabei wurde die Subdimension „transformationale Führung" von neun (Beispielitem: „Die Schulleiterin/Der Schulleiter macht mich stolz darauf, mit ihr/ihm zu tun zu haben." Cronbach's Alpha = 0,93), die Subdimension „transaktionale Führung" durch sechs (Beispielitem: „Die Schulleiterin/Der Schulleiter spricht klar aus, was man erwarten kann, wenn die gesteckten Ziele erreicht worden sind.", Cronbach's Alpha = 0,75) und die Subdimension „laissez-faire" ebenfalls durch sechs (Beispielitem: „Die Schulleiterin/Der Schulleiter vertritt die Ansicht, dass Probleme erst wiederholt auftreten müssen, bevor man handeln sollte.", Cronbach's Alpha = 0,83) Items indiziert.

Zur Erhebung instruktionaler Führung werden durch die Schulinspektion wiederum Skalen genutzt, die im Rahmen der internationalen OECD-Untersuchung TALIS entwickelt und eingesetzt wurden. Erfasst wird dieser Führungsstil mithilfe von insgesamt neun Items, die auf einer vierstufigen Likert-Skala, die von 1 (selten oder nie) bis 4 (sehr oft) reicht, beantwortet werden konnten und ursprünglich aus dem Principal Instructional Management Rating Scale (PIMRS) von Hallinger (1994) stammen. Instruktionale Führung wird hierbei durch die drei Dimensionen *Ziele festlegen, Probleme lösen* und den *Unterricht entwickeln* beschrieben und im Rahmen der Inspektion mithilfe von insgesamt acht Items (Beispielitem: „Die Schulleiterin/Der Schulleiter stellt sicher, dass die Fortbildungsaktivitäten der Pädagoginnen und Pädagogen auf die Lehrziele abgestimmt sind.", Cronbach's Alpha = 0,91) erfasst, die eine Gesamtskala „instruktionale Führung" bilden.

Skalen zur Erhebung der Unterrichtsgestaltung Um international anschlussfähig zu sein, werden im Rahmen der schriftlichen Befragung von Lehrkräften durch die Hamburger Schulinspektion auch Skalen zur Unterrichtsgestaltung aus TALIS eingesetzt. Hierfür müssen Lehrerinnen und Lehrer auf einer Skala von eins bis vier angeben, wie häufig sie eine bestimmte Methode im Unterricht einsetzen (1 = nie oder fast nie, 4 = in ungefähr ¾ der Stunden). Die Gestaltung des Unterrichts wird dabei in drei Dimensionen erfasst: strukturorientierter Unterricht, schülerzentrierter Unterricht und erweiterte Unterrichtsaktivitäten. Dabei korrespondieren diese Skalen mit den drei Basisdimensionen guten Unterrichts Klassenführung (*strukturiert*), Schülerorientierung (*schülerzentriert*) und kognitive Aktivierung (*erweitert*, vgl. Vieluf et al. 2012), die sich in den letzten Jahren im Rahmen der Unterrichtsforschung international etabliert haben (vgl. z. B. Klieme und Rackoczy 2008; Pianta und Hamre 2009). Die erste Dimension bündelt daher fünf Items zur Messung von Lehreraktivitäten, in denen die Steuerung und Organisation des Unterrichts im Fokus steht (Beispielitem: „Ich gebe Lernziele explizit an.", Cronbach's Alpha = 0,51). Die zweite Dimension setzt sich aus vier Items zusammen und berücksichtigt vor allem den Bereich des schülerzentrierten Unterrichtens (Beispielitem: „Ich gebe

unterschiedliche Aufgaben an Schülerinnen und Schüler, die Lernschwierigkeiten haben, und/oder an diejenigen, die schneller vorankommen.“, Cronbach’s Alpha = 0,55)[1]. Die dritte Dimension beinhaltet, indiziert mit Hilfe von vier Items, Merkmale, die es Schülerinnen und Schülern ermöglichen sollen, sich aktiv mit Lerninhalten auseinanderzusetzen (Beispielitem: „Die Schülerinnen und Schüler diskutieren und vertreten einen bestimmten Standpunkt, der nicht unbedingt ihr eigener sein muss.“, Cronbach’s Alpha = 0,70).

Skalen zur Erhebung vermittelnder Faktoren Die vermittelnden Faktoren des oben dargestellten Wirkungsmodells werden wiederum mithilfe der folgenden Skalen erhoben:

Die *Arbeitsbedingungen* der Lehrerinnen und Lehrer werden mit den Skalen „Kooperation“ sowie „Partizipation“ erfasst. Die Kooperationsskala wurde dabei aus Steinert et al. (2006) übernommen und misst die Zusammenarbeit im Kollegium anhand von 20 Items, die auf einer vierstufigen Likert-Skala von 1 (trifft nicht zu) bis 4 (trifft zu) beantwortet werden können. Für die vorliegende Untersuchung wurden hiervon jedoch nur sieben Items genutzt, die einerseits die Dimensionen „Zusammenhalt im Kollegium“ (Beispielitem: „In Konferenzen beteiligen sich die meisten Anwesenden aktiv an den Diskussionen.“) und „programmatische Kooperation“ indizieren (Beispielitem: „Wir erarbeiten gemeinsame Strategien zur Bewältigung beruflicher Schwierigkeit.“), jedoch auf einer gemeinsamen Skala abbildbar sind (Cronbach’s Alpha = 0,83). Bei der Skala Partizipation handelt es sich um eine Eigenkonstruktion, die in Anlehnung an ein Instrument von Wahlstrom und Louis (2008) entwickelt wurde und die die Teilhabe von Lehrkräften an der Steuerung der Schule, im Sinne eines *shared leadership,* mithilfe von sieben Items erfasst, die auf einer vierstufigen Likert-Skala von 1 (trifft nicht zu) bis 4 (trifft zu) beantwortet werden können (Beispielitem: „Die Pädagoginnen und Pädagogen an der Schule haben einen erheblichen Einfluss auf die Ausgestaltung von Fortbildungsmaßnahmen.“, Cronbach’s Alpha = 0,80).

Der Bereich der *Emotion und Motivation* wird mithilfe von Skalen zur Arbeitszufriedenheit sowie zum affektiven Commitment der Lehrerinnen und Lehrer erfasst. Dabei wird die Arbeitszufriedenheit anhand einer eigenkonstruierten Skala erfasst, die sich aus fünf Items zusammensetzt, die auf einer vierstufigen Likert-Skala (1 = trifft nicht zu, 4 = trifft voll zu) beantwortet werden können (Beispielitem: „Ich bin zufrieden mit den allgemeinen Arbeitsbedingungen an der Schule.“, Cronbach’s Alpha = 0,88). Das affektive Commitment wiederum wurde aus Lipowski et al. (2009) übernommen. Dieses Konstrukt wird anhand von vier Items indiziert, die ebenfalls auf einer vierstufigen Likert-Skala (1 = trifft nicht zu, 4 = trifft zu) beantwortet werden können (Beispielitem: „Ich bin ausgesprochen froh, dass ich gerade an dieser Schule arbeite.“, Cronbach’s Alpha = 0,91).

[1] Die Schulinspektion Hamburg setzt zur Messung dieser Dimensionen die Skalen der TALIS-Untersuchung aus dem Jahr 2008 in der österreichischen Adaption ein. Die für dort berichteten Alphas belaufen sich auf: 0,602, 0,642 und 0,701 (vgl. OECD 2010). In der internationalen TALIS-Gesamtstichprobe belaufen sich diese Werte auf: 0,723, 0,733 und 0,634. Insofern kann davon ausgegangen werden, dass das hier berichtete Alpha stichprobenbedingt das dargestellte, teilweise recht niedrige Niveau einnimmt.

Die *Innovationskapazität* der Lehrkräfte wird schließlich mithilfe einer selbstkonstruierten Skala gemessen, die sich an die Vorarbeiten von Leithwood und Jantzi (2006) anlehnt (*capacity beliefs*), aus sechs Items besteht und Aspekte wie Selbstwirksamkeitserwartungen, Selbstvertrauen und Selbstkonzepte beim Umgang mit pädagogischen Innovationen und Veränderungen umfasst. Diese Items können ebenfalls auf einer vierstufigen Likert-Skala (1 = trifft nicht zu, 4 = trifft zu) beantwortet werden und geben zusammengenommen Auskunft darüber, in welchem Maße sich Lehrkräfte befähigt fühlen, mit Veränderung und Innovationen umzugehen bzw. diese in den eigenen Unterricht zu implementieren, wenn dies seitens der Bildungsadministration erwartet wird (Beispielitem: „Ich verfüge in der Regel über die notwendigen Kenntnisse und Fähigkeiten, um behördenseitig initiierte Neuerungen in meinem Unterricht umzusetzen.“, Cronbach's Alpha = 0,82).

4.1.2 Leistungsdaten

Um die Instrumente und Modelle der Schulinspektion mithilfe von Leistungsdaten zu validieren, werden nachfolgend Leistungstests aus den Hamburger KERMIT-Erhebungen genutzt. Dieses Instrument erweitert die bundesweit durchgeführten Vergleichsarbeiten (VERA) in den Klassenstufen 3 und 8 um weitere Erhebungen in den Klassenstufen 2, 5, 7 und 9 (Lücken et al. 2014). Durch eine wiederholte Ermittlung von Kompetenzen im Verlauf des Bildungswegs wird die Dokumentation der individuellen Leistungsentwicklung der Schülerinnen und Schüler mittels „echter Längsschnitte“ ermöglicht.

Mit den KERMIT-Erhebungen werden zentrale Leistungsindikatoren erfasst, die ebenfalls in den nationalen und internationalen Schulleistungsuntersuchungen (z. B. PISA, IQB-Ländervergleich) zur Untersuchung der Leistungsfähigkeit von Bildungssystemen herangezogen werden (vgl. Stanat et al. 2012; Pant et al. 2013). Zum Einsatz kommen standardisierte, normierte und aufeinander abgestimmte Schulleistungstests für die Fächer und Testbereiche Deutsch Leseverstehen, Mathematik, Englisch Lese- und Hörverstehen sowie die Naturwissenschaften. Dabei werden die Tests zum Deutsch Leseverstehen sowie zur Mathematik flächendeckend eingesetzt. Auf diese Weise werden objektive diagnostische Informationen gewonnen, mit denen die aktuellen schulischen Leistungen der Hamburger Schülerinnen

Tab. 1 Kennwerte der KERMIT-Erhebungen

KERMIT 5						
Jahr	Testbereich	Items pro Testheft	Innere Konsistenz (Cronbachs Alpha)	Aufgabenschwierigkeit (in Prozent)	Trennschärfe	Intra-klassen-korrelation (ICC)
2010	Deutsch	35	0,80	18–91	0,18–0,56	0,33
–	Mathematik	29	0,83	16–88	0,17–0,58	0,31
2011	Deutsch	26–28	0,81	20–95	0,23–0,63	0,27
–	Mathematik	26	0,80	20–91	0,23–0,58	0,32
2012	Deutsch	28	0,83	24–93	0,20–0,56	0,34
–	Mathematik	29	0,85	17–91	0,19–0,61	0,38

Tab. 2 Kennwerte der KERMIT-Erhebungen

KERMIT 7						
Jahr	Testbereich	Items pro Testheft	Innere Konsistenz (Cronbachs Alpha)	Aufgabenschwierigkeit (in Prozent)	Trennschärfe	Intra-klassen-korrelation (ICC)
2012	Deutsch	32–33	0,84	29–97	0,29–0,58	0,40
–	Mathematik	32–33	0,86	12–87	0,19–0,57	0,47
2013	Deutsch	29–30	0,83	23–90	0,28–0,57	0,31
–	Mathematik	29–31	0,88	17–90	0,18–0,79	0,35
2014	Deutsch	31–32	0,82	10–96	0,21–0,58	0,34
–	Mathematik	33–34	0,84	10–91	0,20–0,64	0,36

Tab. 3 Kennwerte der KERMIT-Erhebungen

KERMIT 9						
Jahr	Testbereich	Items pro Testheft	Innere Konsistenz (Cronbachs Alpha)	Aufgabenschwierigkeit (in Prozent)	Trennschärfe	Intra-klassen-korrelation (ICC)
2013	Deutsch	28–42	0,89	12–92	0,20–0,65	0,39
–	Mathematik	34–52	0,90	09–92	0,20–0,60	0,36
2014	Deutsch	33–35	0,83	20–93	0,20–0,56	0,38
–	Mathematik	30–33	0,86	16–92	0,22–0,66	0,31
2015	Deutsch	26–30	0,80	19–93	0,20–0,57	0,31
–	Mathematik	32–34	0,89	15–91	0,20–0,68	0,35

und Schüler sowie die Entwicklung ihrer Leistungen im Verlauf der Schulzeit zuverlässig abgebildet werden.

Bei KERMIT 5, 7 und 9 verteilen sich die einzelnen Aufgaben der Testbereiche auf verschiedene Testheftversionen. Für die beiden Schulformen Stadtteilschule und Gymnasium gibt es pro KERMIT-Erhebung jeweils zwei Parallelversionen, um ein Abschreiben zu vermeiden. Die mittlere Lösungshäufigkeit bei allen vier Testheften liegt bei ca. 55 %. Diese vier Testhefte sind für jeden Testbereich mit mindestens zehn gleichen Items verankert (vgl. hierzu z. B. Robitzsch et al. 2011). Gleichzeitig wurde auch darauf geachtet, dass bei den Testheften der verschiedenen KERMIT-Erhebungen genauso viele Anker verwendet wurden, um die einzelnen Testbereiche im Längsschnitt von Jahrgang 5 nach Jahrgang 7 sowie von 7 auf 9 abbilden zu können.

Da die KERMIT-Erhebungen in den Jahrgangsstufen 5, 7 und 9 durch externe Testleiterinnen und Testleiter administriert werden und so die Durchführungsobjektivität weitestgehend gesichert ist, lassen sich gerade aus diesen Ergebnissen besonders aussagekräftige Erkenntnisse über das Hamburger Bildungssystem und die Leistungsentwicklung an den Schulen gewinnen. In den Tab. 1, 2 und 3 sind die wichtigsten Kennwerte für die einzelnen Testbereiche für die KERMIT-Erhebungen abgebildet.

Die Intra-Klassen-Korrelationen (ICC) liegen auf Ebene der Klassen in der Regel bei Werten in Höhe von 0,30 bis 0,40. Dasselbe gilt für die ICC auf Ebene der Schule.

Die Leistungen der Schülerinnen und Schüler in den einzelnen Klassenstufen sind dabei zwischen den berichteten Jahrgangskohorten äußerst stabil. Auf Schulebene korrelieren die Leistungen in den Domänen Lesen und Mathematik je Domäne mit r = 0,92 bis r = 0,98 zwischen den Jahrgangskohorten.

4.2 Analysemethoden

4.2.1 Modellierung der latenten Konstrukte sowie des Strukturgleichungsmodells

Im Folgenden werden Schätzungen von Strukturgleichungs- sowie Pfadmodellen berichtet, die mithilfe der Software MPLUS 6.0 (vgl. Muthén und Muthén 2006) ermittelt wurden. Dabei ging es im ersten Schritt primär darum, zu prüfen, wie tragfähig das oben dargestellte Modell zur indirekten Wirkung von Schulleitungen auf den Unterricht von Lehrkräften im Rahmen von Schulinspektionen ist. Daher wurde ein mehrschrittiges, sequentielles Verfahren eingesetzt, das Vorschlägen von Anderson und Gerbig (1998) folgt. Dieses Vorgehen erlaubt es, da das Pfadmodell im Gesamtmodell geschachtelt ist, den Gesamtfit des Modells in die Komponenten des Mess- sowie des Pfadmodells zu unterteilen. Auf diesem Wege kann dem Problem begegnet werden, dass (McDonald und Ho 2002, S. 75) „*the fit of the composite model can appear satisfactory when the few constraints implied by the path model are not, in fact, correctly specified.*“ Die Berechnung eines isolierten Fitindex für das spezifizierte Pfadmodell bietet dann den Vorteil, dass explizit „*the appropriateness of (...) paths and relations among latent variables as proposed by theory*“ (Williams und O'Boyle 2011, S. 18) geprüft und artifizielle Fehlschlüsse zur Tragfähigkeit der theoretischen Annahmen im Strukturgleichungsmodell (die modellierten Pfade und deren Richtung) vermieden werden können (vgl. McDonald und Ho 2002; O'Boyle und Williams 2011; Williams und O'Boyle 2011).

Hierfür wurden erstens die spezifischen Messmodelle der einzelnen latenten Variablen separat geschätzt. Zweitens wurden dann die so ermittelten Item-Parameter fixiert und ein Gesamtmessmodell geschätzt. Drittens wurde anschließend das theoretisch beschriebene Pfadmodell spezifiziert. Dabei wurde stets die Komplexität des Erhebungsdesigns in den Analysen berücksichtigt und entsprechend hinsichtlich der Lehrerinformationen Maximum-Likelihood-Schätzer mit robusten Standardfehlern sowie mit Blick auf den Modellfit robuste χ^2-Statistiken berechnet. Um die Güte des Modells einzuordnen, werden klassische Fit-Maße herangezogen. Demnach kann von einer guten Modellanpassung in etwa bei CFI > 0,90, RMSEA < 0,06 und SRMR < 0,08 ausgegangen werden (vgl. Hu und Bentler 1999; Marsh et al. 2004). Darüber hinaus wird der Fit des Pfadmodells, der RMSEA-*P*, berichtet, der über die χ^2-Werte der einzelnen Modelle, deren Freiheitsgrade sowie die Stichprobengröße ermittelt werden kann (vgl. O'Boyle und Williams 2011). Der RMSEA-*P* lässt sich dann wie folgt berechnen:

$$\sqrt{(((\chi_g^2 - \chi_m^2) - (df_g - df_m))/((df_g - df_m) * (n - 1))).}$$

χ_g^2 entspricht hierbei dem Wert des Gesamt- und χ_m^2 demjenigen des Messmodells. df_g und df_m sind die Freiheitsgrade der jeweiligen Modelle; n bezeichnet die Stich-

probengröße. Auch hier gilt ein Wert von RMSEA-P < 0,06 als Hinweis auf eine gute Modellpassung.

4.2.2 Propensity Score Matching

Weiterhin sollen Schulen belastbar anhand ihrer Leistungsentwicklung verglichen werden. Da es sich hierbei um ein ex-post-facto-Design handelt, bei dem von Unterschieden in Modellparametern zwischen Gruppen auf Differenzen in der Leistungsentwicklung derselben geschlossen wird, besteht die Gefahr, dass Fehlschlüsse auftreten, wenn nicht ausgeschlossen wird, dass andere Faktoren als das Megaevent (in diesem Falle das Schulleitungshandeln) im engen Zusammenhang mit diesen Unterschieden stehen.

Um diese Fehleranfälligkeit zu verringern, wird daher im Folgenden ein so genanntes Propensity-Score-Matching genutzt (vgl. Rosenbaum und Rubin 1983). Die Grundidee dieses Verfahrens besteht darin, für eine Treatmentgruppe eine Kontrollgruppe zu ermitteln, die dieser in relevanten Merkmalen möglichst ähnlich ist. Auf diesem Wege soll sichergestellt werden, dass die zentrale Annahme der kontrafaktischen Kausalanalyse, die Annahme der konditionalen Unabhängigkeit (*conditional independence assumption*, CIA), erfüllt wird. Dies bedeutet, dass nach der Kontrolle von Kovariaten die Verteilung der Analyseeinheiten über Treatment- und Kontrollgruppe möglichst zufällig ist.

Die Schätzung der Propensity Scores sowie das Matching wurden mithilfe des SPSS-Plugins PS Matching (Thoemmes 2014), das auf das R-Paket „match-it“ zurückgreift, durchgeführt. Hierbei wurde sowohl die Mehrebenenstruktur der Daten berücksichtigt als auch ein so genanntes *across cluster matching* vollzogen, also gestattet, für die Bildung der Kontrollgruppe eine Stichprobe aus verschiedenen Schulen zu synthetisieren (vgl. Steiner et al. 2013).

Die Güte des Matchings lässt sich anhand verschiedener Indikatoren überprüfen: Einerseits sollen sich die beiden Gruppen hinsichtlich der kontrollierten Variablen nach dem Matching möglichst nicht mehr statistisch nachweisbar unterscheiden. Andererseits sollte der Standardisierte Bias (SB), der Unterschied in den Mittelwerten der Kovariaten unter Berücksichtigung der Varianz dieser Merkmale (vgl. Rosenbaum und Rubin 1985), auf einen Wert von deutlich unter 20 (vgl. Rosenbaum und Rubin 1985) und idealerweise von 5 % oder weniger reduziert worden sein (vgl. Caliendo und Kopeinig 2008). Weiterhin sollte sich der vorhandene Bias nach dem Matching (*Percent Bias Reduction*, PBR) um mindestens 80 % reduziert haben (vgl. Pan und Bai 2015). Darüber hinaus muss die nicht-parametrische Identifikation des Treatmenteffekts durch eine Überlappung der Propensity Scores (*Common Support*)

Tab. 4 Latente Korrelationen der einzelnen Führungsfacetten

	Transformational	Transaktional	Laissez-faire	Instruktional
Transformational	1,000	–	–	–
Transaktional	0,978	1,000	–	–
Laissez-faire	–0,844	–0,916	1,000	–
Instruktional	0,785	0,825	–0,761	1,000

der beiden Gruppen garantiert werden (vgl. Rubin 2001). Letzteres lässt sich sowohl mithilfe eines globalen χ^2-Tests zur Prüfung der Stichprobenbalance (vgl. Hansen und Bowers 2008) als auch der Statistik $\mathcal{L}_1$-Statistik (vgl. Iacus et al. 2009) prüfen. Dabei sollte der χ^2-Test nach dem Matching idealerweise eine Nichtsignifikanz nachweisen und sich der Wert von $\mathcal{L}_1$ im Vergleich zur ungematchten Stichprobe verringert haben ($A\mathcal{L}_1 < 0$).

5 Befunde

5.1 Der (indirekte) Einfluss des Schulleitungshandelns auf den Unterricht

In einem ersten Schritt wurden, wie oben beschrieben, die Messmodelle der einzelnen latenten Konstrukte geschätzt. Grundlage hierfür bildeten die Lehrerfragebögen aller n = 1663 Lehrkräfte aus den n = 37 Schulen der Gesamtstichprobe. Wie erwartet und aus anderen Studien bekannt (vgl. z. B. Tejeda et al. 2001; van Knippenberg und Sitkin 2013), korrelieren die einzelnen Führungsfacetten hoch miteinander. Wie in Tab. 4 dargestellt, liegen die Korrelationen zwischen den Führungsfacetten alle bei über r = 0,75. Die einzelnen Subskalen des MLQ weisen dabei stets Korrelationen von über r = 0,84 auf, und selbst die instruktionale Führung korreliert mit diesen Skalen stark, wobei insbesondere der Zusammenhang mit der transaktionalen Führung groß ist (r = 0,83). Dabei zeigt sich zwischen dem Aspekt der Laissez-faire-Führung und allen anderen Führungsfacetten stets ein bedeutsamer negativer Zusammenhang (r = –0,92 bis r = –0,76). Insgesamt liegt der Anteil der gemeinsamen Varianz der einzelnen Führungsaspekte damit im Bereich von 58 bis 96 %, was deutlich für einen zugrunde liegenden Generalfaktor spricht, der für diese ausgeprägten Zusammenhänge zwischen den einzelnen Führungsaspekten verantwortlich ist.

Um den Einfluss der einzelnen Führungspraktiken voneinander sowie von einem möglichen Generalfaktor „Aktive Führung durch Schulleitung“ abgrenzen zu können, wurde daher das Führungsmessmodell als Bifaktormodell geschätzt. Bifaktormodelle sind vor allem dann geeignet, wenn es darum geht, heraus zu finden, ob es a) einen Generalfaktor gibt, auf dessen Existenz sich Kommunalitäten zwischen Items zurückführen lassen und b) verschiedene, voneinander abgrenzbare Subfaktoren gibt, die einen eigenen, vom Generalfaktor unabhängigen Einfluss auf abhängige Variablen im Pfadmodell haben und c) Effekte sowohl des General- als auch der einzelnen Subfaktoren im Modell erwartet werden (vgl. Brunner et al. 2012; Chen et al. 2012). In der Modellierung wird entsprechend angenommen, dass es einen allgemeinen Generalfaktor g gibt, der sich auf alle beobachteten Führungsindikatoren auswirkt und einzelne Führungsfacetten, die dann unabhängig voneinander Einfluss auf spezifische Indikatoren haben und sich zueinander ebenso wie zum Generalfaktor orthogonal verhalten. Mit Blick auf die Skalen zur Unterrichtsqualität wurde wiederum, anders als in der klassischen konfirmatorischen Faktorenanalyse, in deren Rahmen Items gemeinhin exklusiv einem Faktor zugeordnet werden, üblich, eine Mehrfachladungsstruktur zugelassen, die den teilweise geringen Skalenreliabilitäten Rechnung trägt, somit hilft Fehlspezifikationen zu vermeiden und

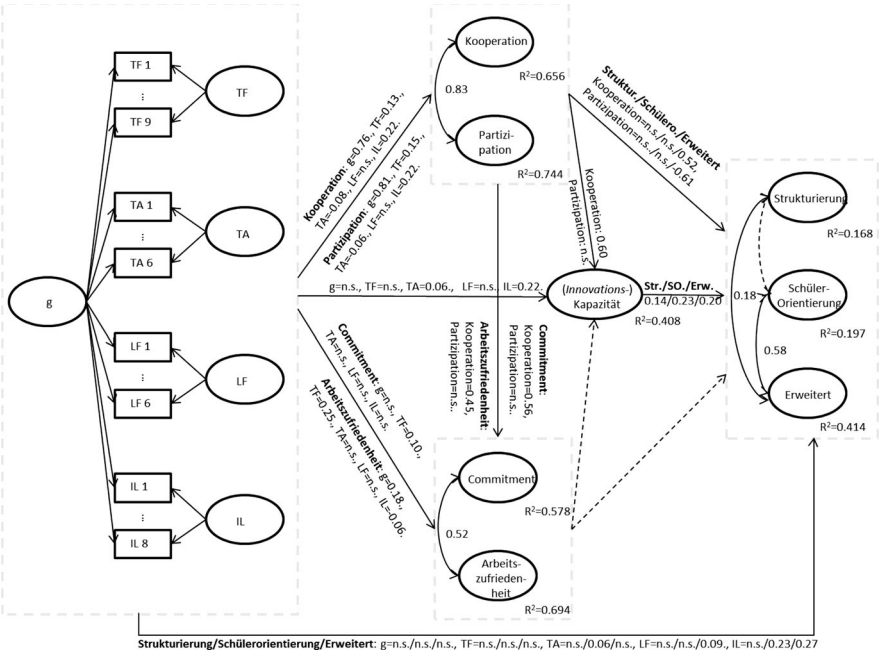


Abb. 3 Strukturgleichungsmodell zur Beschreibung des Einflusses von Schulleitungshandeln auf die Unterrichtsgestaltung von Lehrkräften. 1) *g* Generalfaktor „aktive Führung", *TF* transformationale Führung, *TA* transkationale Führung, *LF* Laissez-faire-Führung, *IL* instruktionale Führung; 2) Modell berechnet unter Kontrolle von Schulform und sozialer Zusammensetzung der Schülerschaft an der Schule; 3) *fett* abhängige Variable, *nicht fett* Einfluss der unabhängigen Variablen auf die abhängige Variable (standardisierte Regressionskoeffizienten); 4) *gestrichelte Linie* keine signifikanten Effekte nachweisbar

zu akkurateren Schätzungen von Korrelationen und Pfadkoeffizienten führt (vgl. Marsh et al. 2013), die Struktur der vorhanden Daten im Vergleich gut abbildet ($RMSEA_{Einfachladung\_Unterricht}$: 0,067, $RMSEA_{Mehrfachladung\_Unterricht}$: 0,053, ΔBIC: –175,369) aber dennoch das theoretisch angenommene Konstrukt einer mehrdimensionalen Struktur von Unterrichtsqualität ausreichend berücksichtigt[2].

Im Anschluss an die Schätzung wurden die einzelnen Messmodelle in einem Gesamtmodell zusammengeführt und in einem ersten Schritt ein Modell geschätzt, in dem alle Korrelationen, mit Ausnahme derjenigen zwischen den einzelnen Führungsfacetten, frei gegeben wurden (χ^2 = 5288,061, df : 1935). Im Anschluss wurde dann das theoretisch definierte Pfadmodell geschätzt, das in Abb. 3 dargestellt ist, wobei mit Thoonen et al. (2011) davon ausgegangen wurde, dass die Innovationskapazität der Lehrkräfte durch die Arbeitsbedingungen sowie ihre Motivation und ihr affektives Commitment beeinflusst wird.

Die Gütekriterien für dieses Modell (χ^2 = 5461,402, df: 2047, RMSEA: 0,032, SRMR: 0,043, CFI: 0,916) weisen auf eine insgesamt gute Passung hin. Die Entflechtung der Fit-Indices zeigt wiederum, dass auch das Pfadmodell eine sehr gute

[2] Die interne Konsistenz einer eindimensionalen Gesamtskala Unterrichtsqualität läge bei einem Cronbach's Alpha in Höhe von 0,74.

Passung aufweist. Der RMSEA-*P* liegt bei 0,018, was dafür spricht, dass der theoretisch postulierte Wirkungszusammenhang zwischen Schulleitungshandeln resp. Führung durch Schulleitungen und der Gestaltung von Unterricht durch Lehrkräfte sich auch in den Daten der Schulinspektion wieder findet.

Auffällig ist, dass das Schulleitungshandeln einen sehr großen Einfluss auf die Arbeitsbedingungen von Lehrkräften ($R^2_{Kooperation}$ = 0,656, $R^2_{Partizipation}$ = 0,744) hat. Die Arbeitsbedingungen werden dabei in erster Linie durch Führung im Allgemeinen (g = 0,76 und 0,81) beeinflusst – wichtig ist demnach vor allem, ob eine Schulleitung überhaupt aktiv gestaltend handelt oder nicht. Jedoch sind auch inkrementelle Effekte sowohl für die transformationale als auch für die transaktionale sowie die instruktionale Führung nachweisbar. Transformationale Führung (TF = 0,13 und 0,15) hat diesbezüglich ebenso wie instruktionale Führung (IL = 0,22 und 0,22) einen nachweisbar positiven Einfluss auf die Kooperation sowie die Partizipation im Kollegium, transaktionale Führung hingegen einen leicht negativen (TA = –0,06). Für den Bereich der Laissez-faire-Führung lassen sich hingegen keine eigenständigen Effekte nachweisen. Dabei ist die Korrelation zwischen den latenten Variablen Kooperation und Partizipation mit r = 0,83 vergleichsweise hoch ausgeprägt.

Auch die Arbeitszufriedenheit und das affektive Commitment der Lehrkräfte lassen sich in weiten Teilen durch das Leitungshandeln in Verbindung mit den schulischen Rahmenbedingungen erklären ($R^2_{Arbeitszufriedenheit}$ = 0,694, $R^2_{Commitment}$ = 0,578). Den größten Einfluss übt aufseiten der Schulleitung demnach die transformationale Führung aus (TF = 0,10 und 0,25). Mit Blick auf die Arbeitszufriedenheit lassen sich darüber hinaus noch positive Effekte durch Führung im Allgemeinen (g = 0,18) aber auch leicht negative Effekte durch instruktionale Führung (IL = –0,06) nachweisen. Stärkster Prädiktor ist jedoch die Kooperation im Kollegium. Mit zunehmender Kooperation zwischen Lehrkräften geht sowohl eine gesteigerte Identifikation (Kooperation = 0,56) mit der Schule als auch eine höhere Arbeitszufriedenheit (Kooperation = 0,45) der einzelnen Lehrkraft einher. Die Partizipation am schulischen Steuerungshandeln hat hingegen keine nachweisbaren Effekte auf diese Variablen.

Auch die Innovationskapazität der Lehrkräfte, also die durch sie berichtete Befähigung, Neuerungen in den eigenen Unterricht zu implementieren, hängt maßgeblich davon ab, wie stark die Kooperation im Kollegium ausgeprägt ist (0,60), wohingegen die Partizipation am Steuerungshandeln in keinem Zusammenhang steht. Aber auch direkte Effekte des Schulleitungshandelns lassen sich nachweisen. So haben hier sowohl die instruktionale (0,22) als auch die transaktionale (0,06) Führung einen nachweisbaren Einfluss darauf, ob Lehrkräfte sich in der Lage sehen, Neuerungen und Innovationen in ihren Unterricht zu implementieren.

Das Unterrichtshandeln der Lehrkräfte wiederum wird sowohl direkt durch das Schulleitungshandeln als auch indirekt über die Arbeitsbedingungen sowie die Innovationskapazität beeinflusst. Hinsichtlich der Arbeitszufriedenheit sowie der Verbundenheit der Lehrkräfte mit der Schule lassen sich jedoch wider Erwarten keinerlei Zusammenhänge zum Unterrichtshandeln der befragten Lehrerinnen und Lehrer nachweisen. Nur begrenzt Einfluss haben Schulleitungen und die vermittelnden Faktoren den Analysen zufolge auf die Klassenführung der Lehrkräfte (R^2 = 0,168).

Dies liegt u. a. daran, dass dieser Unterrichtsaspekt ausschließlich indirekt, über die Innovationskapazität (0,14), also die Befähigung von Lehrkräften, Innovationen und Veränderungen im eigenen Unterricht angemessen umzusetzen, beeinflusst wird. Eine direkte Beziehung des Schulleitungshandelns zur Klassenführung durch Lehrkräfte lässt sich hingegen nicht feststellen. Etwas mehr Einfluss können Schulleitungen und vermittelnde Faktoren dann auf den Bereich der Schülerorientierung ausüben ($R^2 = 0{,}197$). Auch hier spielt die Innovationskapazität als vermittelnde Variable wiederum eine große Rolle (0,23), wobei Schulleitungen jedoch auch direkten instruktionalen (0,23) und transaktionalen (0,06) Einfluss ausüben. Den größten Einfluss haben Schulleitungen und vermittelnde Faktoren jedoch darauf, ob im Unterricht häufig kognitiv aktivierende Methoden eingesetzt werden können ($R^2 = 0{,}414$). Dies wird sowohl direkt, durch instruktionale (0,27) als auch durch Laissez-faire-Führung (0,09), als auch indirekt über die Arbeitsbedingungen (Kooperation = 0,52, Partizipation = –0,61) der Lehrkräfte an der Schule sowie deren Innovationskapazität (0,20) beeinflusst.

5.2 Führungshandeln an Schulen mit unterschiedlicher Leistungsperformanz

5.2.1 Ermittlung von Schulen mit unterschiedlicher Leistungsperformanz

Um zu klären, inwieweit das Führungshandeln mit der Performanz der Schülerinnen und Schüler an einer Schule zusammenhängt, wurden in einem zweiten Schritt Schulen miteinander verglichen, an denen sich die Lernzuwächse unterschiedlich gestalten, Schulleitungen und Lehrkräfte jedoch unter ähnlichen Bedingungen arbeiten.

Hierfür wurden zuerst für jeden Schüler bzw. jede Schülerin individuelle Leistungsentwicklungen in den Lese- sowie den Mathematiktests der KERMIT-Erhebungen in einem Zweischuljahres-Zeitraum (Klassenstufe 5 auf 7 bzw. 7 auf 9) ermittelt. Insgesamt lagen für alle 37 Schulen Leistungsentwicklungen von $n = 23.943$ Schülerinnen und Schüler für die Domänen Mathematik und Lesen und somit insgesamt $n = 47.886$ Lernentwicklungsinformationen vor. Anschließend wurden diese Informationen je Test auf Ebene der Schule aggregiert. Pro Schule lagen demnach folgende 12 Leistungsinformationen vor: 1) Leistungsentwicklung jeweils in Mathematik und Deutsch Leseverstehen von der Klassenstufe 5 auf 7 für die Schuljahre 2010/11 auf 2012/13, 2011/12 auf 2013/14 und 2012/13 auf 2014/15, 2) Leistungsentwicklung in Mathematik und Deutsch Leseverstehen von der Klassenstufe 7 auf 9 für die Schuljahre 2010/11 auf 2012/13, 2011/12 auf 2013/14 und 2012/13 auf 2014/15.

In der Folge wurde eine dichotome Kodierung dieser Variablen vorgenommen und die einzelnen Leistungsentwicklungen dahingehend unterteilt, ob sie auf Schulebene jeweils im Vergleich über- oder unterdurchschnittlich ausgefallen waren. Abschließend wurde aus diesen Informationen ein Summenscore gebildet und auf Basis dieser Information dann diejenigen 25 % der Schulen herausgesucht (75 %-Perzentil), an denen es sowohl über Schüler- und Alterskohorten als auch über Domänen hinweg, und somit stetig und regelmäßig, gelungen war, überdurchschnittlich hohe Lernzuwächse zu erzielen. Hierbei lag der empirisch ermittelte Schwellenwert bei 75 %: an diesen Schulen war die Leistungsentwicklung der Schülerinnen und

Schüler seit dem Jahr 2010 in neun von 12 Fällen überdurchschnittlich hoch. Insgesamt wurden auf diese Art und Weise neun Schulen, allesamt Gymnasien, mit auffallend hohen Lernzuwächsen ermittelt. Grundlage dieser Auswahl bildeten die Leistungsentwicklungsinformationen von $n = 4694$ Schülerinnen und Schülern für die Domänen Lesen und Mathematik. Mit Blick auf die Führung sowie den Unterricht unterscheiden sich die Angaben der Lehrkräfte an diesen hoch performanten Schulen ($n = 326$) von denen der restlichen Stichprobe ($n = 1337$) folgendermaßen: So wird demnach an den Schulen mit hohen Lernzuwächsen seltener im Laissez-faire-Stil ($\text{Ą}MW = -0{,}10$, $p < 0{,}01$), aber dafür häufiger instruktional ($\text{Ą}MW = 0{,}13$, $p < 0{,}01$) geführt. Darüber hinaus kommen häufiger struktur- ($\text{Ą}MW = 0{,}08$, $p < 0{,}01$) und seltener schülerorientierte ($\text{Ą}MW = -0{,}13$, $p < 0{,}01$) Methoden im Unterricht zum Einsatz. Mittelwertunterschiede in der transformationalen und der transaktionalen Führung lassen sich ebenso wenig nachweisen wie beim Einsatz kognitiv aktivierender Methoden im Unterricht.

5.2.2 Propensity Score Matching

Der obige Vergleich berücksichtigt jedoch nicht, dass sowohl Schulform als auch der soziale Hintergrund der Schülerinnen und Schüler einen Einfluss auf die Leistungsstände sowie das Schulleitungshandeln haben können (vgl. z. B. Hallinger 2011). So erklären allein die Variablen Schulform und sozialer Hintergrund bezogen auf die KERMIT-Daten, über die verschiedenen Erhebungen (Domänen, Kohorten und Jahrgänge) hinweg, 81 bis 90 % der Leistungsunterschiede auf Ebene der Einzelschule. Entsprechend wurde ein Propensity Score Matching für die Lehrkräfte an den Schulen durchgeführt um diese Faktoren angemessen zu berücksichtigen. Das Matching erfolgte dabei für den sozialen Hintergrund der Schülerinnen und Schüler, indiziert über den Sozialindex (vgl. hierzu Schulte et al. 2014) der Schule. Hierbei wurde die dichotome Variable „hoch performante Schule vs. nicht-hoch-performante Schule“ als Auswahlkriterium herangezogen. Eingesetzt wurde ein Nearest-Neighbor-Matching mit einer 1-zu-1-Zuordnung und einem Caliper von 0,2 Standardabweichungen. Zusätzlich wurde die Schulform als restringierendes, exaktes Matchingkriterium genutzt, so dass nachfolgend ausschließlich Gymnasien miteinander verglichen werden. Weiterhin wurde die Prozedur ohne Zurücklegen durchgeführt. Nicht zugeordnete Lehrerinnen und Lehrer aus der potenziellen Kontrollgruppe wurden aus dem Datensatz entfernt, so dass, wie beabsichtigt, ebenso viele Lehrkräfte aus der Kontroll- wie aus der Treatmentgruppe im Analysedatensatz verblieben.

Insgesamt lagen vor dem Matching Daten für $n = 1663$ Lehrkräfte vor, wovon $n = 326$ an den neun ausgewählten Gymnasien arbeiteten und $n = 1337$ an den anderen 28 Schulen, die im Datensatz vorhanden waren. Die beiden Lehrergruppen unterschieden sich mit Blick auf die soziale Zusammensetzung der Schülerschaften, die sie unterrichteten, deutlich voneinander ($\text{Ą}MW = 0{,}375$, $p < 0{,}01$). Durch das Propensity Score Matching konnte dieser Unterschied aufgehoben werden ($p > 0{,}10$), wobei jedoch auch die Anzahl der Lehrkräfte im Datensatz reduziert wurde, da nicht für jede der 326 Lehrkräfte ein adäquater Matchingpartner gefunden werden konnte. Letztlich verblieben insgesamt $n = 246$ Lehrkräfte an den Schulen mit hoher Performanz im Datensatz, denen wiederum $n = 246$ Lehrkräfte an vergleichbaren Schulen

Tab. 5 Unterschiede zwischen hoch performanten Schulen und Schulen der Vergleichsgruppe in den Bereichen Führung und Unterricht (nach dem Matching)

	Hoch performante Schulen		Vergleichsgruppe		Signifikanz
	MW	(SE)	MW	(SE)	*p*
Führungsstile					
Transformational	3,00	(0,05)	2,98	(0,05)	>0,10
Transaktional	2,72	(0,03)	2,71	(0,03)	>0,10
Laissez-faire	1,80	(0,04)	1,87	(0,04)	>0,10
Instruktional	2,61	(0,04)	2,47	(0,04)	<0,01
Unterrichtsmethoden					
Strukturorientiert	3,28	(0,03)	3,20	(0,03)	<0,05
Schülerorientiert	2,69	(0,04)	2,61	(0,04)	>0,10
Erweitert	2,06	(0,04)	1,97	(0,04)	>0,10

zugeordnet wurden. Der Unterschied in der kontrollierten Variable (standardisierter Bias) konnte dabei um 92 % abgebaut werden und liegt im gematchten Datensatz bei unter 9,2 % (vor dem Matching: 91 %). Der globale χ^2-Test zur Prüfung der Stichprobenbalance ist darüber hinaus nicht signifikant ($p > 0{,}10$) und der Wert der $\mathcal{L}_1$-Statistik konnte deutlich reduziert werden ($A\mathcal{L}_1 = -0{,}15$).

5.2.3 Befunde

Für die weiteren Analysen lagen demnach $n = 492$ Fälle vor. Wie Tab. 5 zeigt, werden die Lehrmethoden im Unterricht durch die Lehrkräfte an den Schulen mit hohen Leistungszuwächsen häufiger variiert als an den Vergleichsschulen, auch wird häufiger instruktional und seltener im Laissez-faire-Stil geführt. Jedoch unterscheiden sich die Unterrichts- ebenso wie die Führungspraktiken in beiden Gruppen nur in zwei Punkten statistisch nachweisbar voneinander. Einerseits führen Schulleitungen an Schulen mit hohen Lernzuwächsen demnach häufiger instruktional als an Schulen der Vergleichsgruppe. Andererseits lässt sich auch für den Bereich der Klassenführung, hier mit *strukturorientierter Unterricht* benannt, ein leichter Unterschied feststellen. Lehrkräfte an hoch performanten Schulen geben an, häufiger Methoden im Unterricht einzusetzen, die dazu beitragen, dass die vorhandene Lernzeit effektiv genutzt und Störungen im Unterricht vorgebeugt wird.

Mit den Daten der 492 ausgewählten Lehrkräfte wurde weiterhin das unter 5.1. beschriebene Modell für die beiden Gruppen erneut gerechnet. Aufgrund der geringeren Fallzahl in Verbindung mit der großen Anzahl an Items wurde nur das Pfadmodell geschätzt und die Messmodelle wurden als manifeste Variablen berücksichtigt. Entsprechend wurden die Faktorscores, die im Gesamtmodell für die einzelnen latenten Variablen je Lehrkraft ermittelt wurden, für die Analyse eingelesen und nachfolgend im Rahmen der Pfadmodellierung genutzt.

Die Tab. 6 zeigt nun folgendes: Der Einfluss der Modellvariablen auf die Gestaltung des Unterrichts durch Lehrkräfte ist in beiden Gruppen ähnlich ausgeprägt. Den geringsten Einfluss haben die Modellvariablen, wie auch im oben dargestellten Gesamtmodell, grundsätzlich auf die Klassenführung der Lehrerinnen und Lehrer

Tab. 6 Effekte des Schulleitungshandelns auf den Unterricht an Schulen mit hoher Performanz und Schulen der Vergleichsgruppe (nach dem Matching)

			Vergleichsgruppe: Effekte auf			
	Strukturierung		Schülerorientierung		Erweiterte Aktivitäten	
	Direkt	Insgesamt	Direkt	Insgesamt	Direkt	Insgesamt
g	0,515	–0,064	0,597	–0,080	0,771	–0,087
TF	–	–0,159	–	–0,134	0,203	–0,032
TA	–	0,037	–	0,035	–	0,000
LF	–	–0,014	0,095	0,060	0,071	0,036
IL	–	0,018	0,425	0,352	0,458	0,286
R2	0,228		0,431		0,363	
			Hoch performante Schulen: Effekte auf			
	Strukturierung		Schülerorientierung		Erweiterte Aktivitäten	
	Direkt	Insgesamt	Direkt	Insgesamt	Direkt	Insgesamt
g	–	–0,051	0,568	–0,066	0,700	0,012
TF	–	–0,078	–	–0,100	0,185	0,093
TA	–	–0,004	0,098	0,103	–	–0,036
LF	–	0,030	–	0,030	0,145	0,145
IL	0,195	0,601	0,559	0,467	0,525	0,406
R2	0,198		0,414		0,416	

Alle Effekte sind auf dem Niveau von $p < 0{,}05$ signifikant

($R^2 = 0{,}228$ und $R^2 = 0{,}198$). Dabei ist nur an leistungsstarken Schulen ein direkter Einfluss der instruktionalen Führung der Schulleitungen auf den Unterricht feststellbar (0,20), wobei dieser durch indirekte Effekte noch deutlich verstärkt wird (0,60). Auffällig ist darüber hinaus, dass an den Schulen der Vergleichsgruppe vor allem transformationale Führung eine Rolle spielt, die jedoch einen negativen Effekt nach sich zieht (–0,16). Für die instruktionale Führung lassen sich hingegen nur äußerst geringe Effekte nachweisen (0,02).

Die Schülerorientierung im Unterricht wiederum wird an beiden Schulgruppen in gleichem Maße erklärt ($R^2 = 0{,}431$ und $R^2 = 0{,}414$). Bei beiden Gruppen ist es die instruktionale Führung, die die diesbezügliche Unterrichtsgestaltung durch die Lehrkräfte in erster Linie beeinflusst (0,35 und 0,47), wobei auffällig ist, dass auch hier die transformationale Führung in beiden Gruppen einen negativen Effekt nach sich zieht (–0,13 und –0,10).

Hinsichtlich der kognitiv aktivierenden Unterrichtsgestaltung bestätigen sich die Befunde des oben vorgestellten Gesamtmodells: Die Kovariaten im Modell haben demnach einen großen Einfluss darauf, ob eine solche Unterrichtsgestaltung häufiger oder seltener stattfindet ($R^2 = 0{,}363$ und $R^2 = 0{,}416$). Auch hier ist es vor allem die instruktionale Führung, die Effekte nach sich zieht, wobei der diesbezügliche Einfluss an hoch performanten Schulen deutlich größer ist als an Schulen der Vergleichsgruppe (0,29 und 0,41). Auffällig ist darüber hinaus, dass die instruktionale Einflussnahme der Schulleitungen an leistungsstarken Schulen mit einer Laissez-faire-Führung einhergeht, die auch positiv dazu beiträgt, dass häufiger kognitiv aktivierende Methoden im Unterricht eingesetzt werden (0,15).

Tab. 7 Effekte des Schulleitungshandelns auf vermittelnde Faktoren Schulen mit hoher Performanz und Schulen der Vergleichsgruppe. (Nach dem Matching)

			Vergleichsgruppe: Direkte Effekte auf		
	Kooperation	Partizipation	Commitment	Arbeitszufriedenheit	Innovationsfähigkeit
g	0,855	0,895	0,271	–	–
TF	0,112	0,136	0,176	0,269	0,158
TA	–	–	–	–	0,126
LF	–0,076	–0,050	0,087	–	–
IL	0,212	0,208	–	–	0,270
R2	0,796	0,868	0,748	0,846	0,520
			Hoch performante Schulen: Direkte Effekte auf		
	Kooperation	Partizipation	Commitment	Arbeitszufriedenheit	Innovationsfähigkeit
g	0,802	0,861	0,245	0,227	0,118
TF	0,199	0,189	0,095	0,259	–
TA	–0,096	–0,066	–0,072	–	–
LF	–	–	–	–	0,126
IL	0,294	0,274	–	–	0,297
R2	0,782	0,859	0,703	0,800	0,601

Alle Effekte sind auf dem Niveau von $p < 0{,}05$ signifikant

Aus der Tab. 7 kann letztlich abgelesen werden, dass der Einfluss von Schulleitungen auf die Rahmenbedingungen, unter denen Lehrkräfte unterrichten, zwischen den beiden Gruppen in etwa gleich groß ist. Der Einfluss auf das affektive Commitment und die Arbeitszufriedenheit ist an hoch performanten Schulen jedoch etwas geringer. Dafür jedoch ist ihr Einfluss auf die Innovationskapazität, also die Befähigung der Lehrerinnen und Lehrer, Innovationen und Veränderungen im eigenen Unterricht umzusetzen, wiederum größer.

Insbesondere mit Blick auf diesen Punkt unterscheidet sich die Führung in beiden Gruppen deutlich voneinander. Während Schulleitungen an weniger leistungsstarken Schulen instruktionale Führung (0,27) mit transformationaler (0,16) und transaktionaler (0,13) Führung koppeln, um Einfluss auf die Innovationskapazität der Lehrkräfte zu nehmen, flankieren Schulleitungen an leistungsstarken Schulen ihre instruktionalen Führungspraktiken (0,30) durch ein grundsätzlich aktives Schulleitungshandeln (0,12) in Verbindung mit einem Laissez-Faire-Stil (0,13). Auch mit Blick auf die Arbeitsbedingungen der Lehrerinnen und Lehrer ist das Leitungshandeln an hoch performanten Schulen deutlicher durch einen instruktionalen Führungsstil geprägt als an den Schulen der Vergleichsgruppe (0,29 und 0,27 vs. 0,21 und 0,21), wobei jedoch auch deutlich wird, dass an ersteren Schulen der gesteigerte Einfluss dieses Führungsstils mit einem ebenfalls stärkeren Einfluss (0,20 und 0,19 vs. 0,11 und 0,14) transformationaler Führung einhergeht.

6 Zusammenfassung und Diskussion

Im Rahmen des Beitrages ging es darum, zu validieren, inwieweit Aussagen von Lehrkräften zum Schulleitungshandeln in Fragebögen der Schulinspektion dazu ge-

eignet sind, hieraus belastbare Entscheidungen zur Schulentwicklung und in der Konsequenz zur Verbesserung von Schülerleistungen abzuleiten.

Die Befunde machen deutlich, dass diese Konsequenz nicht ohne weiteres zu erwarten ist. So zeigen die Analysen, dass es möglich ist, das Schulleitungshandeln auf Basis der gewonnenen Informationen (sozial) zu vergleichen und somit Bewertungen zu ermöglichen. Auch ließ sich nachweisen, dass die eingesetzten Führungsskalen, selbst bei Einsatz eines Bi-Faktor-Modells, empirisch nachweisbar voneinander diskriminieren und somit einzelne relevante Führungsfacetten ausreichend gut repräsentieren. Grundsätzlich lässt sich damit theorie- sowie evidenzkonform festhalten: Schulleitungen nutzen unterschiedliche Führungsstile, sie wirken auf den Unterricht[3] von Lehrkräften sowohl direkt als auch indirekt und insbesondere der Aspekt der Innovationskapazität von Lehrkräften spielt eine entscheidende Rolle als Mediator. Darüber hinaus führen Schulleitungen an hoch performanten Schulen auch häufiger instruktional und die Lehrkräfte an diesen Schulen legen besonderen Wert auf das Klassenmanagement.

Dies allein ist jedoch für das intendierte Ziel der inspektionsbasierten Schulentwicklung nicht hinreichend. Denn eine Verallgemeinerung bzw. Extrapolation ist wiederum nicht oder nur begrenzt möglich. So lassen sich zwar für die einzelnen Skalen die theoretisch postulierten und empirisch bekannten Zusammenhänge mit anderen Konstrukten nachweisen. Diese variieren jedoch in Stärke, Einfluss und Komplexität zwischen verschiedenen Kontexten, was darauf hindeutet, dass Schulleitungen an Schulen mit hoher Performanz grundsätzlich anders führen als ihre Kolleginnen und Kollegen an Schulen mit weniger hohen Lernzuwächsen. Mit anderen Worten: Schulleitungshandeln erfolgt weniger standardisiert als vielmehr situiert. Dies wiederum macht es in der Konsequenz dann auch schwierig aus einem standardisierten und mit Blick auf den Kontext invarianten Stärken-Schwächen-Profil, wie es im Rahmen von Schulinspektionen genutzt wird, zielgerichtete Entscheidungen zur Schulentwicklung zu treffen, die dann letztlich in verbesserten Schülerleistungen münden.

Grundsätzlich bedeutet dies mit Blick auf die Effektivität von Schulinspektionen bzw. deren Effektivierung zweierlei: Einerseits können quantitative, bewährte Instrumente zur Ermittlung von Schulleitungshandeln durch die Befragung von Lehrkräften belastbar eingesetzt werden, um erste, vergleichbare Diagnosen zur Qualität der Führung an Schulen zu ermöglichen. Andererseits sind diese Informationen als Grundlage für eine wissensbasierte Schulentwicklung jedoch nicht hinreichend, wenn sie, wie im Rahmen von Inspektionen üblich, primär im Rahmen eines Stärken-Schwächen-Profils verarbeitet und berichtet werden. Vielmehr bedarf es im Rahmen

[3] Einschränkend ist dabei festzuhalten, dass die Effekte, die die latenten Variablen im Modell auf den Unterricht haben, teilweise leicht unterhalb derjenigen liegen, die im Rahmen der Literatur (20–40 % erklärte Varianz, vgl. Leithwood und Jantzi 2006) berichtet werden. Dies kann jedoch u. a. an der teilweise geringen internen Konsistenz der aus TALIS stammenden Skalen zum Unterricht bzw. daran, dass durch die Berücksichtigung einer Mehrfachladungsstruktur Zusammenhänge grundsätzlich geringer als in klassischen Strukturgleichungsmodellen geschätzt werden, liegen. Robustere Ergebnisse hätten womöglich berichtet werden können, wenn die ebenfalls im Rahmen der Schulinspektion vorhandenen Informationen zum Unterricht aus Beobachter- und/oder Schülerperspektive den Lehrkräften zuordenbar gewesen wären. Diese Daten konnten jedoch aus Gründen des Datenschutzes nicht miteinander verknüpft werden.

von Inspektionen einer umfassenderen Auseinandersetzung mit den Bedingungen vor Ort sowie der Analyse und Vermittlung der komplexen Zusammenhänge und Interaktionen zwischen einzelnen Merkmalen auf Ebene der Schule. Nur so können mit Hilfe von zielgerichteten und aufeinander abgestimmten Veränderungen die intendierten Entwicklungen auch wirklich mit hoher Wahrscheinlichkeit eintreten.

Was diesbezüglich konkret wirksam ist, dazu lassen die hier vorgelegten querschnittlich angelegten ex-post-facto-Analysen keinen Schluss zu. Zu erwarten wäre jedoch, dass Inspektionen, selbst dann, wenn es ihnen gelingt, Informationen valide zu erheben, voraussichtlich erst in dem Augenblick wirklich effektiv werden, ab dem ihre Arbeit durch ein entsprechendes Beratungs- und Unterstützungssystem für Schulen flankiert wird. Um diese Annahme zu prüfen, bedarf es jedoch alternativer Studiendesigns, die den Rückmeldeprozess und die innerschulische Verarbeitung von Rückmeldeinformationen an sich in den Blick nehmen, wie z. B. Interventionsstudien und/oder Begleituntersuchungen, die Schulen im Längsschnitt direkt im Anschluss an die Inspektion oder gar schon im Vorfeld des Inspektionsbesuchs über einen längeren Zeitraum wissenschaftlich begleiten.

Literatur

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Bass, B. M., & Avolio, B. J. (1995). *MLQ Multifactor leadership questionnaire*. Technical report. Redwood City: Mind Garden.

Behörde für Schule und Berufsbildung. (2012). *Orientierungsrahmen Schulqualität und Leitfaden*. Hamburg: Behörde für Schule und Berufsbildung.

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: Springer VS.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796–846.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. *Journal of Personality*, *80*, 219–251.

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools*. New York: Routledge.

Denis, J.-L., Langley, A., & Sergi, V. (2012). Leadership in the plural. *The Academy of Management Annals*, *6*(1), 211–283.

Diedrich, M. (2015). Der zweite Zyklus der Schulinspektion Hamburg: Ein Ausblick. In M. Pietsch, B. Scholand, & K. Schulte (Hrsg.), *Schulinspektion in Hamburg, Der erste Zyklus 2007–2013: Grundlagen, Befunde, Perspektiven* (S. 419–436). Münster: Waxmann.

Döbert, H., Rürup, M., & Dedering, K. (2008). Externe Evaluation von Schulen in Deutschland – die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede. In H. Döbert & K. Dedering (Hrsg.), *Externe Evaluation von Schulen. Historische, rechtliche und vergleichende Aspekte* (S. 63–152). Münster: Waxmann.

Ehren, M. C. M., & Pietsch, M. (2016). Validation of inspection frameworks and methods. In M. C. M. Ehren (Hrsg.), *Methods and modalities of effective school inspections* (S. 39–53). London: Springer.

Ehren, M. C. M., & Scheerens, J. (2015). Evidenzbasierte Referenzrahmen zur Schulqualität als Grundlage von Schulinspektion. In M. Pietsch, B. Scholand, & K. Schulte (Hrsg.), *Schulinspektion in Hamburg, Der erste Zyklus 2007–2013: Grundlagen, Befunde, Perspektiven* (S. 233–272). Münster: Waxmann.

Felfe, J. (2006). Validierung einer deutschen Version des „Multifactor Leadershipnaire" (MLQ Form 5 x Short) von Bass und Avolio (1995). *Zeitschrift für Arbeits- und Organisationspsychologie*, *50*, 61–78.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: Springer VS.

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 aus empirischer Sicht* (S. 9–32). Münster: Waxmann.

Gilroy, P., & Wilcox, B. (1997). OFSTED, criteria and the nature of social understanding: a Wittgensteinian critique of the practice of educational judgement. *British Journal of Educational Studies*, *45*(1), 22–38.

Hallinger, P. (1994). *A resource manual for the Principal Instructional Management Rating Scale (PIRMS manual 2.2)*. Nashville: Center for the Advanced Study of Educational Leadership.

Hallinger, P. (2003). Leading educational change. Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, *33*(3), 329–351.

Hallinger, P. (2011). Leadership for learning: lessons from 40 years of empirical research. *Journal of Educational Administration*, *49*(2), 125–142.

Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980-1995. *School Effectiveness and School Improvement*, *9*(2), 157–191.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*, 219–236.

Harazd, B., & van Ophuysen, S. (2011). Transformationale Führung in Schulen: Der Einsatz des „Multifactor Leadership Questionnaire" (MLQ 5 x Short). *Journal for Educational Research Online*, *3*(1), 141–167.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Iacus, S. M., King, G., & Porro, G. (2009). CEM: coarsened exact matching software. *Journal of Statistical Software*, *30*, 1–27.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: a meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*(5), 755–768.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Klieme, E., & Rackoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*(2), 222–237.

van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic – transformational leadership research: back to the drawing board? *The Academy of Management Annals*, *7*(1), 1–60.

Kotter, J. F. (1990). *A force for change: How leadership differs from management*. New York: Free Press.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Measurement: Issues and Practice*, *21*(1), 23–30.

Leithwood, K., & Jantzi, D. (2006). Transformational school leadership for largescale reform: effects on students, teachers, and their classroom practices. *School Effectiveness and School Improvement*, *17*, 201–227.

Leithwood, K., Jantzi, D., & Mascall, B. (2002). A framework for research on largescale reform. *Journal of Educational Change*, *3*, 7–33.

Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership and Management*, *28*(1), 27–42.

Lipowski, F., Faust, G., & Greb, K. (2009). *PERLE-Instrumente. Schüler, Lehrer, Eltern (Messzeitpunkt 1)*. Frankfurt a. M.: DIPF.

Lücken, M., Thonke, F., Pohlmann, B., Hofmann, H., Golecki, R., Rosendahl, J., Benzing, M., & Poerschke, J. (2014). KERMIT – Kompetenzen ermitteln. In D. Fickermann & N. Maritzen (Hrsg.), *Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung* (S. 127–154). Münster: Waxmann.

De Maeyer, S., Rymenans, R., van Petegem, P., van den Bergh, H., & Rijlaarsdam, G. (2007). Instructional leadership and pupil achievement: the choice of a valid conceptual model to test effects in school effectiveness research. *School Effectiveness and School Improvement*, *18*(2), 125–145.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: an integration of transformational and instructional leadership. *Educational Administration Quarterly*, *39*, 370–397.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right. camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*, 257–284.

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: from research to results*. Alexandria: Association for Supervision and Curriculum Development.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus software (Version 6)*. Los Angeles: Muthén & Muthén.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: SAGE.

OECD (2009) = Organisation for Economic Co-operation and Development. (2009). *Creating effective teaching and learning environments. first results from TALIS*. Paris: OECD.

OECD (2010) = Organisation for Economic Co-operation and Development. (2010). *TALIS 2008*. Technical report. Paris: OECD.

O'Boyle Jr., E. H., & Williams, L. J. (2011). Decomposing model fit: measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology*, *96*(1), 1–12.

Pan, W., & Bai, H. (2015). *Propensity score analysis: fundamentals and developments*. New York: Guilford.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012 – Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Pietsch, M. (2014a). Wirksame Schulleitungen: Annahmen, Befunde und der Blick der Schulinspektion. *Hamburg macht Schule*, *26*(1), 6–11.

Pietsch, M. (2014b). Was wissen wir über wirksame Schulleitungen? Eine Zusammenschau und praxisorientierte Einordnung von Best-Evidence-Forschungsbefunden der letzten 10 Jahre. *Journal für Schulentwicklung*, *18*(2), 15–23.

Pietsch, M., van den Ham, A.-K., & Köller, O. (2015). Wirkung von Schulinspektion: Ein Rahmen zur theoriegeleiteten Analyse von Schulinspektionseffekten. In M. Pietsch, B. Scholand, & K. Schulte (Hrsg.), *Schulinspektion in Hamburg, Der erste Zyklus 2007–2013: Grundlagen, Befunde, Perspektiven* (S. 117–136). Münster: Waxmann.

Preuß, B., Brüsemeister, T., & Wissinger, J. (2015). Einführung der Schulinspektion: Struktur und Wandel regionaler Governance im Schulsystem. In H. J. Abs, T. Brüsemeister, M. Schemmann, & T. Wissiger (Hrsg.), *Governance im Bildungssystem: Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 117–142). Wiesbaden: Springer VS.

Robinson, V. M. J., Lloyd, C., & Rowe, K. J. (2008). The impact of leadership on student outcomes: an analysis of the differential effects of leadership type. *Educational Administration Quarterly*, *44*(5), 635–674.

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, *43*, 213–227.

Rosenbaum, P., & Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P., & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistican*, *39*, 33–38.

Rubin, D. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188.

Scheerens, J. (2012). *School leadership effects revisited: review and meta-analyses of empirical studies*. Dordrecht: Springer.

Schmich, J., & Schreiner, C. (2008). *TALIS 2008. Schule als Lernumfeld und Arbeitsplatz*. Graz: Leykam.

Schulte, K., Hartig, J., & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann & N. Maritzen (Hrsg.), *Grundlagen für eine daten- und theoriegestützte Schulentwicklung – Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ)* (S. 67–80). Münster: Waxmann.

Shatzer, R. H., Caldarella, P., Hallam, P. R., & Brown, B. L. (2013). Comparing the effects of instructional and transformational leadership on student achievement: implications for practice. *Educational Management Administration & Leadership*, *42*(4), 445–459.

Sireci, S. G., & Sucin, T. (2013). Test validity. In K. F. Geisinger (Hrsg.), *APA Handbook of Testing and Assessment in Psychology. Bd. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology* (S. 61–84). Washington DC: American Psychological Association.

Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik*. Münster: Waxmann.

Steiner, P. M., Kim, J.-S., & Thoemmes, F. (2013). *Matching strategies for observational multilevel data. JSM proceedings*. Alexandria: American Statistical Association.

Steinert, B., Klieme, E., Maag Merki, K., Döbrich, P., Halbheer, U., & Kunz, A. (2006). Lehrerkooperation in der Schule: Konzeption, Erfassung, Ergebnisse. *Zeitschrift für Pädagogik*, *52*, 185–204.

Tejeda, M., Scandura, T., & Pillai, R. (2001). The MLQ revisited: psychometric properties and recommendations. *The Leadership Quarterly*, *12*, 31–52.

Thoemmes, F. (2014). *Propensity score matching in SPSS*. http://sourceforge.net/projects/psmspss/. Zugegriffen: 13. Juli 2016.

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., & Geijsel, F. P. (2011). How to improve teaching practices: the role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly*, *47*(3), 496–536.

Vieluf, S., Kaplan, D., Klieme, E., & Bayer, J. (2012). *Teaching practices and paedagogical innovation: evidence from TALIS*. Paris: OECD.

Wahlstrom, K. L., & Louis, K. S. (2008). How teachers experiences principal leadership. The roles of professional community, trust, efficiency, and shared responsibility. *Educational Administration Quarterly*, *44*, 458–495.

Williams, L., & O'Boyle, E. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods*, *14*(2), 350–369.

Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Instructional leadership and student achievement: the elusive search for an association. *Educational Administration Quarterly*, *39*(3), 398–425.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT, 2017
VOL. 28, NO. 4, 629–649
https://doi.org/10.1080/09243453.2017.1363787

Routledge
Taylor & Francis Group

ARTICLE

Disentangling school leadership and its ties to instructional practices – an empirical comparison of various leadership styles

Marcus Pietsch [a] and Pierre Tulowitzki [b]

[a]Department of Education, Leuphana University Lueneburg, Lueneburg, Germany; [b]Department of International Educational Leadership and Management, Ludwigsburg University of Education, Ludwigsburg, Germany

ABSTRACT

This paper investigates the direct and indirect ties between various leadership styles, namely, instructional, transformational, transactional, and laissez-faire leadership, and the instructional practices of teachers by applying a structural equation model. For this purpose, we analyzed survey data of n = 3,746 teachers from 126 schools collected by the Hamburg school inspection in Germany between 2012 and 2015. The underlying model is based on Leithwood's framework for guiding research on leader effects on learning and instruction. First, the results show that a bi-factor model seems to be the best measurement model. Next, it is shown that mediating variables are influenced by a leadership core as well as by different leadership facets. Third, we found that for influencing complex instructional practices like cognitive activation with challenging content, a combination of leadership styles is most promising, while for classroom management instructional leadership is the only and, thus, the primary determinant.

ARTICLE HISTORY
Received 21 November 2016
Accepted 1 August 2017

KEYWORDS
Bi-factor model; instructional leadership; instructional practices; school leadership; transformational leadership

Introduction

School leadership has been deemed an important factor for creating and sustaining "functional" schools (Robinson, Lloyd, & Rowe, 2008). There are nowadays a myriad of leadership styles and models of leadership, with instructional leadership and transformational leadership being two of the more popular ones in the educational discourse.

Empirical evidence indicates that both leadership styles have an effect on student learning as well as on learning preconditions within schools. However, until now only few studies have attempted to investigate the differential effects of instructional and transformational leadership using one coherent design. Robinson et al. (2008) compared those two leadership styles in a meta-analysis and found that the effect size for instructional leadership on student achievement (r = .42) was nearly 4 times as high as the effect size for transformational leadership (r = .11). Day, Gu, and Sammons (2016) looked at the impact of leadership practices on student outcomes using a mixed-methods design. The quantitative portion of the study employed a structural equation modeling (SEM) analysis with data from 309 secondary schools and 363 primary schools. Regarding

CONTACT Marcus Pietsch pietsch@leuphana.de

leadership, they found "neither instructional leadership strategies nor transformational leadership strategies alone were sufficient to promote improvement identified by the SEM model" (Day et al., 2016, p. 238). Instead, they found successful principals to employ "layered leadership" consisting of strategies and actions that were both transformational and instructional (Day et al., 2016, p. 245). By contrast, a study by Shatzer, Caldarella, Hallam, and Brown (2014), which compared both leadership styles directly, revealed that instructional leadership explained more of the variance in student achievement and achievement gains than transformational leadership, supporting the conclusion that instructional leadership has a slight advantage over transformational leadership in relation to student achievement.

Therefore, Marks and Printy (2003) may be right when they argue that transformational leadership is of great importance for school reform, but that the practices related to this type of leadership do not lead to improvements in student outcomes as they lack a clear focus on teaching and learning. But until now, this assumption has more or less remained untested due to intermediate variables with regard to teaching and other organizational factors being absent in many leadership effect studies (Scheerens, 2012). Studies researching the effects of leadership on teaching practices and other school variables are often each exclusively focused on a specific leadership style (Leithwood & Jantzi, 2006; Thoonen, Sleegers, Oort, Peetsma, & Geijsel, 2011). A study directly comparing instructional and transformational leadership and their ties to the instructional practices of teachers within a single model-driven framework (and sample) has yet to be conducted.

Addressing these matters, the overall purpose of the present study was to simultaneously test the differential effects of different leadership styles on organizational conditions, teachers, and their instructional practices. For this purpose, this paper investigates the factor structure of various leadership styles, namely, instructional, transformational, transactional, and laissez-faire leadership, as well as the direct and indirect ties between these leadership styles and the instructional practices of teachers. This is done by applying a complex structural equation model, which draws on Leithwood's model of leadership influence on teaching and student learning (Leithwood & Jantzi, 2006).

Instructional and transformational leadership: an overview

While there exists a great multitude of leadership styles (MacBeath, 2003), with many of them having been the subject of empirical enquiry, two styles have gained a certain enduring prominence in the domain of educational leadership research: instructional leadership and transformational leadership.

Instructional leadership

Instructional leadership can be viewed as being centered on the quality of teaching in classrooms. It "typically assumes that the critical focus for attention by leaders is the *behaviours of teachers as they engage in activities directly affecting the growth of students"* (Leithwood, Jantzi, & Steinbach, 1999, p. 8, emphasis in the original text). Emphasis is put – as the name suggests – on the principal having a succinct understanding of instruction in general, but also of the curriculum so as to be able to judge what is taught and how

and to provide appropriate feedback. Thus, from an instructional leadership perspective, the principal is responsible for the quality of teaching of her/his staff and is influential in this regard. Common areas of activity of instructional leadership include (Krug, 1992, pp. 433–434):

- defining the schools' mission;
- managing curriculum and instruction;
- supervising teaching;
- monitoring student progress;
- promoting instructional climate.

These areas are close to areas often associated with the tasks of teachers, highlighting how instructional leadership activities can often cross paths with typical teacher activities. Instructional leadership and matters of curriculum as well as curriculum research have been linked on several occasions. This even led to a temporary rise of the term "curriculum leadership", often used similarly to "instructional leadership" (e.g., in Fidler, 1997; Lee & Dimmock, 1999) though never gaining the latter's predominance. Following Hallinger (2003), all facets of instructional leadership can be combined into three dimensions: (a) defining the school's mission, (b) managing the instructional program, and (c) promoting a positive school learning climate.

Recapping its history and looking at its current state, Hallinger and Wang conclude that "instructional leadership has become increasingly accepted globally as a normative expectation in the principalship", acknowledging that while other models have come and gone, "scholarly interest in instructional leadership has remained surprisingly consistent and strong" (Hallinger & Wang, 2015, p. 15). Its roots are often linked to the effective schools movement arising from the US in the 1960s, which led to increased research in the domain of instructional leadership. Once the notion that schools did not matter (Coleman et al., 1966, although it was never expressed this drastically in said report) had been refuted (Rutter, Maughan, Mortimore, & Ouston, 1979), attention quickly turned towards also looking at school principals as influential factors with regard to improving student success. Evidence suggested that in schools that were improving in challenging circumstances, the school principal was more likely to be an instructional leader (see, e.g., Edmonds, 1979). This led to increased research efforts in this area, often attempting to assess and characterize effective instructional principals (Hallinger & Murphy, 1986; Leithwood, Begley, & Cousins, 1990). This was complemented by research on the work activity as well as the time use of school principals (Kmetz & Willower, 1982; Martin & Willower, 1981), indicating that in many cases principals did not spend much time on instructional leadership due to a myriad of other activities. These findings dampened the enthusiasm for the principal as omnipresent chief instructor (among many other things). Later, studies from various contexts solidified these results, often finding that administrative duties overshadowed curriculum and instruction (Horng, Klasik, & Loeb, 2010; Huber, Gördel, Kilic, & Tulowitzki, 2016; Spillane & Hunt, 2010; Tulowitzki, 2013; Wildy & Dimmock, 1993).

Transformational leadership

The term transformational leadership is often traced back to Burns (1978), who distinguished between transactional and transforming leadership. He explained that

transforming leadership "occurs when one or more persons engage with others in such a way that leaders and followers raise one another to higher levels of motivation and morality" (p. 20). Burns originally considered transforming leadership to stand in opposition to transactional leadership (both being ends on a spectrum). Bass (1990) expanded upon this concept; he described transformational leadership as something that:

> occurs when leaders broaden and elevate the interests of their employees, when they generate awareness and acceptance of the purposes and mission of the group, and when they stir their employees to look beyond their own self-interest for the good of the group. (p. 21)

According to Bass (1988), leaders employ both transactional and transformational leadership. His conceptualization of transactional and transformational leadership consisted of six leadership factors:

- inspirational leadership;
- intellectual stimulation;
- individualized consideration;
- contingent reward;
- management-by-exception; and
- laissez-faire leadership.

However, it should be noted that transformational leadership is far from a unitary concept. As an example, the concept of transformational leadership (including transactional practices) used by Leithwood and Jantzi (2000, p. 114) is comprised of the following dimensions:

- building school vision and goals;
- providing intellectual stimulation;
- offering individualized support;
- symbolizing professional practices and values;
- demonstrating high performance expectations;
- developing structures to foster participation in school decisions;
- staffing;
- instructional support;
- monitoring school activities; and
- community focus.

While these two conceptualizations partly differ, they both have in common a focus on leaders interacting with the people around them. In addition, the two definitions by Burns (1978) and Bass (1990) both underscore that this kind of leadership is primarily focused on leaders building up the capacity of those they work with, motivating them towards instilling change and transformation.

Transformational leadership has been suggested as an ideal leadership style for principals of schools considering substantial reforms as change management is considered a strength of transformational leaders (Leithwood & Jantzi, 2006). With regard to a possible impact on student achievement, more research is needed. Reviews of research have found transformational leadership to be inconclusive with regard to direct or

indirect impact on student achievement (Leithwood & Jantzi, 2005), and have found that transformational leadership in studies using direct effect designs had "small but positive and practically meaningful effects on student achievement" as well as having inconclusive results in indirect effect designs (Sun & Leithwood, 2012, p. 442).

Differences and paths towards the integration of instructional and transformational leadership

An often-made distinction regarding the aforementioned leadership theories is the more direct involvement of instructional school leaders in teaching and learning processes while transformational leaders typically seek to generate second-order effects (Hallinger, 2003), trying to improve the capacity of staff, who in turn produce first-order effects on learning. Day et al. (2016, p. 224) offer a similar distinction, viewing transformational leadership as emphasizing vision and inspiration and instructional leadership as establishing educational goals, planning the curriculum, and evaluating teachers and teaching.

Although the question whether there is a "best" kind of educational leadership is still being debated and studied, there is evidence with regard to the impact of instructional as opposed to transformational leadership on school organizations as well as student learning. In a meta-analysis, Robinson et al. (2008) found direct leaders' involvement in teaching and teacher learning to have the highest impact on student outcomes. More recently, Shatzer et al. (2014) demonstrated that instructional leadership accounts for more of the variance in educational achievement (45%) than transformational leadership (29%) and that instructional leadership also explains slightly more of the variation in progress scores (27%) than transformational leadership (22%).

Overall, however, the evidence can still be considered inconclusive. Whereas some researchers have found instructional leadership to be "effective" (Robinson, Hohepa, & Lloyd, 2009), others (e.g., Hallinger, 2003) esteem a focus on instructional leadership to be of limited value. Similarly, a criticism regarding transformational leadership is that such an approach alone is "an insufficient condition for measurable school improvement", lacking "a specific orientation towards student learning" (Hopkins, Stringfield, Harris, Stoll, & Mackay, 2014, p. 266). Marks and Printy (2003) view transformational leadership as vital for school reform, but view transformational leadership practices as having no bearing on student achievement as matters of teaching and learning are not necessarily front and center in a transformational leadership approach. Printy, Marks, and Bowers (2009) argue, like Hallinger (2003) as well, for a more holistic, integral leadership approach:

> In truth, although quantitative methods such as surveys permit the isolation of transformational and instructional forms based on the content of questions, these forms are likely to cohere in practice. The basic assumption of integral leadership therefore is that distinguishing between instructional leadership and other leadership facets is not very effective, primarily because it leads to fragmentation and segmentation. (p. 511)

Differential effects of leadership through inner school mediators

On the one hand, literature has frequently portrayed school principals as the decisive factor for matters of school improvement (Day & Sammons, 2013). On the other hand,

there is ample evidence that school principals are restricted by a myriad of demands, by often fragmented workdays or unplanned events, and that, consequently, their influence is limited or mediated by inner school factors (Easley & Tulowitzki, 2016; Pont, Nusche, & Moorman, 2008). Regarding how to model this for empirical studies, a strong case can be made for indirect effect models. The underlying hypothesis is that there is not (only) a direct effect of school leaders on student performance, but a mediated one. The school principal directly affects what could be considered intermediate variables, such as school culture, school climate, or the organizational structure within a school. These in turn are linked to student performance.

One of the dominant models of leadership influences on student achievement is Leithwood's model of leadership influences on student learning, originally developed as a framework for research on large-scale reform (Leithwood, Jantzi, & Mascall, 2002). It has been used for various empirical studies (Leithwood & Jantzi, 2006; Thoonen et al., 2011) and gone through several iterations (Leithwood, Harris, & Hopkins, 2008) and expansions in scope (Leithwood, Louis, Anderson, & Wahlstrom, 2004; Louis, Leithwood, Wahlstrom, & Anderson, 2010).

It presumes that school leadership influences working conditions as well as capacity beliefs and the motivation and commitment of school staff (see Figure 1). These three are viewed as (interdependent) key variables that in turn influence teaching practices, which in turn influence student learning outcomes. They were developed by Leithwood et al. (2002) based on previous modeling efforts (Rowan, 1996) as well as through a review of "theoretical and empirical accounts of the conditions influencing educators' motivations and capacities, as well as features of their work settings which facilitate the successful implementation of largescale reform" (Leithwood et al., 2002, p. 18).

Motivation and commitment

Motivation can be viewed as a collection of processes or qualities of a person which are future oriented and tend to be evaluative in nature (Ford, 1992, pp. 72ff). They are tied to a person's goals and beliefs. Motivation and commitment are viewed as similar concepts, both with the potential to be "energizing forces with implications for behavior" (Meyer, Becker, & Vandenberghe, 2004, p. 994). However, commitment is viewed as more binding in nature while motivation is seen as a broader term.

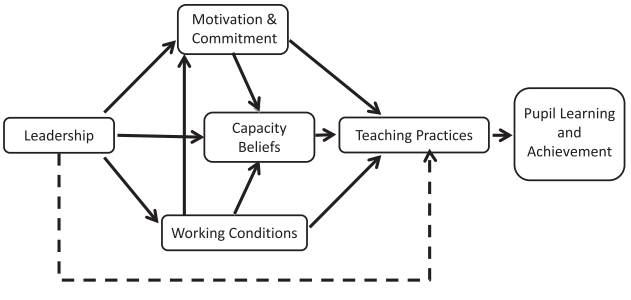


Figure 1. Model used in this study (based on Leithwood & Jantzi, 2006).

Capacity beliefs

Capacity refers to the belief of a person to be able to perform the tasks required for the job. The individual capacity belief is tied to the notion of self-concept and perceived self-efficacy. The actual capacity is tied to individual learning but also to the conditions for learning for the staff in schools (capacity building).

Working conditions

The working conditions describe the organizational and cultural context in which the persons operate and that are relevant for the collective learning in the organization. This includes facets like teacher interactions and cooperation and the school climate.

Thus, a corresponding model would assume an indirect effect of leadership on student learning with leadership directly impacting motivation and commitment, capacity beliefs, as well as working conditions in the school.

Design and methods

Source of data and context

The data for this study were collected by the Hamburg school inspection between 2012 and 2015 within the German federal state and city of Hamburg. In Germany, every child between the age of 6 and 15 is required to attend school. For the first years of school, they attend a comprehensive primary school (*Grundschule*). Following the primary school, the German school system allocates students of varying abilities into different types of schools.

In Germany, the responsibility for the education system, and hence its detailed organization, lies primarily with the federal states (*Bundesländer*, see Tulowitzki, 2015). In the federal state of Hamburg, children are entitled to attend nursery education from 1 until the age of 6 and to attend pre-schooling at the age of 5. Afterwards, primary education provides schooling from the age of 6 to 10. Following this, students are enrolled in the secondary schools at the lower secondary level (fifth grade) at an age of about 11. Since 2010, two types of secondary schools exist in Hamburg: (what could be considered) advanced secondary (*Gymnasium*) and comprehensive schools (*Stadtteilschule*).

During the 2015/2016 school year, there were 190 primary schools, 60 grammar schools, and 58 comprehensive schools in the federal state of Hamburg. During that school year, 161,789 students (including pre-school students) were enrolled in these schools. The sample for the subsequent analyses consists of $n = 126$ Schools (74 primary schools, 31 advanced secondary schools, *Gymnasium* in German, and 21 comprehensive schools, *Stadtteilschule* in German) and thus comprises nearly 40% of all schools within the system. All teachers of the inspected schools were asked to complete an online questionnaire on their instructional practices and motivational factors as well as on their perceptions of working conditions and school leadership at their school. A total of $n = 3{,}746$ teachers participated in the survey, resulting in a response rate of 65.2% ($N = 5{,}745$). The teaching staff within a single school ranged in size from 7 to 85 ($m = 30$).

Measures and instruments

Teachers were asked to complete a questionnaire with a total of 114 items which were divided into seven sections. Detailed in-depth information concerning the measures and instruments can be found in the methodological documentation of the Hamburg school inspection (Pietsch, Scholand, Graw, Hengstmann, & Kulin, 2013). For the present study, five sections of the teacher questionnaire are relevant.

School leadership

For gathering information related to leadership behavior, two instruments were used. First, a 21-item short form of the Multifactor Leadership Questionnaire (MLQ; see Bass & Avolio, 1995) was administered to measure transformational leadership within the full range of leadership approach. Thus, teachers rated six to nine items on a 4-point Likert-type scale (1 = *(almost) never*, 4 = *(almost) always*) with regard to the three main scales of the MLQ transformational (Cronbach's alpha = .93, McDonald's omega = .94), transactional (Cronbach's alpha = .75, McDonald's omega = .75), and laissez-faire/avoidant (Cronbach's alpha = .83, McDonald's omega = .86) leadership.

Second, a scale from the Teaching and Learning International Survey (TALIS; see Organisation for Economic Co-operation and Development, 2009) was adopted and employed to measure instructional leadership. This eight-item scale is a derivative of Hallinger's (1994) Principal Instructional Management Rating Scale (PIMRS) and enables the identification of leadership behavior regarding the management of school goals, instructional management, and direct supervision of teachers. These items were rated by the teachers on a 4-point Likert-type scale (1 = *(almost) never*, 4 = *(almost) always*) and statistically combined into a single variable named instructional leadership (Cronbach's alpha = .91, McDonald's omega = .91).

Instructional practices

For measuring instructional practices, the teacher questionnaire draws on teacher effectiveness research (Pianta & Hamre, 2009) and distinguishes three dimensions: (a) structure and classroom management; (b) supportive, student-oriented classroom climate; and (c) cognitive activation with challenging content. These dimensions are also surveyed by using scales from TALIS, which are named structuring, student orientation, and enhanced activities. Here, structuring is indicated by five items (Cronbach's alpha = .51, McDonald's omega = .62)[1] like "I explicitly state learning goals." Student orientation in turn is measured by four items (Cronbach's alpha = .55, McDonald's omega = .73). An example item is: "I give different work to the students that have difficulties learning and/or to those who can advance faster." Finally, enhanced activities are measured with four items too (Cronbach's alpha = .70, McDonald's omega = .82). An example item is: "Students hold a debate and argue for a particular point of view which may not be their own." Each item had to be rated on a 4-point Likert-type scale (1 = *never or hardly never*, 4 = *in about three quarters of my lessons*). Due to the fact that the Hamburg school inspection also conducts systematic random-sample-based 20-min classroom observations at every school, it was also possible to validate the information provided by the teachers in the questionnaire. The classroom observations were conducted using a 30-item grid (Pietsch, 2010). A total of n = 8,271 (average n per school = 66) classroom observations were available for analysis at

the school level. As previous research on assessing instructional processes has demonstrated (Clausen, 2002; Kunter & Baumert, 2006), the correlations of the different perspectives can be expected (and turned out) to be moderate, with correlation coefficients ranging from $r = .15$ to $r = .40$.

The unidimensional item response theory-based scale of the observations correlated with classroom management at $r = .35$ ($p < 0.001$), with student orientation at $r = .19$ ($p < 0.05$), and with enhanced activities at $r = .18$ ($p < 0.05$). The correlations of the grids marker item indicating the quality of classroom management ("Instructions and explanations in class are given in an appropriate, clear, and precise manner") with the three scales from the teacher survey were: classroom management ($r = .28$, $p < 0.001$), student orientation ($r = .24$, $p < 0.05$), and enhanced activities ($r = .04$, $p > 0.05$). The correlations of the grids marker item indicating the quality of student orientation ("The experiences and/or interests of students are taken into account when planning classes") with the three scales from the teacher survey were: classroom management ($r = .17$, $p < 0.05$), student orientation ($r = .19$, $p < 0.05$), and enhanced activities ($r = -.16$, $p > 0.05$). The correlations of the grids marker item, indicating the quality of cognitive activation with challenging content ("Classroom instructions are designed to allow room for flexibility and are not fixated on one right answer/lesson") with the three scales from the teacher survey were: classroom management ($r = .06$, $p > 0.05$), student orientation ($r = .14$, $p > 0.05$), enhanced activities ($r = .27$, $p < 0.001$).

Working conditions

Two constructs were used for the assessment of the working conditions within schools: a scale measuring teacher collaboration and a scale measuring teacher participation, in the sense of shared leadership. The original 20-item teacher collaboration scale has been developed by Steinert et al. (2006), from which only seven items are used in this study. These items are indicating the dimensions *programmatic cooperation* and *social cohesion* on a unidimensional scale (Cronbach's alpha = .83, McDonald's omega = .83). Example items are: "We work together to create common strategies to overcome professional difficulties" and "It's a natural part of our work to visit each other in the classroom to see each other teach." Responses were given on a 4-point Likert-type scale (1 = *I do not agree at all*, 4 = *I strongly agree*).

For measuring teacher participation, a seven-item scale was developed by the Hamburg school inspection. This scale surveys teachers' perceptions of their influence over and participation in schoolwide decisions in accordance with the definition of shared leadership as provided by Wahlstrom and Louis (2008). The items were rated on a 4-point Likert-type scale (1 = *I do not agree at all*, 4 = *I strongly agree*). Example items are: "Teachers have an effective role in school-wide decision making" and "Teachers have significant input into plans for professional development and growth." Internal consistency was Cronbach's alpha = .80 and McDonald's omega = .81.

Motivational factors

For capturing motivational factors, two scales were used. The items of both scales were rated on a 4-point Likert-type scale (1 = *I do not agree at all*, 4 = *I strongly agree*). One

scale captures the affective commitment of teachers with three items drawn from the German translation of the Organizational Commitment Questionnaire (OCQ; Maier & Woschée, 2002) as adapted for schools by Lipowsky, Faust, and Greb (2009). Cronbach's alpha was .91, McDonald's omega was .92, and an example item is: "I am extremely glad that I chose this school over others I was considering at the time I joined." The other scale captures the job satisfaction of teachers with six items and has been developed by the Hamburg school inspection. Cronbach's alpha was .88, McDonald's omega was .88, and an example item is: "Overall, I am satisfied with the working conditions at my school."

Capacity (beliefs)

Six items were used for measuring the capacity (beliefs) of teachers, their beliefs of being capable of accomplishing goals regarding innovations and improvement when it comes to their own classroom practices. This scale was developed by the Hamburg school inspection with reference to the groundwork of and instruments developed by Leithwood and Jantzi (2006). The items were rated on a 4-point Likert-type scale (1 = *I do not agree at all*, 4 = I *strongly agree*). Example items are: "Usually, I have the knowledge and skills I need to implement recommended teaching innovations and improvements by the school authority into my own classroom practices" and "Usually, I have sufficient opportunities to think and talk about how to implement recommended teaching innovations and improvements by the school authority into my own classroom practices." Internal consistency was Cronbach's alpha = .82 and McDonald's omega = .83.

Method of analysis

In the current study, data were analyzed using structural equation models in Mplus 7 (Muthén & Muthén, 2012). For testing the hypothesized model as presented in Figure 1 in the first phase of data analysis, a full maximum likelihood estimation which also took the complex structure of the data as well as missing values into account has been conducted. The analysis followed the suggestions made by Anderson and Gerbing (1988). Thus, a sequential testing procedure was conducted: First, the measurement models for the latent variables were specified and afterwards constrained within a saturated structural model, where all possible paths are linked among the latent variables. Next, the theory-based model (composite) was calculated. That provided the opportunity to decompose the overall global fit indices into a measurement and a path component (O'Boyle & Williams, 2011).

To assess the fit of the models, the classic fit indices comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) as provided by Mplus 7 were used. Acceptable fit would be indicated by a CFI over .90, a RMSEA less than .08, and a SRMR less than .08 (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). Furthermore, by following the suggestions made by McDonald and Ho (2002), the RMSEA-P was calculated as the fit of the path model by using the sample size, the chi-square values, and degrees of freedom. For the RMSEA-P, values less than .05 represent a close approximate fit, values between .05 and .08 suggest a reasonably approximate fit, values between .08 and .10 are indicative of a mediocre fit, and values greater than .10 suggest a poor fit (O'Boyle & Williams, 2011).

Results

Modeling issues

Modeling leadership

Table 1 presents the latent correlations of the leadership facets. All facets are highly correlated, with *r* coefficients from – .77 to .97 (all $p < 0.01$), demonstrating that they are in all likelihood not orthogonal.

Hence, it was first examined what kind of measurement model would best fit the data. For this purpose, five alternative models were tested:

- *One leadership factor*: This model assumes that a principal either does or does not exhibit leadership. All items, the ones of the MLQ indicating the full range of leadership as well as the ones from the TALIS scale indicating instructional leadership, are expected to load on a global general factor *g*.
- *Two correlated factors – MLQ versus instructional leadership*: It is assumed that principals could act highly person oriented within the full range of the leadership framework as well as primarily task oriented as instructional leaders. Thus, this model assumes a simple, unidimensional structure of the MLQ that is correlated with the instructional leadership facet.
- *Four correlated factors – MLQ facets versus instructional leadership*: This model takes into account the three leadership styles as described within the framework of the MLQ, and their particular correlations among each other, as well as the ones with the instructional leadership facet. Thus, it is assumed that a principal could act transformational, transactional, and passive-avoidant, as well as instructional.
- *Second-order g factor – g versus MLQ facets versus instructional leadership*: It is implied that a higher order factor "active leadership" accounts for the commonality among the four lower order factors transformational, transactional, passive-avoidant, and instructional leadership. Hence, it is first assumed that a higher degree of general leadership activity is attended by higher values on all four leadership facets simultaneously. Second, it is assumed that this influence is mediated by each of the facets. And third, it is assumed that each first-order leadership facet also represents facet-specific characteristics that account for individual differences of principals.
- *Bi-factor – g versus MLQ facets versus instructional leadership*: This model assumes that principals can exhibit domain-specific leadership behavior that is independent from a global (*g*) factor dubbed "leadership core" as well as active leadership on its own. As the (*g*) factor and the facets are orthogonal (i.e., uncorrelated) to each other, it is expected that the first-order leadership facets are still existing after partialling out the common variance and may have incremental effects on potential covariates that go above the ones deriving from the global factor.

Table 1. Latent correlations of leadership facets.

	Transformational	Transactional	Laissez-faire	Instructional
Transformational	1			
Transactional	.97***	1		
Laissez-faire	–.87***	–.91***	1	
Instructional	.78***	.83***	–.77***	1

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 2. Summary of overall fit measures for each of the five factor models.

	1 factor	2 factors: MLQ vs. instructional	4 factors: MLQ facets vs. instructional	second order: *g* and MLQ facets and instructional	bifactor: *g* and MLQ facts and instructional
Chi2/*df*	10.018/377	7.972/376	7.137/371	7.130/373	4.285/348
TLI	0.811	0.850	0.865	0.866	0.916
CFI	0.824	0.861	0.877	0.877	0.928
RMSEA	0.084	0.074	0.071	0.070	0.056
SRMR	0.065	0.057	0.055	0.056	0.043

Fit indices for these five models are presented in Table 2. The bi-factor model produced the best fit, whereas the unidimensional solution produced the worst absolute model fit. The χ^2-difference tests revealed that the two-factor model was a major improvement compared to the unidimensional solution ($p < 0.01$). Modeling school leadership as a bi-factor model by taking into account that the facets representing the specific leadership constructs are nested within the general factor was another major improvement over this model ($p < 0.01$).

Modeling instructional practices

In our sample, 12 out of 13 items were significantly loading ($p < 0.05$) on more than one dimension of instructional practice. Thus, a highly restrictive independent cluster model, in which cross loadings of the items between the dimensions are set to zero, seems inappropriate. If high cross loadings do exist and are not represented by the measurement model, the model fit should not be acceptable (Asparouhov, Muthén, & Morin, 2015), and correlational as well as regression coefficients may be overestimated (Zhang, 2012). Hence, due to the multidimensionality of the items indicating the teaching practices, three different models were tested for describing the dependent variable(s) in the theoretically best possible way while taking into account the statistical challenges.

- *One instructional practice factor*: All 13 items are loading on a single factor. Thus, it is assumed that it is not possible to distinguish between the three proposed factors measuring teaching practices.
- *Three instructional practices factors without cross loadings*: The three developed measures, structure and classroom management, supportive, student-oriented classroom climate, and cognitive activation with challenging content, were kept and modeled as a traditional independent clusters confirmatory factor analysis. Hence, cross loadings were not allowed, and it is assumed that all three factors are unidimensional.
- *Three instructional practices factors with cross loadings*: The three developed measures, structure and classroom management; supportive, student-oriented classroom climate, and cognitive activation with challenging content, were kept and modeled as an exploratory structural equation model with three factors, which is equivalent to a three-dimensional exploratory factor analysis. Hence, cross loadings were allowed, and it is assumed that all three factors are multidimensional.

Fit indices for the one-factor model were: $\chi^2 = 1863.93$, $df = 65$, $p < 0.01$; and a CFI, RMSEA, and SRMR of 0.64, 0.09, and 0.07, respectively. Fit indices for the three-factor model without cross loadings were: $\chi^2 = 1493.63$, $df = 62$, $p < 0.01$; and a CFI, RMSEA,

and SRMR of 0.71, 0.08, and 0.06, respectively. And fit indices for the three-factor model with cross loadings were: $\chi^2 = 378.76$, $df = 42$, $p < 0.01$; and a CFI, RMSEA, and SRMR of 0.93, 0.05, and 0.03, respectively. Thus, the three-dimensional model with cross loadings fitted the data best by a large margin. As χ^2-difference tests revealed, this model was a major improvement compared to the unidimensional as well as to the three-dimensional independent clusters confirmatory factor model (both $p < 0.01$).

Testing the hypothesized causal model

Next, the hypothesized causal model was tested. First, all hypothesized paths were modeled as directed regressions, with the exception of the relations between the latent variables measuring the working conditions as well as the one measuring motivational aspects, which were modeled as correlations. In this regard, it is noteworthy to mention that the model fits the correlational data and thus causality cannot be established even if the model fits well. As the context of a school may have its own influence on leadership behavior as well as on other inner-school variables like teacher satisfaction and commitment (Hallinger & Murphy, 1986; Leithwood & Jantzi, 2009), the size of the school, as measured by the number of students and the number of teachers, as well as the type of school (which in Germany is highly correlated with the social composition of the school's student body, its socioeconomic status, see Ehmke & Jude, 2010), were included as control variables. Figure 2 presents the results for the structural model. For better readability, only statistically significant ($p < 0.05$) paths are visualized.

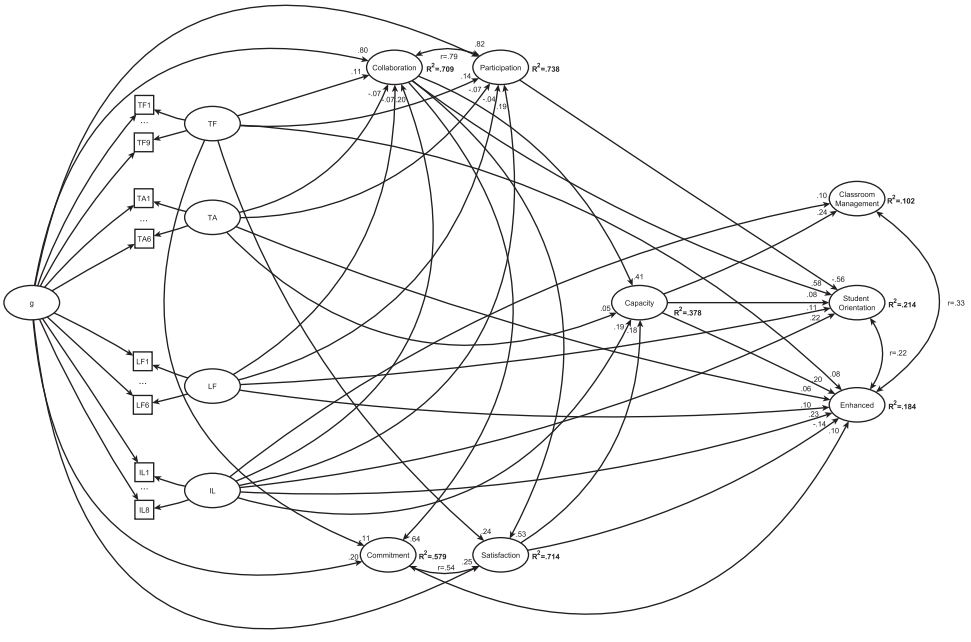


Figure 2. Path diagram with standardized parameter estimates and coefficients of determination.

Notes: Leadership facets: *g* = *g*-factor/leadership core, TF = transformational, TA = transactional, LF = laissez-faire, IL = instructional. Model controlled for school size and school type. All reported paths are least significant at the level of $p < 0.05$.

Fit indices of the model were: $\chi^2 = 9363.43$, $df = 1991$, $p < 0.01$; and a CFI, RMSEA, and SRMR of 0.92, 0.03, and 0.04, respectively, indicating a good model fit. By also modeling the saturated model ($\chi^2 = 9383.11$, $df = 1999$, $p < 0.01$), it was possible to calculate the fit index RMSEA-P for examining the quality of the path model independently from the one of the measurement models. RMSEA-P was 0.08, indicating a reasonably approximate fit of the path model.

Results of the structural model show that principals have a strong influence on the working conditions of teachers ($R^2_{Collaboration} = .709$, $R^2_{Participation} = .738$) and in interaction with those factors have a somewhat smaller influence on their staff's motivation ($R^2_{Commitment} = .579$, $R^2_{Satisfaction} = .714$). The impact of the model variables on the capacity beliefs of teachers is reasonably high ($R^2_{Capacity} = .378$). According to Cohen (1988), all these effects can be considered large. The instructional practices of teachers are, however, influenced less by the model variables ($R^2_{Classroom\ Management} = .102$, $R^2_{Student\ Orientation} = .214$, $R^2_{Enhanced\ Practices} = .184$); nonetheless, the effects can be considered to be medium effects as defined by Cohen.

The paths are in line with Leithwood's (Leithwood & Jantzi, 2006) as well as Hallinger's (2011) models and findings and, thus, can be viewed as a(nother) good validation. Motivational factors, for example, are influenced directly by a general active leadership, a leadership core (*g*), as well as by a transformational leadership behavior. The working conditions of teachers are directly influenced by all of the leadership facets as well as an overall leadership core behavior (*g*).

However, theory assumes principals to become effective directly as well as indirectly. Thus, also indirect effects as well as total effects were tested next. Table 3 first presents the total effects of the leadership factors on the mediators.

The leadership core (*g*) has the biggest total effect on all mediators with β coefficients from .419 to .822. But there are also incremental effects of the first-order leadership facets on the mediators detectable: As a result, transformational leadership has a strong incremental total effect on the job satisfaction of teachers ($\beta = .304$) and instructional leadership on teacher collaboration ($\beta = .202$), as well as on their innovation capacity ($\beta = .281$). It is also noteworthy that a transactional as well as a passive-avoidant leadership behavior of principals always goes along with negative effects on the mediators (if statistically significant).

The effects of leadership behavior on the instructional practices of teachers are in turn predominantly not verifying the theoretical assumptions as the mediator analysis reveals (see Table 4). Only for instructional leadership, small statistically significant

Table 3. Total effects of leadership facets and leadership core on mediation variables.

	Collaboration	Participation	Commitment	Job Satisfaction	Capacity (Beliefs)
g	0.802***	0.822***	0.650***	0.715***	0.419***
TF	0.111***	0.140***	0.171***	0.304***	0.127***
TA	−0.066**	−0.072**	−0.063**	−0.074**	0.007
LF	−0.070**	−0.044*	−0.020	−0.017	0.015
IL	0.202***	0.188***	0.109***	0.099***	0.281***
R^2	0.709	0.738	0.579	0.714	0.378

Notes: Leadership facets: *g* = *g*-factor/leadership core, TF = transformational, TA = transactional, LF = laissez-faire, IL = instructional.
*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 4. Direct and indirect effects of leadership facets and leadership core on instructional practices.

	Classroom Management		Student Orientation		Enhanced Activities	
	indirect	direct	indirect	direct	indirect	direct
g	0.073	0.079	−0.012	0.102	0.021	0.085
TF	0.005	−0.001	−0.024	0.018	−0.012	0.081**
TA	0.005	0.053	0.008	0.004	0.009	0.059**
LF	0.002	−0.010	−0.014	0.109**	−0.003	0.101***
IL	0.069**	0.096**	0.027	0.217***	0.052***	0.232***
R^2	0.102		0.214		0.184	

Notes: Leadership facets: *g* = *g*-factor/leadership core, TF = transformational, TA = transactional, LF = laissez-faire, IL = instructional.
*$p < 0.05$ **$p < 0.01$. ***$p < 0.001$.

indirect effects were detectable in relation to classroom management practices ($\beta = .069$) and enhanced activities ($\beta = .052$).

According to the analysis, the effects of leadership behavior on the instructional practices of teachers are first and foremost a result of direct interactions. Again, the largest direct effect on all practice types stems from instructional leadership ($\beta_{Classroom\ Management} = .096$, $\beta_{Student\ Orientation} = .217$, $\beta_{Enhanced\ Activities} = .232$). For the leadership core (*g*), neither direct nor indirect effects were detectable. Finally, when it comes to cognitive activation with challenging content, it seems to be necessary that all first-order leadership facets fall into place. But once again, effects are mainly direct ones with β coefficients from .059 for transactional to .232 for instructional leadership behavior.

Discussion and conclusion

The findings revealed that the behavior of principals generally affects the instructional practices of teachers directly as well as indirectly. On the one hand, instructional leadership had strong direct as well as small indirect effects on all instructional dimensions, specifically classroom management, supportive, student-oriented classroom climate, and cognitive activation with challenging content. On the other hand, the transactional, transformational, and laissez-faire leadership facets only had moderate significant effects on cognitive activating instruction. Statistically significant indirect effects for those leadership facets were, with the exception of laissez-fair leadership on student activation, not detectable. Leadership, along with work setting, innovation capacity, and teacher motivation explained approximately 10 to 21% of the variation in teachers' instructional practices.

The leadership core appears to be of vital importance, having a sizeable effect on all mediators along the path towards educational achievement. What this leadership core is comprised of precisely is currently undeterminable. It echoes the general assumptions and findings of the presence of what has been referred to as the "basics of successful school leadership" (Leithwood et al., 2004, pp. 23ff) or "basic leadership practices" (Leithwood et al., 2008). However, the so-called "basics of successful leadership" can be seen more in line with a transformational approach to leadership, consisting of setting directions, developing people, and redesigning the organization as categories of leadership activities. These kinds of activities are already accounted for through the MLQ in our study. Our findings demonstrate that the leadership core is something that appears not to be mapped by the previously mentioned (concepts of) basics of successful leadership or basic leadership practices.

The findings further demonstrate that school principals have a strong influence on the work setting, innovation capacity, and motivation and a considerably smaller influence on the instructional practices of their staff, with mechanisms being first and foremost direct ones. Thus, it is obvious that school leaders also have an influence on the likelihood that teachers will change their instructional practices, particularly if they are directly involved in the design and implementation of curriculum, instruction, and assessment practices and if they directly exert influence on the classroom practices of their teachers. It is also evident that for changing more complex practices, a well-balanced and coordinated combination of leadership styles is necessary – one may call this "integrated leadership".

The results of the study make a strong case for a (more) holistic leadership approach in practice, but also in the training of school leaders. Instead of focusing on a certain kind of leadership – whatever it may be – or viewing one kind of leadership as superior, an approach that is inclusive of various facets of leadership seems more prone to success.

With regard to promising avenues of research, a recommendation is to use model-driven research and a design that allows for the analysis of distinct facets of leadership. Furthermore, the search for a "leadership core" that transcends various leadership practices seems promising to better understand what makes a school principal an effective leader.

Limitations of the study

This article presents a model-driven research on the effects of leadership on teaching practices and thus seeks to describe and analyze how certain leadership behaviors influence the key determinant of student's experience and outcomes of schooling. The study thereby addresses several criticisms and desiderata of current research into the effects of educational leadership (Scheerens, 2012). Hence, intermediary variables were aptly chosen, and the measures were well defined and measured by applying established, well researched, and validated instruments. Furthermore, it was possible to validate the self-reported teaching styles of teachers by contrasting them with external observer ratings on the instructional quality at the school level.

Nonetheless, the study is limited by two main factors: (a) It is a correlational study; therefore, causality is inferred but cannot be demonstrated. (b) No achievement data were used, and thus it was not possible to link leadership and teaching to students' outcomes. Consequently, it was not possible to model and analyze school improvement or improvement of instructional quality as the used data are derived from a cross-sectional survey. Furthermore, it was impossible to check whether differences of leadership are accompanied by a change in students' achievement, as measured, for example, by value-added scores.

Another limitation is that modeling distinct instructional practices as an independent cluster model seems inappropriate as indicator items are multidimensional. Thus, if one is interested in evaluating the influence of school variables on specific teaching practices as well as the effects of specific practices on student outcomes, it seems necessary to take the corresponding statistical challenges into account. Possible models within a SEM framework could be CFA models with cross loadings, bi-factor-models, exploratory structural equation models, or Bayesian structural equation models with cross loadings. In the study presented herein, neither a bi-factor nor a Bayesian model with cross loadings of instructional practices has been tested. A promising avenue of further

research could therefore be that future studies evaluate the differential effects of such models in contrast to the one reported in this study.

Note

1. In contrast to the other scales, the differences of alpha and omega reliability are strikingly large. The omega reliability of structure and classroom management was ω = .62, the omega reliability of supportive, student-oriented classroom climate was ω = .73, and the omega reliability of cognitive activation with challenging content was ω = .82. Thus, between 62 and 82% of the factor-specific variance is attributable to the particular indicators. As the internal consistency as measured by coefficient alpha refers to the homogeneity of the items in the measure or the extent to which item responses correlate with the total test score of the dimension of instructional practice, the aforementioned mediocre internal consistency of coefficient alpha points towards the instability of the traits as measured by the particular items. However, one would expect some degree of cross loadings because the factors are conceptually related (Hamre, Hatfield, Pianta, & Jamil, 2014). In such cases, the reliability of the scales may be dramatically underestimated by coefficient alpha (Graham, 2006), whereas coefficient omega will lead to a more accurate correction (Revelle & Zinbarg, 2009). In our sample, 12 out of 13 items were significantly loading ($p < 0.05$) on more than one dimension. Thus, because of the multidimensional data, coefficient omega (ω) by McDonald (1999) can be considered a more accurate measure of reliability for the three scales measuring instructional practices.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Marcus Pietsch is a senior researcher at the Institute for Education Monitoring Hamburg and currently holds a postdoctoral scholarship in Empirical Educational Research awarded by the Leuphana University Lueneburg in Germany. His main research interests include accountability, educational as well as teacher effectiveness, and data-driven school improvement, primarily with a focus on measurement and modeling issues. He is a chair of the International Congress for School Effectiveness and Improvement (ICSEI) Methods of Researching Educational Effectiveness (MoRE) network.

Pierre Tulowitzki is an assistant professor and the head of the Department of International Educational Leadership and Management at the Ludwigsburg University of Education in Germany. He is the German director of the accredited international Master program "International Education Management", a joint program of the Ludwigsburg University and the Helwan University, Egypt. His research interests include educational leadership, school improvement and accountability, and educational change. He is a member of the board of the International Congress for School Effectiveness and Improvement (ICSEI).

ORCID

Marcus Pietsch http://orcid.org/0000-0002-9836-6793
Pierre Tulowitzki http://orcid.org/0000-0001-8809-2541

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, *41*, 1561–1577. doi:10.1177/0149206315591075

Bass, B. (1988). The inspirational processes of leadership. *Journal of Management Development*, *7*(5), 21–31. doi:10.1108/eb051688

Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, *18*(3), 19–31. doi:10.1016/0090-2616(90)90061-S

Bass, B. M., & Avolio, B. J. (1995). *MLQ Multifactor Leadership Questionnaire. Technical report*. Redwood City, CA: Mind Garden.

Burns, J. M. (1978). *Leadership*. New York, NY: Harper & Row.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?: Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität* [Instruction: A question of perspective? Empirical analyses regarding fit, construct and validity]. Münster, Germany: Waxmann.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: US Department of Helath, Education, and Welfare, Office of Education/National Center for Education Statistics.

Day, C., Gu, Q., & Sammons, P. (2016). The impact of leadership on student outcomes: How successful school leaders use transformational and instructional strategies to make a difference. *Educational Administration Quarterly*, *52*, 221–258. doi:10.1177/0013161X15616863

Day, C., & Sammons, P. (2013). *Successful leadership: A review of the international literature*. Berkshire, UK: CfBT Education Trust.

Easley, J., II, & Tulowitzki, P. (Eds.). (2016). *Educational accountability – International perspectives on challenges and possibilities for school leadership*. London, UK: Routledge.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, *37*, 15–18, 20.

Ehmke, T., & Jude, N. (2010). Soziale Herkunft und Kompetenzerwerb [Social background and the acquisition of competencies]. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, … P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* [PISA 2009. A look back after a decade] (pp. 231–254). Münster, Germany: Waxmann.

Fidler, B. (1997). School leadership: Some key ideas. *School Leadership & Management*, *17*, 23–38. doi:10.1080/13632439770140

Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: SAGE.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930–944. doi:10.1177/0013164406288165

Hallinger, P. (1994). *A resource manual for the Principal Instructional Management Rating Scale (PIRMS Manual 2.2)*. Nashville, TN: Center for the Advanced Study of Educational Leadership.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, *33*, 329–352. doi:10.1080/0305764032000122005

Hallinger, P. (2011). Leadership for learning: Lessons from 40 years of empirical research. *Journal of Educational Administration*, *49*, 125–142. doi:10.1108/09578231111116699

Hallinger, P., & Murphy, J. F. (1986). The social context of effective schools. *American Journal of Education*, *94*, 328–355. doi:10.1086/443853

Hallinger, P., & Wang, W.-C. (2015). *Assessing instructional leadership with the Principal Instructional Management Rating Scale*. London, UK: Springer.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*, 1257–1274. doi:10.1111/cdev.12184

Hopkins, D., Stringfield, S., Harris, A., Stoll, L., & Mackay, T. (2014). School and system improvement: A narrative state-of-the-art review. *School Effectiveness and School Improvement*, *25*, 257–281. doi:10.1080/09243453.2014.885452

Horng, E. L., Klasik, D., & Loeb, S. (2010). Principal's time use and school effectiveness. *American Journal of Education*, *116*, 491–523. doi:10.1086/653625

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

Huber, S. G., Gördel, B.-M., Kilic, S., & Tulowitzki, P. (2016). Accountability in the German school system: More of a burden than a preference. In J. Easley II & P. Tulowitzki (Eds.), *Educational accountability – International perspectives on challenges and possibilities for school leadership* (pp. 165–183). London, UK: Routledge.

Kmetz, J. T., & Willower, D. J. (1982). Elementary school principals' work behavior. *Educational Administration Quarterly*, *18*(4), 62–78. doi:10.1177/0013161X82018004006

Krug, S. E. (1992). Instructional leadership: A constructivist perspective. *Educational Administration Quarterly*, *28*, 430–443. doi:10.1177/0013161X92028003012

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231–251. doi:10.1007/s10984-006-9015-7

Lee, J. C.-K., & Dimmock, C. (1999). Curriculum leadership and management in secondary schools: A Hong Kong case study. *School Leadership & Management*, *19*, 455–481. doi:10.1080/13632439968970

Leithwood, K. A., Begley, P. T., & Cousins, J. B. (1990). The nature, causes and consequences of principals' practices: An agenda for future research. *Journal of Educational Administration*, *28*(4), 5–31. doi:10.1108/09578239010001014

Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership & Management*, *28*, 27–42. doi:10.1080/13632430701800060

Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration*, *38*, 112–129. doi:10.1108/09578230010320064

Leithwood, K., & Jantzi, D. (2005). A review of transformational school leadership research 1996–2005. *Leadership and Policy in Schools*, *4*, 177–199. doi:10.1080/15700760500244769

Leithwood, K., & Jantzi, D. (2006). Transformational school leadership for large-scale reform: Effects on students, teachers, and their classroom practices. *School Effectiveness and School Improvement*, *17*, 201–227. doi:10.1080/09243450600565829

Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. *Review of Educational Research*, *79*, 464–490. doi:10.3102/0034654308326158

Leithwood, K., Jantzi, D., & Mascall, B. (2002). A framework for research on large-scale reform. *Journal of Educational Change*, *3*, 7–33. doi:10.1023/A:1016527421742

Leithwood, K., Jantzi, D., & Steinbach, R. (1999). *Changing leadership for changing times*. Philadelphia, PA: Open University Press.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. New York, NY: The Wallace Foundation.

Lipowsky, F., Faust, G., & Greb, K. (2009). *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE) – Teil 1* (Vol. 1) [Documentation of the instruments used for data collection in the project "Development of personalities and abilities to learn among primary school students" (PERLE) – Part 1 (Vol. 1)]. Frankfurt am Main, Germany: GFPF, DIPF.

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Learning from leadership: Investigating the links to improved student learning: Final report of research findings*. New York, NY: The Wallace Foundation.

MacBeath, J. (2003). The alphabet soup of leadership. *Inform, 2*. Retrieved from https://www.educ.cam.ac.uk/centres/lfl/about/inform/PDFs/InForm_2.pdf

Maier, G. W., & Woschée, R. M. (2002). Die affektive Bindung an das Unternehmen: Psychometrische Überprüfung einer deutschsprachigen Fassung des Organizational Commitment Questionnaire (OCQ) von Porter und Smith (1970) [The affective bond to the enterprise: Psychometric validation of the German version of the organizational commitment questionnaire (OCQ) of Porter and Smith (1970)]. *Zeitschrift für Arbeits- und Organisationspsychologie*, *46*, 126–136. Retrieved from https://pub.uni-bielefeld.de/publication/2492558

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly*, *39*, 370–397. doi:10.1177/0013161X03253412

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 320–341. doi:10.1207/s15328007sem1103_2

Martin, W. J., & Willower, D. J. (1981). The managerial behavior of high school principals. *Educational Administration Quarterly*, *17*(1), 69–90. doi:10.1177/0013161X8101700105

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. doi:10.1037/1082-989X.7.1.64

Meyer, J. P., Becker, T. E., & Vandenberghe, C. (2004). Employee commitment and motivation: A conceptual analysis and integrative model. *Journal of Applied Psychology*, *89*, 991–1007. doi:10.1037/0021-9010.89.6.991

Muthén, L. K., & Muthén, B. O. (2012). Mplus software (Version 7) [Computer Software]. Los Angeles, CA: Authors.

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology*, *96*, 1–12. doi:10.1037/a0020539

Organisation for Economic Co-operation and Development. (2009). *Creating effective teaching and learning environments: First results from TALIS*. Paris, France: Author.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. doi:10.3102/0013189X09332374

Pietsch, M. (2010). Evaluation von Unterrichtsstandards [Evaluation of classroom teaching standards]. *Zeitschrift für Erziehungswissenschaft*, *13*, 121–148. doi:10.1007/s11618-010-0113-z

Pietsch, M., Scholand, B., Graw, S., Hengstmann, E., & Kulin, S. (2013). *Skalenhandbuch der Schulinspektion Hamburg: Fragebögen für Pädagoginnen und Pädagogen, Eltern und Schülerinnen und Schüler* [Scale handbook of the Hamburg school inspection: Questionnaires for teachers, parents and students]. Hamburg, Germany: Institut für Bildungsmonitoring und Qualitätsentwicklung.

Pont, B., Nusche, D., & Moorman, H. (2008). *Improving school leadership. Volume 1: Policy & Practice*. Paris, France: OECD.

Printy, S. M., Marks, H. M., & Bowers, A. J. (2009). Integrated leadership: How principals and teachers share transformational and instructional influence. *Journal of School Leadership*, *19*, 504–532.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Robinson, V., Hohepa, M., & Lloyd, C. (2009). *School leadership and student outcomes: Identifying what works and why: Best evidence synthesis iteration*. Wellington, New Zealand: Ministry of Education.

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*, 635–674. doi:10.1177/0013161X08321509

Rowan, B. (1996). Standards as incentives for instructional reform. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 195–225). San Francisco, CA: Jossey-Bass.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.

Scheerens, J. (Ed.). (2012). *School leadership effects revisited: Review and meta-analysis of empirical studies*. Dordrecht, The Netherlands: Springer.

Shatzer, R. H., Caldarella, P., Hallam, P. R., & Brown, B. L. (2014). Comparing the effects of instructional and transformational leadership on student achievement: Implications for practice. *Educational Management Administration & Leadership*, *42*, 445–459. doi:10.1177/1741143213502192

Spillane, J. P., & Hunt, B. R. (2010). Days of their lives: A mixed-methods, descriptive analysis of the men and women at work in the principal's office. *Journal of Curriculum Studies*, *42*, 293–331. doi:10.1080/00220270903527623

Steinert, B., Klieme, E., Maag Merki, K., Döbrich, P., Halbheer, U., & Kunz, A. (2006). Lehrerkooperation in der Schule: Konzeption, Erfassung, Ergebnisse [Collaboration of teachers in school: Conceptualization, measurement, results]. *Zeitschrift für Pädagogik*, *52*(2), 185–204.

Sun, J., & Leithwood, K. (2012). Transformational school leadership effects on student achievement. *Leadership and Policy in Schools*, *11*, 418–451. doi:10.1080/15700763.2012.681001

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly*, *47*, 496–536. doi:10.1177/0013161X11400185

Tulowitzki, P. (2013). Leadership and school improvement in France. *Journal of Educational Administration*, *51*, 812–835. doi:10.1108/JEA-03-2012-0026

Tulowitzki, P. (2015). The development of educational leadership and teaching professions in Germany. *ECPS – Educational, Cultural and Psychological Studies*, *11*, 45–55. doi:10.7358/ecps-2015-011-tulo

Wahlstrom, K. L., & Louis, K. S. (2008). How teachers experience principal leadership: The roles of professional community, trust, efficacy, and shared responsibility. *Educational Administration Quarterly*, *44*, 458–495. doi:10.1177/0013161x08321502

Wildy, H., & Dimmock, C. (1993). Instructional leadership in primary and secondary schools in Western Australia. *Journal of Educational Administration*, *31*, 43–62. doi:10.1108/09578239310041873

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, *36*, 375–398. doi:10.1177/0146621612445904

Educational
Administration Quarterly

On the differential and shared effects of leadership for learning on teachers' organizational commitment and job satisfaction: A multi-level perspective

Journal:	*Educational Administration Quarterly*
Manuscript ID	EAQ-18-0036.R1
Manuscript Type:	Original Manuscript
Keywords:	context, job satisfaction, organizational commitment, leadership for learning, Multilevel Modeling

Abstract:

Purpose
Over the past years "Leadership for Learning" (LFL) has become popular among educational scholars. LFL refers to the idea that effective leaders demonstrate a contextually contingent mix of instructional, transformational, and shared leadership practices which may have differential effects at various organizational levels. These assumptions have rarely been investigated within a coherent empirical design. We examine the shared and differential effects of LFL on teachers' job satisfaction and organizational commitment, which are relevant antecedents for learning, improvement and change on all levels of a school.
Methods
Drawing on survey data (nteachers=3,746, nschools=126) from Germany and on well-established instruments like the MLQ or TALIS, multi-level associations of leadership for learning and teachers' job satisfaction and organizational commitment were explored. This was done by applying doubly-latent structural equation models.
Findings
Our results indicate that 1) it is statistically necessary to model perceived leadership behavior as a multi-level construct, 2) shared leadership is a strong predictor of individual and shared job satisfaction and organizational commitment of teachers whereas 3) individual consideration only shows significant associations on the individual level 4) that leadership for learning is contextually sensitive.
Implications for Research and Practice
Findings make a strong case for studying LFL within a multi-level framework and also for applying complex study and analytical designs, which should take the complexity of the theoretical assumptions into consideration all the way along from questionnaire design, through the process of data collection up to the point of data analysis.

On the differential and shared effects of leadership for learning on teachers' organizational commitment and job satisfaction: A multi-level perspective

1. **Introduction**

Over the past years the term "Leadership for Learning" (LFL) has become increasingly popular among scholars (OECD, 2016b; Townsend & MacBeath, 2011). LFL refers to the idea that effective leaders demonstrate a contextually contingent mix of "instructional leadership, transformational leadership, and shared leadership" practices (Hallinger, 2011, p. 126). Thus, effective principals exhibit three kinds of leadership: (1) instructional leadership behavior centered on the quality of teaching in classrooms, (2) transformational leadership behavior focused on building up the capacity of those they work with and motivating them towards instilling change, and (3) shared leadership behavior comprised of a range of strategies for involving teachers in school-wide decision-making processes (Day, Gu, & Sammons, 2016). While the influence of principals on students and student achievement is considered to be (mostly) indirect, they have a direct impact on teachers as well as their working conditions (Day & Sammons, 2013; Leithwood & Riehl, 2003; Leithwood, Seashore Louis, Anderson, & Wahlstrom, 2004; Marzano, Waters, & McNulty, 2005; OECD, 2016b; Seashore Louis, Leithwood, Wahlstrom, & Anderson, 2010).

Referring to the antecedents and mediating variables of learning on all levels of a school a large body of research suggests that teachers' commitment and job satisfaction are two key factors for improvement and change (Humphreys, 2010; Nielsen & Daniels, 2012; Wahlstrom & Louis, 2008), particularly of students' learning capacities as well as teachers' well-being, motivation, and teaching practices (Dou, Devos, & Valcke, 2017; Federici & Skaalvik, 2012; Newmann, 2002). Furthermore, commitment and job satisfaction are negatively linked to teacher turnover (Billingsley & Cross, 1992; Ostroff, 1992), which in turn can have negative effects on student achievement and a school's social resources (Hanselman, Grigg, Bruch, &

Gamoran, 2016; Ronfeldt, Loeb, & Wyckoff, 2013), notably in low SES-schools (Kushman, 1992) and schools located in low-income neighborhoods (Linnansaari-Rajalin et al., 2015).

With regard to the connection of leadership and the job attitudes of a schools' personnel, there is ample evidence suggesting that specific types of leadership behavior (i.e. instructional, transformational and shared leadership) are positively associated with teachers' job satisfaction as well as teachers' organizational commitment (Bogler, 2001; Bogler & Somech, 2004). Consequently Dou, Davos and Valcke (2017) argue that principal leadership plays an important role in influencing teachers' job satisfaction and organizational commitment and Ware and Kitsantas (2007) even go so far as to state that teachers' commitment can be seen as a direct reflection of a principal's leadership performance.

Nevertheless, most studies within this area of research conducted so far do not incorporate all three leadership facets – instructional, transformational and shared – into one coherent study design. Thus, it is not fully clear whether the three leadership facets have differential effects on teachers' organizational commitment and job satisfaction. Moreover, a vast majority of studies do not take into account the multilevel structure of the underlying measurement design (Batistic, Cerne, & Vogel, 2017). This is problematic because leadership (for learning) is by nature a phenomenon that can have different consequences at various levels, e.g. between a principal and an individual teacher, and/or group of teachers and/or the entire school (Batistic et al., 2017; Dionne et al., 2014). Consistent with this line of thinking, a strong case can also be made to use the LFL model for viewing principal-teacher relations as embedded in the organizational and the social context as well as the culture of a school (Hallinger, 2011, 2016). Thus, modeling LFL only on one particular measurement level (i.e. within or across schools) is not recommended and may even lead to false conclusions (Boyce & Bowers, 2016). The goal of the current study is to fill this gap and take a closer look at the differential and shared effects of LFL on teachers' organizational commitment and job satisfaction within

and across schools. To the best of our knowledge, this is the first study in Germany that analyzes effects of LFL on those job attitudes using a multi-level design.

The article is organized as follows. First, the concepts of (organizational) commitment and job satisfaction as well as their importance for the success of an organization are discussed, drawing first on literature from the non-educational sector, then on literature from the educational realm. Next, the notion of LFL is briefly introduced. Next, drawing on survey data ($n_{teachers}$=3,746, $n_{schools}$=126) from Germany, multi-level associations of leadership for learning and teachers' job satisfaction and organizational commitment are explored. This is done by applying doubly-latent structural equation models. The results are discussed in the context of current research as well as with regard to their implications for practice and future research.

2. The Relations of Leadership for Learning, Job Satisfaction and Organizational Commitment

Defining Organizational Commitment and Job Satisfaction

According to Locke (1976), job satisfaction can be defined as "a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" (p. 1304). Similar definitions have been proposed in recent years (e.g. Robbins & Judge, 2017, p. 116). Locke's definition implies that job satisfaction is not a purely cognitive matter. In line with Hulin and Judge (2003, pp. 255–256), we conceive job satisfaction as multidimensional psychological responses to one's job that have cognitive (evaluative), affective (or emotional), and behavioral components.

Organizational commitment can be understood as the "relative strength of an individual's identification with and involvement in a particular organization" (Mowday, Steers, & Porter, 1978, p. 4), characterized by a strong belief and acceptance of the organization's goals and values, a willingness to exert considerable effort on behalf of the organization and a strong desire to maintain membership in the organization. While there is research suggesting that

commitment is threefold, this understanding has also been criticized (Solinger, van Olffen, & Roe, 2008).

In sum, job satisfaction and organizational commitment are considered two of the central job attitudes and – taken together – form a well-researched classic domain of organizational psychology (Judge & Kammeyer-Mueller, 2012, p. 342). Ample evidence suggests that both are linked (Mathieu & Zajac, 1990), though the exact nature of the relationship remains vague (Martin & Bennett, 1996).

Leadership as Antecedent of Organizational Commitment and Job Satisfaction

In general organizational research (not focused on schools), personal, job and organizational characteristics as well role states and group-leader relations have been considered antecedents to job satisfaction (Wagner III, Leana, Locke, & Schweiger, 1997) and organizational commitment of employees (Mathieu & Zajac, 1990, p. 175; for a more restricted model focusing on personal characteristics, see Meyer, Stanley, Herscovitch, & Topolnytsky, 2002).

Regarding correlates, various aspects of leadership have been found to have a significant link to organizational commitment, chiefly leaders initiating structure and consideration, leader communication as well as participatory or shared leadership, although they appear to be contingent on other factors in the immediate context (Mathieu & Zajac, 1990, p. 180). Results of a meta-analysis show a strong positive correlation between transformational leadership and organizational commitment (Meyer et al., 2002). With regard to job satisfaction, previous research points to a strong link between job satisfaction and participation at work. In a sub study, Wagner et al. (1997) conducted a meta-analysis of 69 participation-correlations (from various studies), arriving at a weighted mean correlation of .3 as well as evidence of substantial unexplained variance (Wagner III et al., 1997, p. 57). And, finally, longitudinal research provides some evidence, that leadership can leverage organizational commitment as well as job satisfaction of employees. For example, Bateman and Strasser (1984) found a

positive change in the commitment and job satisfaction of employees over time when exceptional job performance was recognized and rewarded by their leaders. And Epitorpaki and Martin (2005) demonstrated that the change of an employee's organizational commitment and job satisfaction is an indirect result of perceived leadership behavior filtered through a dyadic leader–employee interaction.

The Leadership for Learning Model as Framework for Educational Research

According to Bush and Glover (2014), the most dominant conceptions of leadership in education are instructional, transformational and distributed/ shared leadership with educational leadership being contingent to the setting. Thus, there is no single best type of leadership, but instead, different leadership skills can be an appropriate way to effectively handle certain issues or situations.

Taking these aspects into consideration, the LFL model aims to connect school leadership to student achievement and the learning organization with the ultimate goal of providing an integrative analytical framework for educational leadership research as well as making available a clearly represented organizing device that is useful for practitioners (Swaffield & MacBeath, 2009). Unlike models focusing on specific leadership styles, the framework emphasizes the relationship between school leadership, collaborative leadership practices/participation, context and learning at various organizational levels. The framework further suggests that principals become effective (mostly) indirectly and that leadership behavior as well as its connections to learning and its antecedents are shaped by a schools' context and culture (Hallinger, 2011).

The LFL approach is closely related to idea of instructional leadership, but goes beyond its limitations, namely principal- and teaching-centering, by incorporating a broader range of leadership activities to support learning and learning outcomes (Bush & Glover, 2014, p. 556). Consequently, the LFL model subsumes features of instructional leadership,

transformational leadership, and shared leadership (Hallinger, 2011, p. 126) practices and it is assumed that effective leadership is contextually contingent (Hallinger & Heck, 2010).

Leadership Styles: Instructional, Transformational and Shared Leadership

With respect to the specific facets of the LFL model, there are findings from various correlational studies suggesting that all of the facets of the LFL model might be connected with teacher' job satisfaction and organizational commitment: Looking at instructional leadership, findings from various studies from the international realm suggest a link between the instructional leadership of principals and the satisfaction and commitment of teachers (Al-Mahdy, Emam, & Hallinger, 2018; Krug, 1992; Sheppard, 1996; for a similar study where such a link could not be determined, see Kouali, 2017). Similarly, findings from various studies point to a connection between transformational leadership and the organizational commitment and job satisfaction of teachers (Bogler, 2001; Leithwood & Jantzi, 2005). Finally, there are also indications that a cooperative leadership team as well as the leadership support for sharing and distributing tasks among between principals and teachers positively predict the organizational commitment of teachers (Hulpia, Devos, & Keer, 2009). Thus, there are empirical findings for all three facets of LFL that point to a connection of leadership in education and the job attitudes of teachers.

Culture and Context

Advocates of the LFL approach argue "that leadership is enacted within an organizational and environmental context" (Hallinger, 2011, p. 127), with context referring to features of the broader organizational and environmental setting within which the school and the principal are located (Hallinger, 2016).

Prior research indicates that school contextual and compositional factors may have effects on all three leadership styles incorporated into the LFL model (Hallinger & Murphy, 1986; Liu, Bellibas, & Printy, 2016; Smith & Bell, 2011). It underscores that a school's context can

influence the way the school is led and/or its priorities. In other words, the national as well as the social and organizational context of a school can be considered important moderators of school leadership and its consequences.

An important contextual aspect to consider is the socio-economic status (SES) of a school. Previous findings show that a school's SES can have a major impact on how that school "functions", on its priorities, working conditions and on strains of school leaders and staff (Brauckmann & Böse, 2017; Levine & Lezotte, 1990; Sirin, 2005). For instance, Hallinger and Murphy (1986), in their groundbreaking study on the topic, found that the SES of a school affected how principals perceived their work and acted in everyday working. The main result from this study was that the principal leadership style in low-SES schools was high in regard to control of instruction and task orientation, whereas it was moderate in non-low-SES schools. Similarly, a study employing mainly end-of-day logs showed that principals tended to work in a more focused manner in schools in challenging circumstances (Goldring, Huff, May, & Camburn, 2008, p. 349). Focus tended to either be put on instruction or student affairs; eclectic principals (without a clear focus) tended to be working in schools that were not challenging circumstances.

German Research On Educational Leadership Styles, Effects And Context

German research analyzing possible links between educational leadership and motivation or commitment as well as research looking into the relationship between educational leadership and context is scarce. Regarding the first field (links between leadership and commitment and motivation), there is some relevant research that underscores the benefits of leadership that incorporates aspects of LFL, i.e. shared or transformational leadership practices. For example, Schaarschmidt and colleagues found that a participatory and supportive leadership style led to more intact interpersonal relationships among staff and acted as a "buffer" for stressors of the day-to-day work (Schaarschmidt & Kieschke, 2013, p. 93). Similarly, a study conducted in

North-Rhine Westphalia found evidence that transformational leadership, participation (in other words sharing of tasks and responsibilities) as well as the work climate in schools correlate highly with the affective commitment of teachers (Harazd, Gieske, & Gerick, 2012). Findings from a mixed-methods study (Gieske, 2013), also conducted in North-Rhine Westphalia, echo this: Data indicated that teaching staff had a stronger organizational commitment in schools with what the author dubbed "rational school principals", who tried to lead by presenting issues in a transparent manner, winning people over through arguments and tried to involve staff in the decision-making process (Gieske, 2013, p. 131ff).

With regard to the relationship between leadership and context, there's hardly any German research that makes use of a quantitative design. However, the existing findings indicate that, similarly to the US context, context matters. For example, Schwarz & Brauckmann (2015) drew upon survey data to show that the area close to school (ACTS) influences among other things school principals' perceptions of student-related challenges at school, workload and what is done during the work time.

Using teacher survey data from the federal state of Hamburg, Author (2016, 2017) investigated the direct and indirect ties between various leadership styles, namely, instructional, transformational, transactional, and laissez-faire leadership, and the instructional practices of teachers by applying a structural equation model. Results revealed that mediating variables are influenced by a leadership core as well as by all leadership facets and that the leadership behavior varied systematically with a schools' achievement context

Leadership as Multi-level Phenomenon

Leadership is a multi-level phenomenon per se (Bliese, Halverson, & Schriesheim, 2002). Nonetheless the vast majority of studies conducted so far focus on one specific level of analysis: either the individual (level 1) or the organizational (level 2) level (Batistic et al., 2017). This is problematic as leadership is an integral part of most fundamental group

processes as well as of individual principal-teacher interactions and, thus, may have differential effects on individual and group-level outcomes (Nielsen & Daniels, 2012).

In this respect teacher-focused leadership behavior (level 1) is targeted at each individual teacher and may lead to the development, change and improvement of his or her individual skills, attitudes, beliefs, performance and outcomes. Group-focused principal leadership behavior (level 2) in contrast is targeted at the whole group of teachers within a single school and is logically similar to all of them (Hofmann, 1997).

From a more methodological point of view the individual level (level 1) reflects the effects of inter-individual differences in perceptions of a principals' leadership behavior, whereas the school level (level 2) represent a climate or a context construct (Marsh et al., 2012). As Marsh et al. (2012) elaborate, for context constructs, based on individual data, the referent is the individual at level 1 and the level 2 construct is an aggregation of the different characteristics within a school. For climate constructs in contrast the referent is a stimulus or construct on the level 2 unit and thus, represent the shared perceptions of group members.

According to Bureau et al. (2017), associations are especially meaningful at level 2 and less so at level 1 for climate constructs, where they represent the effects of differences in the perceptions of the level 2 construct by individual teachers (relative to group averages). For contextual constructs, associations at level 1 are, however, readily interpretable and represent the effects of each individual's unique experience on an individual construct. At level 2, associations involving contextual constructs represent how between-group differences in group aggregates predict group outcomes, above and beyond individual experiences.

3. Design & Method

Purpose

This study aims to examine the differential effects of LFL on teachers' organizational commitment and job satisfaction using a multi-level modeling approach. In particular, the study addresses the following research questions:

1. How much variance of the teacher ratings is accounted by school differences? How reliable are the expected school means estimated? To what degree do teachers' evaluations converge in their assessment of specific LFL facets within schools? The answers to these questions resolve around the overarching question, whether or not LFL should be treated as a multi-level construct.
2. Does the factorial structure of the leadership measures differ within and across schools? This question refers to the factorial validity of the leadership facets.
3. What facets of LFL predict teachers' job satisfaction and teachers' organizational commitment at the individual level (i.e. within schools) and at the school level (i.e. across schools)? This question refers to the differential and shared effects of LFL on the teachers' job satisfaction and commitment.
4. To what degree does the particular social environment or the context of a school affect the associations between leadership behavior and teachers' job satisfaction and organizational commitment at the school level? This question refers to the impact of the school context on the relationship between LFL and the outcome variables.

Sample

This study is a secondary analysis of teacher survey data that has been collected by the Hamburg school inspection between 2012 and 2015 within its second inspection cycle (2012-2019) in the federal state of Hamburg, Germany.

In Germany, the responsibility for the education system, and hence its detailed organization, lies primarily with the federal states (see Author, 2015). While the German education system has many unique features compared to other European countries (for a full presentation of the German education system, see Döbert, 2007), the leadership practices and their possible effects are of universal nature and can also be found in schools all over the world (see for example OECD, 2014, which among other things also indicates that there are aspects of educational leadership that can be found internationally). One major difference with regard to

most other countries, however, is, that in Germany principals only have little autonomy over teacher recruitment and appointment as well as over teacher salaries and teacher promotion, as principals hold less than 20 percent of the responsibility for resources (the OECD average is 38 percent, OECD, 2016a).

The Hamburg school inspection collects data at every school in the federal state every four to six years for assessing a school's quality in terms of school and teacher effectiveness. Schools are annually quota-sampled (reflecting school type and social composition) at random from all schools within the system that have not been inspected during the second inspection cycle up to that point in time. This can be understood as a successive sampling with continuously increasing sample size, meaning that at the end of a cycle all 100% of schools within the federal state will have been inspected. The inspection relies on a multimethod investigation including classroom observations, structured interviews as well as student, teacher and parent surveys. The teacher survey is conducted as a full population survey on the level of each school, thus, no sampling takes place within a school and all teachers of a school are requested to participate in the online survey.

Verified and validated inspection data for the second inspection cycle was available for the period from 2012 to 2015. During that period the inspection gathered data of n=3,746 teachers within n=126 schools. The school sample consists of 74 primary schools, 31 upper secondary schools, ("Gymnasium" in German) and 21 comprehensive schools ("Stadtteilschule" in German). The available school sample represents a proportion of 40 percent of all schools within the Hamburg school system. The teaching staff within a single school ranged in size from 7 to 85 (*mean*=30). The response rate of the teacher online survey was 65.2 percent (N=5,745). The percentage of missing values in the data set was 16.2, with missing values at the item-level ranging between seven and 34 percent.

Measures and Instruments

The teacher questionnaire comprises a total of 114 items that is divided into seven sections: a) teacher collaboration, b) (perceived) leadership behavior, c) teachers' instructional practices, d) inner-school conditions for school and instructional improvement, e) teachers' innovation capacity, f) job satisfaction of teachers and g) teachers' organizational commitment. Detailed in-depth information concerning the measures and instruments can be found in the methodological documentation of the Hamburg school inspection (e.g. Author, 2013).

Transformational leadership ($\omega = .94$)[1] was measured using three subscales of the Multifactor Leadership Questionnaire (MLQ, Bass & Avolio, 1995). First, teachers answered three items measuring Individualized Influence attributed (IIa, $\omega = .89$), a leadership behavior in which charisma is used to foster strong positive emotional bonds with teachers (e.g. "The principal acts in ways that foster my respect."). Second, they answered three items indicating Individual Consideration (IC, $\omega = .90$), which occurs when a principal tries to understand the intrinsic needs and to recognize the abilities of each individual teacher with the goal of developing and empowering each of them individually (e.g. "The principal considers me as having different needs, abilities, and aspirations from others."). And third, teachers responded to another three items capturing the facet Inspirational Motivation (IM, $\omega = .80$), that aims to influence follow teachers through providing meaning, optimism and enthusiasm for a vision/ a set of goals that then becomes viewed as universally valuable to achieve (e.g. "The principal articulates a compelling vision of the future.").

Instructional leadership ($\omega = .91$) respectively its facets in turn are measured by applying three scales form the Teaching and Learning International Survey (TALIS, OECD, 2009), that are derivatives of Hallinger's (1994) Principal Instructional Management Rating Scale (PIMRS). Thus, the first facet, called management for school goals ($\omega = .79$), is similar to the

[1] For all scales McDonald Omega coefficient was computed as a measure of the overall scale reliability, as the Cronbach's Alpha coefficients is just the lower bound of the reliability in case of congeneric variables (Dunn, Baguley, & Brunsden, 2014).

first dimension of PIMRS in that it "highlights principals' explicit management of instruction through school goals" (Lee, Walker, & Chui, 2012, p. 589, e.g. "The principal ensures that teachers work according to the school's educational goals."). The other both facets are overlapping with the second dimension of the PIMRS – called: Manages the Instructional Program – and partitions this dimension by proposing instructional management and direct supervision of instruction (Lee et al., 2012). Thus, the items of the second facet, which is called instructional management ($\omega = .78$), focus on principals' actions to improve teachers' instruction by encountering and solving instructional challenges and problems (e.g. "When a teacher has problems in his/her classroom, the principal takes the initiative to discuss the matter."). And the third facet, which is called direct supervision of instruction in the school ($\omega = .76$), refers to "actions to directly supervise teachers' instruction and learning outcomes" (OECD 2009, p. 194, e.g. "The principal or someone else in the management team observes teaching in classes.").

Shared leadership ($\omega = .81$), finally, is measured by a three item scale which is a German translation by the Hamburg school inspection of items from a scale capturing *Shared Leadership Among Principal and Others* that has been developed by Wahlstrom and Louis (2008). Thus, the construct follows the idea that principals use less of their controlling power and are more willing to share their positional power with others through a participatory form of decision-making, eventually resulting in the empowerment of teachers. Thus, the surveyed teachers rated items indicating their influence over and participation in schoolwide decisions (e.g. "Teachers have an effective role in school-wide decision making.").

All of the 20 leadership items used within this study were rated on a four point rating scale (Transformational and Instructional Leadership: 1=(almost) never, 4=(almost) always; Shared Leadership: 1=(strongly agree), 4=(strongly disagree)).

As the referent for all leadership scales is on the school level (level 2), the school level aggregate represents leadership culture.

Organizational Commitment ($\omega = .92$) of teachers is captured by four items drawn from a derivative of the Organizational Commitment Questionnaire adapted for schools (OCQ, Lipowsky, Faust, & Greb, 2009). Thus, the teachers rated items reflecting their belief in and the acceptance of school goals and values, the willingness to exert effort towards school goal accomplishment, and their desire to maintain school membership (Mowday et al., 1978). In this respect, the scale *grosso modo* assess the identification of teachers with their school as well as their willingness to work towards and to accept school goals (e.g. "I am extremely glad that I chose this school to work for over others I was considering at the time I joined.").

Job Satisfaction ($\omega = .88$) of teachers is measured by a five-item scale developed by the Hamburg school inspection. The construct is based upon the assumption that job satisfaction is an internal state that is "expressed by affectively and/or cognitively evaluating an experienced job with some degree of favor or disfavor" (Brief, 1998, p. 86). Referring to this the scale captures and combines the satisfaction of teachers in five dimensions within a single measure: their satisfaction with a) work in general, b) promotion/ development opportunities, c) co-workers/communication opportunities, d) decision-making/ participating opportunities and e) work climate (e.g. "Overall I am satisfied with the working conditions at my school."). The items of both scale were rated on a four-point rating scale (1= I do not agree at all., 4=I strongly agree.).

As the referent for both scales is located on the individual level (level 1) the school level aggregates represent shared attitudes of teachers and should be considered context variables.

Further context-control variables, that potentially confound or moderate the association of leadership behavior and the organizational commitment as well as the job satisfaction of teachers include:

Social index (Author, 2014) of a Hamburg school, that is based on data collected from parents and students within each school every five years as well as on information annually reported by the Statistical Office for Hamburg and Schleswig-Holstein, and that allows to differentiate

among schools drawing their students from highly socio-economically disadvantaged areas and schools with a more privileged context. The index is unidimensional and covers facets like the socio-economic and the socio-cultural background of the students' families as well as their migration background and furthermore information on social indicators of the neighborhood at their families' places of residence (for Details see Appendix A). Where ever possible international standardized and comparable sub-scales were calculated. Thus, the social index contains information on the parental stage of qualification, as defined within the International Standard Classification of Education (ISCED, UNESCO Institute for Statistics, 2012) as well as information on the occupational classes of parents as defined within the Erikson–Goldthorpe–Portocarero (EGP, Erikson & Goldthorpe, 1992) scheme. To help policy makers interpret what an index score means, the scale of the Hamburg social index is divided into six levels. Levels one and two are defined as extremely challenging conditions for teaching and schooling, indicating that a school has a high proportion of students from disadvantaged backgrounds. Thus, schools with an index of one or two receive a higher level of language promotion/training and are limited to maximum class size of 21 students. Within the sample n=33 schools (26%) enroll the majority of their students from such socially disadvantaged backgrounds. For the analyses, the variable was dummy-coded (0= school in challenging circumstances, 1=school in no particular circumstances).

School type is indicated by the three types of schools in our sample: primary schools (n=75), upper secondary schools, "Gymnasium" in German (n=34) and comprehensive schools, "Stadtteilschule" in German (n=21). It is important to note that this stratification includes the level of schools, that may have its own influence on inner-school variables (Seashore Louis & Lee, 2016), as in Germany the educational level and the type of school are partially confounded (primary schools = elementary; comprehensive schools and upper secondary = secondary schools). The variable was dummy-coded.

Analytic Strategy

First, we computed the intra-class correlations in terms of ICC(1) (i.e., the proportion of between-group variance to the total variance) and ICC(2) (i.e., the reliability of a schools mean) as well as the $r_{wg}(j)$-index (i.e. a measure of the rater agreement among teachers within schools; see James, Demaree, & Wolf, 1984) for all leadership measures. The calculation was done using the r package `multilevel` by Bliese (2016).

Following the recommendations by Bliese (2000), ICC(1)-values greater than .05 and ICC(2)-values greater than .70 indicate that the given construct is a group-level construct, and thus, a multi-level modeling approach should be used (see also Lüdtke et al., 2008). According to Brown and Hauenstein (2005), $r_{wg}(j)$-values ranging between 0 to .59 suggest an unacceptable level of rater agreement, values between .60 and .69 indicate a weak level of rater agreement, values between .70 to .79 indicate a moderate level of rater agreement, and values equal or greater than .80 indicate as a strong level of rater agreement.

Next, multi-level confirmatory factor analyses (MCFA) were estimated to ensure that the proposed leadership measurement model on each level fits the data well and to check if the constructs are the same and equally related to each other on the individual as well as on the school level.

To assess the fit of the models the classic fit indices comparative fit Index (CFI), root mean square error of approximation (RMSEA) and standardizes root mean square residual (SRMR) as provided by MPLUS 7 were used. Acceptable fit would be indicated by a CFI over .90, a RMSEA less than .08 and a $SRMR_w$ / $SRMR_b$ less than .08 (e.g. Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). In addition, we compared different models using the Akaike Information Criterion (AIC), and Bayes Information Criterion (BIC), preferring models with lower values (Bollen, Harden, Ray, & Zavisca, 2014).

After identifying the best fitting model, we related the different leadership facets to the outcome variables (i.e. teachers' commitment and job satisfaction) using a multi-level (or

doubly latent) structural equation approach (MSEM, see Marsh et al., 2009). The advantage of this modeling approach is that it allows researchers to investigate the differential and shared effects at the individual (within) and the aggregated (between) level, while accounting for measurement and sampling error influences.

Fourth, context variables were entered at the school level to examine the stability and sensitivity of the estimates of the relationship between the leadership facets and the outcomes variables after controlling for the particular context of a school. Missing data was handled using full information maximum likelihood estimation implemented in M*plus* 7 (Muthén & Muthén, 2012).

For all multi-level models we report the standardized latent regression coefficients together with their corresponding standard errors. The standardized latent regression coefficients can be interpreted as level-specific effect sizes. That is, they show the increase in the dependent variable (i.e. job satisfaction and organizational commitment) in standard deviations if the independent variable is increased by one standard deviation on a particular measurement level (holding all other independent variables constant). Further, we calculated and report effect sizes (*ES*) from formulas provided by Tymms (2004). Those effect sizes are based on the residual variance of the dependent variables (i.e. job satisfaction and organizational commitment) at level 1 (e.g. the teacher level) and are equivalent to Cohen's d (Cohen, 1988), where an effect size of 0.20–0.30 is taken to be a small effect, 0.50 a medium effect and greater than 0.80 a large effect.

Limitations

The current study is limited in the following aspects. First, this study comprised only three out of five dimensions of transformational leadership as described within the MLQ. Hence, the two dimensions Intellectual Stimulation (IS) and Idealized Influence behavior (IIb) were not included in our analyses. Further the LFL framework makes a case that the exercise of

leadership is moderated by personal characteristics of principals, an assumption we were not able to test within our study as no data on this topic was available.

Second, the individual leadership facets were highly correlated in this study (see Appendix B). This was particularly the case at the school level, which may have caused higher standard errors in the estimation of the latent regression parameters on level 2. However, in order to examine the differential and shared effects, it was necessary to include all leadership facets in our analysis.

Third, potential confounders were only tested by including them as control variables in our MSE-model. Therefore, it was impossible to investigate the direction of the association within different school settings using a sophisticated multi-group MSEM modeling approach. Future studies may examine the differential effects of leadership facets on different measurement levels and different groups.

4. Results

Research Question 1: Should LFL be treated as a multi-level construct?

Table 1 summarizes the intra-class coefficients ICC(1) and ICC(2) as well as the $r_{wg}(j)$ for each leadership facet. The results show that all leadership facets can be considered as multi-level constructs, as the ICC(1)- and ICC(2)-values consistently exceed the above mentioned cutoff criteria (i.e. ICC(1)<.05 and ICC(2)<.70). More specifically, a large amount of variance of the leadership measures was attributable to school differences (ICC(1): 22-37%). The ICC(2) values ranged between .89 (Individual Consideration) and .95 (Inspirational Motivation). The $r_{wg}(j)$-indices suggested a moderate to strong level of agreement among teacher ratings within schools (ranging from .78 for Shared Leadership to .89 for Transformational Leadership and from .67 for Individual Consideration to .81 for Inspirational Motivation). Overall these results indicate a substantial amount of variation of leadership behavior across schools and thus support the fact that leadership behavior should be treated as a multi-level (within-and-across-school) phenomenon, whereby the facet

Individual Consideration showed the lowest $r_{wg}(j)$-index and, thus, should be viewed with some caution as level 2 construct.

Table 1: Estimates of Variance, Reliability and Agreement

Research Question 2: Does the factorial structure of the leadership measures differ within and across schools?

Next, we scrutinized measurement models of the leadership facets at both levels by applying multi-level confirmatory factor analyses (MCFA). We fitted the following measurement models to the data:

- 1-1-factor model: This model assumes one (unidimensional) general latent factor at the within and the between level. This model is a very restrictive model as it implies that all leadership measures are affected by one global leadership factor (i.e. g-factor) within and across schools.
- 3-1-factor model: This model takes into account the three main leadership dimensions used within this study at the individual level. Thus, it assumes that a principal can act transformational as well as instructional and can also share his respectively her leadership power with others, whereby all leadership dimensions are correlated. On the school level a global leadership factor is implied.
- 3-3-factor-model: This model assumes that a principal can act transformational as well as instructional and can also share his respectively her leadership power with others on the individual as well as on the school level, whereby all leadership dimensions are correlated.
- 7-1-factor-model: This model implies, that, within the three dimensions, transformational, instructional and shared leadership, it could be reliably distinguished between the specific facets respectively sub-dimensions (as far as existing within the specific framework). Hence, it is assumed that a schools' principal could make use of Idealized Influence attributed, Individual Consideration, and Inspirational Motivation

as well as of School Management, Instructional Management, and Supervision and could beyond that share decision power with teachers. On the school level a global leadership factor is implied.

- 7-3-factor model: This model assumes seven latent factors (facets) on the individual level and only three dimensions on the school level, whereby correlations between the facets respectively dimensions were allowed.
- 7-7-factor model: The final and least restrictive model assumed latent factors for each leadership sub-dimension (facet) at the individual and the school level. The correlations between the latent factors at each level were freely estimated.

Insert Table 2

Table 2 summarizes the fit statistics of the above-mentioned measurement models. The results of the multi-level confirmatory factor analyses (MCFA) show, that the measurement model considering one factor on the individual level as well as both measurement models with three factors on the individual level did not produce an adequate fit. According to the fit statistics [$\chi^2 = 1972.753$(df =2305), p<0.001, CFI = 0.95, $RMSEA$ = 0.04, $SRMR_w$ = 0.03, and $SRMR_b$ = 0.07] the 7-7-factor model was chosen as best fitting model. Thus, the factorial structure of the leadership measurement model does not differ across and between schools.

Research Question 3: What facets of LFL predict teachers' job satisfaction and teachers' organizational commitment at the individual level (i.e. within schools) and at the school level (i.e. across schools)?

To answer this research question, we estimated a doubly latent structural equation model (Marsh et al., 2009) based on the 7-7-factor measurement model which is illustrated in Figure 1. The correlations among the leadership facets at both levels as well as reliability coefficients of the different facets are provided in Appendix B. ICC(1), ICC(2) and r_{wg}(j) were .22, .89

and .80 for organizational commitment and .29, .92 and .75 for job satisfaction, indicating that both constructs are multi-level in nature too and represent shared attitudes of teachers at the school-level.

Insert Figure 1

The results of these analyses are given in Table 3a and b (see Model 1). The results indicate that shared leadership significantly predicts teachers' organizational commitment well as teachers' job satisfaction on the individual ($\beta_{Organizational\ Commitment,L1} = .374$, $\beta_{Job\ Satisfaction,L1} = .445$) and on the school-level ($\beta_{Organizational\ Commitment,L2} = .728$, $\beta_{Job\ Satisfaction,L2} = .721$). According to Cohen (1988), the effect size (*ES*) of these effects can be considered as large, ranging between *ES* = 0.84 (commitment) and *ES*= 0.95 (job satisfaction) at the individual, and between *ES* = 0.95 (commitment) and *ES*=1.00 (job satisfaction) at the school-level. On the individual level, the transformational leadership facets Idealized Influence attributed ($\beta_{Organizational\ Commitment,L1} = .205$, ES=0.44) and Individual Consideration ($\beta_{Organizational\ Commitment,L1} = .183$, ES=0.39) have also a significant positive effect on teachers' organizational commitment. With regard to teachers' job satisfaction, the leadership facet Individual Consideration was found to be a further statistically significant predictor at the individual level ($\beta_{Job\ Satisfaction,L1} = .399$, ES= .78). Controlling for all proposed LFL facets, the remaining facets were not statistically significant (even though we find some high effect sizes on both levels of analysis). The shared effect of the leadership facets was examined using the coefficient of multiple determination (R^2). The percentage of explained variance ranged between 54 and 62% on the individual level and between 67 and 78% on the school level. These results suggest that the leadership facets taken together explain a rather large amount of variance in the outcome variables. Thus, the shared effect of LFL on teachers' organizational commitment and job satisfaction is large.

Insert Table 3a

Research Question 4: To what degree does the particular social environment or the context of a school affect the associations between leadership behavior and teachers' job satisfaction and organizational commitment at the school level?

To answer the fourth and last research question, we entered the schools' context variables in two steps. First, the social indices of the educational institutions (see Table 3a and b, Model 2) were entered as control variables into the model. The results indicate that the social background of a schools' student population has a statistically significant impact on teachers' organizational commitment ($\beta_{Organizational\ Commitment,L2} = .111$) and job satisfaction ($\beta_{Job\ Satisfaction,L2} = .094$) at the school level. Thus, teachers who work in schools with a higher amount of socially privileged students are more strongly committed to their schools and more satisfied with their jobs than their colleagues who work at schools in challenging social circumstances. It is worth noting that the regression coefficients of the leadership facets did not differ after entering the control variables into the model.

Insert Table 3b

In a second step, we entered the school type and school size as control variables into the MSEM (Tables 3a and b, Model 3). After including those additional potential confounders into the model, the regression coefficient of the Instructional Management facet at the school level ($\beta_{\text{Organizational Commitment,L2}} = .350$, ES: 0.41, $\beta_{\text{Job Satisfaction,L2}} = .293$, ES: 0.35) became statistically significant (one-tailed: $p<0.05$, two-tailed: $p<.10$) for both dependent variables. These results indicate that the association of an instructional leadership culture and the shared organizational commitment and shared job satisfaction of teachers varies with the social and structural context of a school in its entirety.

5. Discussion

The goal of the present study was to examine the associations of LFL and the job attitudes of teachers, namely job satisfaction and organizational commitment, in a German context from a multi-level perspective. Based on the current study, the need to model leadership within the LFL framework as multi-level construct became clear, as a relevant amount of variance in the perceived leadership behavior is located between schools. The findings also indicate that the commitment and job satisfaction of teachers are virtually a direct reflection of a principals' type of leadership, particularly when leadership is group-focused. These results support findings of earlier studies, that argue that principal leadership is extraordinary relevant for the job attitudes of teachers (Hulpia, Devos, & Rosseel, 2009b; Ware & Kitsantas, 2007) and that leadership is a multilevel construct per se (Yammarino, Dionne, Chung, & Dansereau, 2005).

Our results suggest that mainly shared perceptions of shared leadership predicted teachers' commitment and teachers' job satisfaction at the school level, whereas instructional as well as transformational leadership behavior was not significantly related to teachers' shared organizational commitment and teachers' job satisfaction. Thus, a schools' leadership climate, focusing on sharing controlling power through a participatory form of decision-making, demonstrably shapes the context of the individual job attitudes of teachers within a school. Furthermore, beyond group-level effects of shared leadership, individual teachers perceiving their principal to demonstrate a higher level of shared leadership behavior relative to the school average reported higher levels of job satisfaction and organizational commitment. The shared level 1-effects of LFL on the individual organizational commitment (R^2=0.542 converted into Cohen's d=2.16) and job satisfaction (R^2=0.616 converted into Cohen's d=2.52) of teachers we found in our study are much higher than the average effects reported in the meta-analyses by Wagner et al. (1997, r=0.30 converted into Cohen's d=0.63) and Meyer et al. (2002, r between 0.27 and 0.46 converted into Cohen's d 0.56 to 1.04) for single dimensions of LFL, i.e. transformational leadership. Notwithstanding, the effect of shared

leadership behavior on the individual job attitudes of teachers in our study (Cohen's d between 0.84 and 0.95) is (roughly) in the same range as the effect reported in the meta-analysis by Mathieu and Zajac (1990, r=0.39 converted into Cohen's d=0.85). What our study demonstrates further, is that these effects and effect sizes are not only observable on the individual but also on the organizational, the school level. Those findings shed a new light on LFL research. While former studies and analyses neither took all facets nor all relevant organizational levels of leadership into account, our study indicates that particularly shared leadership behavior has a strong positive effect on those job attitudes when controlled for other aspects of leadership – and that on the individual as well as on the school level and that instructional leadership do not play a vital role with regard to the job attitudes of teachers investigated in our study.

Concerning the different levels of analysis, it should be noted that Individual Consideration, a transformational leadership facet that, according to our data, statistically rather seems to be a level 1 than a level 2 phenomenon, is positively related to both job attitudes of teachers on the individual teacher level, but not on the school level. This may be due to the formulation of the items, that, despite the fact that the referent is at level 2, were operationalized and intended to measure Individual Consideration at the individual level (Avolio & Bass, 1995). In this case the differentiation between contextual and cultural phenomena, as proposed by Marsh et al. (2012), may be too strict and too little flexible and one may have to ask what does this construct represent at level 2: context or culture?

Further, there is current research that suggests that teachers generally tend to perceive instructional leadership to be a single dimension (Urick & Bowers, 2017), a result that our analyses do not echo. Neither on the individual nor on the school-level the amalgamation of factors was statistically justifiable. As the same scales from TALIS were used within our study as within the study by Urick and Bowers (2017), a possible explanation could be that

cultural differences could affect the factorial structure of instructional leadership and that its measurement may be contextually sensitive.

In addition, we were able to demonstrate that the perceived leadership behavior of principals differs by social and structural context as the confounder analysis revealed. On the one hand there seems to be a set of core leadership practices that are related to the job attitudes of teachers, independently from the context of a school; Individual Consideration and Shared Leadership on the individual level and Shared Leadership on the school level. But with regard to the structural and social context of a school, the study also shows that instructional management and its relation to the shared job satisfaction and shared organizational commitment of teachers seems to be contextually contingent. This finding is again in line with the (relatively sparse) empirical research base (Hallinger, 2016), even if our analyses do not reveal the direction of the potential effect.

7. Conclusions

Two conclusions we draw from this study are that LFL should be considered within a multi-level framework and that there is a need for more complex study and analytical designs, that should take the complexity of the theoretical assumptions into consideration all the way along from questionnaire design, through the process of data collection up to the point of data analysis.

The findings further suggest a critical examination of current leadership models, scales and methodological assumptions used within our study, as our results indicate that those teachers who feel that their principals better understand their intrinsic needs and better recognize their abilities with the goal of developing and empowering them individually (Individual Consideration) in contrast to the school mean also experience a higher job satisfaction and are more strongly committed to their school than their colleagues.

Further, the effect sizes are extremely high (about 1.00) for shared leadership behavior on the individual as well as on the school level, indicating that sharing decision-power is remarkably relevant for the individual job satisfaction and organizational commitment of teachers as well as for the shared attitudes. This could indicate that a notion of shared leadership is part of the organizational culture and not dependent on a school leader. While German research dealing with cooperation and sharing of tasks among teachers is scarce (Bauer, 2004), there are indications that themes of collegiality and collaboration are in high esteem among German teachers (Baum, Idel, & Ullrich, 2012). This warrants some further research. Also futurestudies should strive to create models that not only take into account the individual or shared/collective levels of job satisfaction, commitment and leadership, but simultaneously analyze both levels to avoid potential fallacies.

We advocate that these findings should not be taken as a claim that one style of leadership is universally more important or valuable than another. We adopt a contingent view of leadership, depending on the unique cultural makeup of each school as well as the individuals that work in it and various contextual factors. As others have pointed out, there "is no single leadership formula for achieving success" (Day et al., 2016, p. 253). What we can say based on our data, however, is that a shared leadership style is beneficial in many circumstances, on the individual as well as the school level. It therefore constitutes a promising avenue for any school improvement and educational policy development efforts.

As instructional management systematically varies between contexts', it is likely that this LFL facet may be a relevant key for changing and/ or improving schools in which a staff has to work under difficult conditions. This implies that instructional leadership does not seem to produce the (virtually universal) positive impact in Germany it seems to have in other countries. This could be due to school principals in Germany having a different role and different tasks than principals from the UK, the US or other western countries (Author, 2015). Going beyond this aspect, this study makes evident the need for additional research to find out

more about the contextual contingency of this aspect of school leadership – particularly in varying national and/or cultural contexts but also with regard to measurement issues. Regarding implications for practice and policy making, augmenting the time spent on learning about LFL and the culture of shared tasks appears to be a promising avenue of improvement for professional teacher development as well as for the professional development of (future) school principals. By the same token, it seems more crucial than ever for teachers, principals and those who train them to understand leadership as a holistic concept and to keep in mind that instructional leadership seems not to be the promised silver bullet (at least not in all countries), especially when it is viewed as universally effective. Focusing on the LFL approach, that overcomes the limitations of instructional leadership by incorporating a wider spectrum of leadership actions to support learning and learning outcomes on all levels of a school, may be a good starting point for this.

References

Author (2013). Blinded for review.

Author (2014). Blinded for review.

Author (2015). Blinded for review.

Author (2016). Blinded for review.

Author (2017). Blinded for review.

Al-Mahdy, Y. F. H., Emam, M. M., & Hallinger, P. (2018). Assessing the contribution of principal instructional leadership and collective teacher efficacy to teacher commitment in Oman. *Teaching and Teacher Education*, *69*, 191–201. https://doi.org/10.1016/j.tate.2017.10.007

Avolio, B. J., & Bass, B. M. (1995). Individual consideration viewed at multiple levels of analysis: A multi-level framework for examining the diffusion of transformational leadership. *The Leadership Quarterly*, *6*(2), 199–218.

Bass, B. M., & Avolio, B. J. (1995). *MLQ Multifactor Leadership Questionnaire. Technical Report*. Redwood City: Mind Garden.

Bateman, T. S., & Strasser, S. (1984). A Longitudinal Analysis of the Antecedents of Organizational Commitment. *Academy of Management Journal*, *27*(1), 95–112. https://doi.org/10.5465/255959

Batistic, S., Cerne, M., & Vogel, B. (2017). Just how multi-level is leadership research? A document co-citation analysis 1980–2013 on leadership constructs and outcomes. *The Leadership Quarterly*, *28*(1), 86–103.

Bauer, K.-O. (2004). Lehrerinteraktion und -kooperation. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (pp. 813–831). VS Verlag für Sozialwissenschaften, Wiesbaden. https://doi.org/10.1007/978-3-663-10249-6_33

Baum, E., Idel, T.-S., & Ullrich, H. (Eds.). (2012). *Kollegialität und Kooperation in der Schule*. Wiesbaden: VS Verlag für Sozialwissenschaften. Retrieved from http://link.springer.com/10.1007/978-3-531-94284-1

Billingsley, B. S., & Cross, L. H. (1992). Predictors of Commitment, Job Satisfaction, and Intent to Stay in Teaching: A Comparison of General and Special Educators. *The Journal of Special Education*, *25*(4), 453–471. https://doi.org/10.1177/002246699202500404

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

Bliese, P. D. (2016). *Multilevel Modeling in R (2.6)—A Brief Introduction to R, the multilevel package and the nlme package*.

Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods in leadership: The articles, the model, and the data set. *The Leadership Quarterly*, *13*(1), 3–14.

Bogler, R. (2001). The Influence of Leadership Style on Teacher Job Satisfaction. *Educational Administration Quarterly*, *37*(5), 662–683. https://doi.org/10.1177/00131610121969460

Bogler, R., & Somech, A. (2004). Influence of teacher empowerment on teachers' organizational commitment, professional commitment and organizational citizenship behavior in schools. *Teaching and Teacher Education*, *20*(3), 277–289. https://doi.org/10.1016/j.tate.2004.02.003

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and Alternative Bayesian Information Criteria in the Selection of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 1–19. https://doi.org/10.1080/10705511.2014.856691

Boyce, J., & Bowers, A. J. (2016). Different levels of leadership for learning: investigating differences between teachers individually and collectively using multilevel factor analysis of the 2011-2012 Schools and Staffing Survey. *International Journal of Leadership in Education*, *21*(2), 197–225. https://doi.org/10.1080/13603124.2016.1139187

Brauckmann, S., & Böse, S. (2017). Picking up the pieces? Zur Rolle der Schulleitung beim Turnaround – Ansätze und empirische Erkenntnisse. In V. Manitius & P. Dobbelstein (Eds.), *Schulentwicklungsarbeit in herausfordernden Lagen* (pp. 85–103). Münster: Waxmann Verlag.

Brief, A. P. (1998). *Attitudes In and Around Organizations*. Thousand Oaks: SAGE.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater Agreement Reconsidered: An Alternative to the rwg Indices. *Organizational Research Methods*, *8*(2), 165–184. https://doi.org/10.1177/1094428105275376

Bureau, J. S., Gagné, M., Morin, A. J., & Mageau, G. A. (2017). Transformational Leadership and Incivility: A Multilevel and Longitudinal Test. *Journal of Interpersonal Violence*, 0886260517734219.

Bush, T., & Glover, D. (2014). School leadership models: what do we know? *School Leadership & Management*, 1–19. https://doi.org/10.1080/13632434.2014.928680

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J: Lawrence Erlbaum.

Day, C., Gu, Q., & Sammons, P. (2016). The Impact of Leadership on Student Outcomes How Successful School Leaders Use Transformational and Instructional Strategies to Make a Difference. *Educational Administration Quarterly*, *52*(2), 221–258. https://doi.org/10.1177/0013161X15616863

Day, C., & Sammons, P. (2013). *Successful Leadership: a Review of the International Literature*. Nottingham: CfBT Education Trust.

Dionne, S. D., Gupta, A., Sotak, K. L., Shirreffs, K. A., Serban, A., Hao, C., … Yammarino, F. J. (2014). A 25-year perspective on levels of analysis in leadership research. *The Leadership Quarterly*, *25*(1), 6–35. https://doi.org/10.1016/j.leaqua.2013.11.002

Döbert, H. (2007). Germany. In W. Hörner, H. Döbert, B. von Kopp, & W. Mitter (Eds.), *The Education Systems of Europe* (pp. 299–325). Dordrecht: Springer Netherlands.

Dou, D., Devos, G., & Valcke, M. (2017). The relationships between school autonomy gap, principal leadership, teachers' job satisfaction and organizational commitment. *Educational Management Administration & Leadership*, *45*(6), 959–977. https://doi.org/10.1177/1741143216653975

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Epitropaki, O., & Martin, R. (2005). From ideal to real: a longitudinal study of the role of implicit leadership theories on leader-member exchanges and employee outcomes. *Journal of Applied Psychology*, *90*(4), 659–676. https://doi.org/10.1037/0021-9010.90.4.659

Erikson, R., & Goldthorpe, J. H. (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Oxford University Press.

Federici, R. A., & Skaalvik, E. M. (2012). Principal self-efficacy: relations with burnout, job satisfaction and motivation to quit. *Social Psychology of Education*, *15*(3), 295–320. https://doi.org/10.1007/s11218-012-9183-5

Gieske, M. (2013). *Mikropolitik und schulische Führung: Einflussstrategien von Schulleitern bei der Gestaltung origanisationalen Wandels: Einflussstrategien von Schulleitern bei der Gestaltung origanisationalen Wandels*. Bad Heilbrunn: Klinkhardt, Julius.

Goldring, E., Huff, J., May, H., & Camburn, E. M. (2008). School context and individual characteristics: what influences principal practice? *Journal of Educational Administration*, *46*(3), 332–352. https://doi.org/10.1108/09578230810869275

Hallinger, P. (1994). *A Resource Manual for the Principal Instructional Management Rating Scale (PIRMS Manual 2.2)*. Nashville: Center for the Advanced Study of Educational Leadership.

Hallinger, P. (2011). Leadership for learning: lessons from 40 years of empirical research. *Journal of Educational Administration*, *49*(2), 125–142. https://doi.org/10.1108/09578231111116699

Hallinger, P. (2016). Bringing context out of the shadows of leadership. *Educational Management Administration & Leadership*, *46*(1), 5–24. https://doi.org/10.1177/1741143216670652

Hallinger, P., & Heck, R. H. (2010). Leadership for Learning: Does Collaborative Leadership Make a Difference in School Improvement? *Educational Management Administration & Leadership*, *38*(6), 654–678. https://doi.org/10.1177/1741143210379060

Hallinger, P., & Murphy, J. F. (1986). The Social Context of Effective Schools. *American Journal of Education*, *94*(3), 328–355.

Hanselman, P., Grigg, J., Bruch, S. K., & Gamoran, A. (2016). The Consequences of Principal and Teacher Turnover for School Social Resources. In G. Kao & H. Park (Eds.), *Research in the Sociology of Education* (Vol. 19, pp. 49–89). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-353920150000019004

Harazd, B., Gieske, M., & Gerick, J. (2012). Was fördert affektives Commitment von Lehrkräften? Eine Analyse individueller und schulischer (Bedingungs-)Faktoren. *Zeitschrift für Bildungsforschung*, *2*(2), 151–168. https://doi.org/10.1007/s35834-012-0039-z

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*(6), 723–744.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulin, C. L., & Judge, T. A. (2003). Job Attitudes. In I. B. Weiner, W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of Psychology – Volume 12 – Industrial and Organizational Psychology* (Vol. 12, pp. 255–276). Hoboken, NJ: John Wiley & Sons.

Hulpia, H., Devos, G., & Keer, H. V. (2009). The Influence of Distributed Leadership on Teachers' Organizational Commitment: A Multilevel Approach. *The Journal of Educational Research*, *103*(1), 40–52. https://doi.org/10.1080/00220670903231201

Hulpia, H., Devos, G., & Rosseel, Y. (2009a). Development and Validation of Scores on the Distributed Leadership Inventory. *Educational and Psychological Measurement*, *69*(6), 1013–1034. https://doi.org/10.1177/0013164409344490

Hulpia, H., Devos, G., & Rosseel, Y. (2009b). The relationship between the perception of distributed leadership in secondary schools and teachers' and teacher leaders' job satisfaction and organizational commitment. *School Effectiveness and School Improvement*, *20*(3), 291–317. https://doi.org/10.1080/09243450902909840

Humphreys, E. (2010). *Distributed Leadership and its Impact on Teaching and Learning* (phd). National University of Ireland Maynooth. Retrieved from http://eprints.maynoothuniversity.ie/2041/

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85.

Judge, T. A., & Kammeyer-Mueller, J. D. (2012). Job Attitudes. *Annual Review of Psychology*, *63*(1), 341–367. https://doi.org/10.1146/annurev-psych-120710-100511

Kouali, G. (2017). The instructional practice of school principals and its effect on teachers' job satisfaction. *International Journal of Educational Management*, *31*(7), 958–972. https://doi.org/10.1108/IJEM-11-2016-0253

Krug, S. E. (1992). Instructional Leadership: A Constructivist Perspective. *Educational Administration Quarterly*, *28*(3), 430–443. https://doi.org/10.1177/0013161X92028003012

Kushman, J. W. (1992). The Organizational Dynamics of Teacher Workplace Commitment: A Study of Urban Elementary and Middle Schools. *Educational Administration Quarterly*, *28*(1), 5–42. https://doi.org/10.1177/0013161X92028001002

Lee, M., Walker, A., & Chui, Y. L. (2012). Contrasting effects of instructional leadership practices on student learning in a high accountability context. *Journal of Educational Administration*, *50*(5), 586–611. https://doi.org/10.1108/09578231211249835

Leithwood, K., & Jantzi, D. (2005). A Review of Transformational School Leadership Research 1996–2005. *Leadership and Policy in Schools*, *4*(3), 177–199. https://doi.org/10.1080/15700760500244769

Leithwood, K., & Riehl, C. (2003). *What we know about successful school leadership*. Nottingham: National College for School Leadership Nottingham. Retrieved from http://www.leadersdesktop.sa.edu.au/leadership/files/links/School_leadership.pdf

Leithwood, K., Seashore Louis, K., Anderson, S., & Wahlstrom, K. (2004). *Review of research: How leadership influences student learning*. Report Commissioned by the Wallace Foundation. Retrieved from http://conservancy.umn.edu/handle/11299/2035

Levine, D. U., & Lezotte, L. W. (1990). *Unusually Effective Schools: A Review and Analysis of Research and Practice*. Madison: National Center for Effective Schools Research and Development. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0924345900010305

Linnansaari-Rajalin, T., Kivimäki, M., Ervasti, J., Pentti, J., Vahtera, J., & Virtanen, M. (2015). School neighbourhood socio-economic status and teachers' work commitment in Finland: longitudinal survey with register linkage. *Teachers and Teaching*, *21*(2), 131–149. https://doi.org/10.1080/13540602.2014.928128

Lipowsky, F., Faust, G., & Greb, K. (2009). *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE) – Teil 1* (Vol. 1). Frankfurt am Main: GFPF, DIPF.

Liu, Y., Bellibas, M. S., & Printy, S. (2016). How school context and educator characteristics predict distributed leadership: A hierarchical structural equation model with 2013 TALIS data. *Educational Management Administration & Leadership*, 1741143216665839. https://doi.org/10.1177/1741143216665839

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 1293–1349). Chicago: Rand McNally.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in

the Evaluation of Group-Level Effects. *Educational Psychologist*, *47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error. *Multivariate Behavioral Research*, *44*(6), 764–802. https://doi.org/10.1080/00273170903333665

Martin, C. L., & Bennett, N. (1996). The Role of Justice Judgments in Explaining the Relationship between Job Satisfaction and Organizational Commitment. *Group & Organization Management*, *21*(1), 84–104. https://doi.org/10.1177/1059601196211005

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School Leadership that Works: From Research to Results*. Association for Supervision and Curriculum Development.

Mathieu, J. E., & Zajac, D. M. (1990). A review and meta-analysis of the antecedents, correlates, and consequences of organizational commitment. *Psychological Bulletin*, *108*(2), 171.

Meyer, J. P., Stanley, D. J., Herscovitch, L., & Topolnytsky, L. (2002). Affective, Continuance, and Normative Commitment to the Organization: A Meta-analysis of Antecedents, Correlates, and Consequences. *Journal of Vocational Behavior*, *61*(1), 20–52. https://doi.org/10.1006/jvbe.2001.1842

Mowday, R. T., Steers, R. M., & Porter, L. W. (1978). *The measurement of organizational commitment: A progress report*. Graduate School of Management, University of Oregon.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus software (Version 7)*. Los Angeles: Muthén & Muthén.

Newmann, F. M. (Ed.). (2002). *Achieving High- Level Outcomes for all Students: The Meaning of Staff - Shared Understanding and Commitment*. Thousand Oaks, California.

Nielsen, K., & Daniels, K. (2012). Does shared and differentiated transformational leadership predict followers' working conditions and well-being? *The Leadership Quarterly*, *23*(3), 383–397. https://doi.org/10.1016/j.leaqua.2011.09.001

OECD. (2009). *Creating effective teaching and learning environments : first results from TALIS.* Paris: OECD Publishing.

OECD. (2014). *TALIS 2013 Results*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264196261-e

OECD. (2016a). PISA 2015 Results (Volume II). https://doi.org/10.1787/9789264267510-en

OECD. (2016b). *School Leadership for Learning*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264258341-en

Ostroff, C. (1992). The relationship between satisfaction, attitudes, and performance: An organizational level analysis. *Journal of Applied Psychology*, *77*(6), 963.

Robbins, S. P., & Judge, T. (2017). *Organizational behavior* (17th ed.). Boston: Pearson.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. *American Educational Research Journal*, *50*(1), 4–36. https://doi.org/10.3102/0002831212463813

Schaarschmidt, U., & Kieschke, U. (2013). Beanspruchungsmuster im Lehrerberuf – Ergebnisse und Schlussfolgerungen aus der Potsdamer Lehrerstudie. In M. Rothland (Ed.), *Belastung und Beanspruchung im Lehrerberuf: Modelle, Befunde, Interventionen* (2nd ed., pp. 81–97). Wiesbaden: Springer VS.

Schwarz, A., & Brauckmann, S. (2015). *Between facts and perceptions: The area close to school as a context factor in school leadership*. Schumpeter Discussion Papers. Retrieved from https://www.econstor.eu/handle/10419/111690

Seashore Louis, K., & Lee, M. (2016). Teachers' capacity for organizational learning: the effects of school culture and context. *School Effectiveness and School Improvement*, *27*(4), 534–556. https://doi.org/10.1080/09243453.2016.1189437

Seashore Louis, K., Leithwood, K., Wahlstrom, K. L., & Anderson, S. (2010). *Learning From Leadership: Investigating the Links to Improved Student Learning* (Commissioned by The Wallace Foundation). University of Minnesota, University of Toronto.

Sheppard, B. (1996). Exploring the Transformational Nature of Instructional Leadership. *Alberta Journal of Educational Research*, *42*(4), 325–344.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Smith, P., & Bell, L. (2011). Transactional and transformational leadership in schools in challenging circumstances: a policy paradox. *Management in Education*, *25*(2), 58–61. https://doi.org/10.1177/0892020611399608

Solinger, O. N., van Olffen, W., & Roe, R. A. (2008). Beyond the three-component model of organizational commitment. *The Journal of Applied Psychology*, *93*(1), 70–83. https://doi.org/10.1037/0021-9010.93.1.70

Swaffield, S., & MacBeath, J. (2009). Leadership for Learning. In J. MacBeath & N. Dempster (Eds.), *Connecting Leadership and Learning: Principles for Practice* (pp. 32–52). London: Routledge.

Townsend, T., & MacBeath, J. (Eds.). (2011). *International Handbook of Leadership for Learning*. Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-1350-5

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). Slough: NFER.

UNESCO Institute for Statistics. (2012). *International standard classification of education: ISCED 2011*. Montreal, Quebec: UNESCO Institute for Statistics. Retrieved from http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf

Urick, A., & Bowers, A. J. (2017). Assessing International Teacher and Principal Perceptions of Instructional Leadership: A Multilevel Factor Analysis of TALIS 2008. *Leadership and Policy in Schools*, 1–21. https://doi.org/10.1080/15700763.2017.1384499

Wagner III, J. A., Leana, C. R., Locke, E. A., & Schweiger, D. M. (1997). Cognitive and motivational frameworks in US research on participation: A meta-analysis of primary effects. *Journal of Organizational Behavior*, 49–65.

Wahlstrom, K. L., & Louis, K. S. (2008). How Teachers Experience Principal Leadership: The Roles of Professional Community, Trust, Efficacy, and Shared Responsibility. *Educational Administration Quarterly*, *44*(4), 458–495. https://doi.org/10.1177/0013161X08321502

Ware, H., & Kitsantas, A. (2007). Teacher and Collective Efficacy Beliefs as Predictors of Professional Commitment. *The Journal of Educational Research*, *100*(5), 303–310. https://doi.org/10.3200/JOER.100.5.303-310

Yammarino, F. J., Dionne, S. D., Chung, J. U., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *The Leadership Quarterly*, *16*(6), 879–919. https://doi.org/10.1016/j.leaqua.2005.09.002

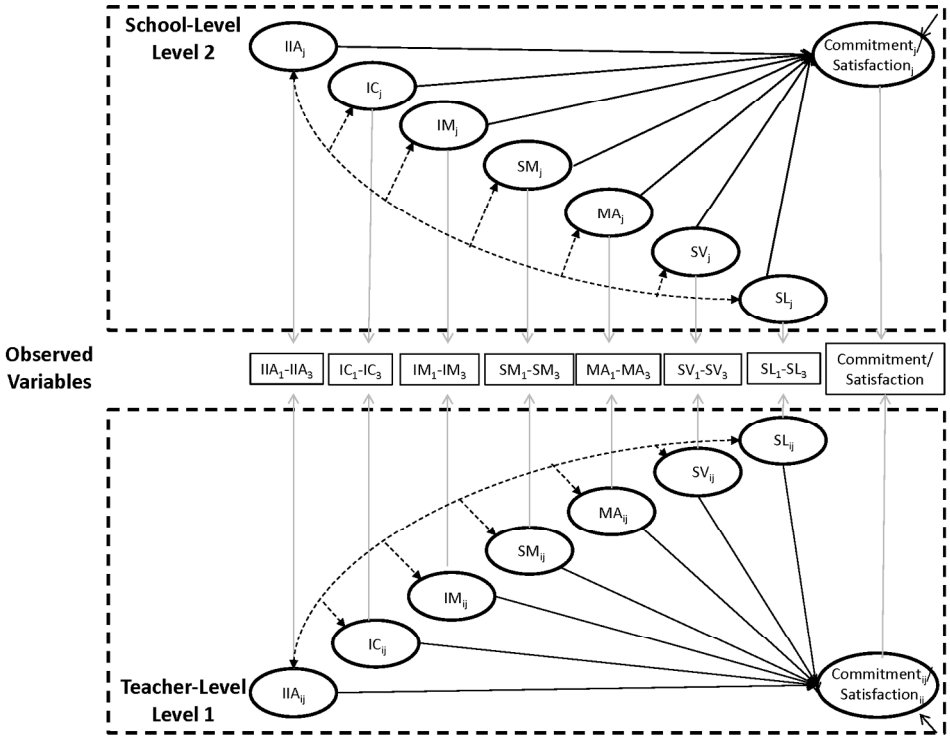


Figure 1: Illustration of the 7-7 Multilevel Structural Equation Model

254x190mm (200 x 200 DPI)

Table 1: Estimates of Variance, Reliability and Agreement

	Mean	SD	ICC1	ICC2	$r_{wg}(j)$
Transformational Leadership (TL)	**3.05**	**.70**	**.291**	**.925**	**.888**
Idealized Influence attrributed (IIa)	2.99	.83	.312	.932	.732
Individual Consideration (IC)	2.97	.85	.217	.893	.668
Inspirational Motivation (IM)	3.19	.69	.374	.947	.813
Instructional Leadership (IL)	**2.60**	**.67**	**.272**	**.918**	**.881**
School Management (SM)	2.84	.74	.239	.904	.744
Instructional Management (MA)	2.72	.78	.338	.939	.718
Supervision (SV)	2.10	.74	.255	.911	.716
Shared Leadership (SL)	**3.03**	**.69**	**.262**	**.914**	**.780**

Table 2: Fit Indices of the scrutinized two-level LFL measurement models

LEVEL 1 **Teacher Level**	**1 factor**	**3 factors**: TL, IL, SL	**3 factors**: TL, IL, SL	**7 factors**: TL (IIa, IM, IC), IL (SM, MA, SV), SL	**7 factors**: TL (IIa, IM, IC), IL (SM, MA, SV), SL	**7 factors**: TL (IIa, IM, IC), IL (SM, MA, SV), SL
LEVEL 2 **School Level**	**1 factor**	**1 factor**	**3 factors**: TL, IL, SL	**1 factor**	**3 factors**: TL, IL, SL	**7 factors**: TL (IIa, IM, IC), IL (SM, MA, SV), SL
Chi2/ d.f.	7440.636/ 341	4520.203/ 338	4370.612/ 337	2362.199/ 320	2203.100/ 319	1972.753/ 305
CFI	.798	.881	.885	.942	.946	.953
RMSEA	.075	.058	.057	.041	.040	.038
$SRMR_w$	.069	.048	.048	.032	.032	.032
$SRMR_b$	.097	.093	.077	.097	.078	.068
AIC	128,782.364	125,872.340	125,729.007	123,795.903	123,646.694	123,467.979
BIC	129,398.450	126,506.549	126,369.543	124,542.157	124,399.167	124,307.516

Table 3a: Results of the Multilevel Structural Equation Models for Teachers' Organizational Commitment (standardized regression coefficients, standard errors, *p*-values and effect sizes)

	Model 1	*SE*	*p*	*ES*	Model 2	*SE*	*p*	*ES*	Model 3	*SE*	*p*	*ES*
within (Individual Teacher Level)												
Transformational Leadership (TL)												
Idealized Influence attrributed (IIa)	.205	.051	*.000*	*0.44*	.205	.051	*.000*	*0.44*	.206	.051	*.000*	*0.44*
Individual Consideration (IC)	.183	.038	*.000*	*0.38*	.183	.038	*.000*	*0.38*	.180	.038	*.000*	*0.38*
Inspirational Motivation (IM)	-.023	.032	*.470*	*-0.06*	-.023	.032	*.467*	*-0.06*	-.025	.032	*.436*	*-0.06*
Instructional Leadership (IL)												
School Management (SM)	.054	.066	*.409*	*0.14*	.054	.066	*.411*	*0.14*	.053	.066	*.420*	*0.14*
Instructional Management (MA)	.019	.097	*.843*	*0.05*	.019	.097	*.842*	*0.05*	.021	.098	*.833*	*0.05*
Supervision (SV)	-.008	.036	*.828*	*-0.02*	-.008	.036	*.828*	*-0.02*	-.008	.036	*.822*	*-0.02*
Shared Leadership (SL)	.374	.045	*.000*	*0.84*	.374	.045	*.000*	*0.84*	.377	.045	*.000*	*0.85*
between (School Level)												
Transformational Leadership (TL)												
Idealized Influence attrributed (IIa)	-.272	.399	*.495*	*-0.37*	-.258	.398	*.517*	*-0.35*	-.267	.364	*.464*	*-0.29*
Individual Consideration (IC)	.036	.373	*.923*	*0.05*	.015	.369	*.967*	*0.02*	.050	.324	*.877*	*0.08*
Inspirational Motivation (IM)	.012	.175	*.945*	*0.02*	.024	.172	*.889*	*0.03*	.056	.169	*.741*	*0.08*
Instructional Leadership (IL)												
School Management (SM)	.236	.249	*.344*	*0.32*	.200	.243	*.409*	*0.27*	.126	.255	*.621*	*0.23*
Instructional Management (MA)	.283	.201	*.159*	*0.39*	.314	.198	*.112*	*0.43*	.350	.182	*.055*	*0.41*
Supervision (SV)	-.193	.182	*.290*	*-0.26*	-.176	.177	*.321*	*-0.24*	-.186	.181	*.303*	*-0.30*
Shared Leadership (SL)	.728	.212	*.001*	*1.00*	.712	.208	*.001*	*0.97*	.696	.231	*.003*	*0.92*
Context Variables												
School in Challenging Circumstances					.111	.053	*.037*		.056	.055	*.307*	
Cumpolsery School[+]									-.273	.090	*.002*	
Higher Secondary School[+]									-.012	.075	*.873*	
School Size									.117	.134	*.383*	
R^2 within	.542	.020	*.000*		.542	.020	*.000*		.542	.020	*.000*	
R^2 between	.671	.070	*.000*		.679	.067	*.000*		.714	.063	*.000*	

+ Reference Group = Primary School
Note: Effect sizes were only calculated for LFL facets

Table 3b: Results of the Multilevel Structural Equation Models for Teachers' Job Satisfaction (standardized regression coefficients, standard errors, *p*-values and effect sizes)

	Model 1	*SE*	*p*	*ES*	Model 2	*SE*	*p*	*ES*	Model 3	*SE*	*p*	*ES*
within (Individual Teacher Level)												
Transformational Leadership (TL)												
Idealized Influence attrributed (IIa)	.009	.053	*.866*	*0.02*	.009	.053	*.866*	*0.02*	.010	.053	*.843*	*0.02*
Individual Consideration (IC)	.399	.032	*.000*	*0.78*	.399	.032	*.000*	*0.78*	.396	.032	*.000*	*0.77*
Inspirational Motivation (IM)	.003	.029	*.928*	*0.01*	.003	.029	*.932*	*0.01*	.001	.029	*.969*	*0.06*
Instructional Leadership (IL)												
School Management (SM)	.030	.060	*.615*	*0.08*	.030	.060	*.616*	*0.08*	.029	.060	*.635*	*0.07*
Instructional Management (MA)	-.062	.089	*.487*	*-0.14*	-.062	.089	*.487*	*-0.14*	-.061	.090	*.495*	*-0.14*
Supervision (SV)	.034	.032	*.285*	*0.09*	.034	.031	*.285*	*0.09*	.034	.032	*.281*	*0.09*
Shared Leadership (SL)	.445	.036	*.000*	*0.95*	.445	.036	*.000*	*0.95*	.447	.036	*.000*	*0.96*
between (School Level)												
Transformational Leadership (TL)												
Idealized Influence attrributed (IIa)	-.363	.330	*.270*	*-0.51*	-.351	.330	*.288*	*-0.49*	-.330	.292	*.313*	*-0.36*
Individual Consideration (IC)	.210	.309	*.496*	*0.29*	.194	.308	*.528*	*0.27*	.190	.298	*.522*	*0.32*
Inspirational Motivation (IM)	.010	.153	*.948*	*0.01*	.021	.137	*.891*	*0.03*	.045	.153	*.771*	*0.06*
Instructional Leadership (IL)												
School Management (SM)	.231	.213	*.279*	*0.32*	.199	.208	*.339*	*0.28*	.158	.216	*.463*	*0.29*
Instructional Management (MA)	.249	.178	*.162*	*0.34*	.276	.173	*.112*	*0.38*	.293	.169	*.083*	*0.35*
Supervision (SV)	-.154	.158	*.330*	*-0.21*	-.138	.154	*.372*	*-0.19*	-.140	.159	*.379*	*-0.22*
Shared Leadership (SL)	.721	.186	*.000*	*1.00*	.704	.181	*.000*	*0.97*	.687	.198	*.001*	*0.91*
Context Variables												
School in Challenging Circumstances					.094	.046	*.041*		.074	.049	*.131*	
Cumpolsery School[+]									-.084	.079	*.289*	
Higher Secondary School[+]									.007	.064	*.913*	
School Size									-.004	.112	*.972*	
R^2 within	.616	.017	*.000*		.616	.017	*.000*		.616	.017	*.000*	
R^2 between	.783	.055	*.000*		.788	.052	*.000*		.789	.051	*.000*	

+ Reference Group = Primary School
Note: Effect sizes were only calculated for LFL facets

Appendix A: Details of the Social Composition for Schools with Social Index 1 and 2 and 3 to 6

	Social Index 1 and 2 $n_{schools}=33$	**Social Index 3 to 6** $n_{schools}=93$
Individual Data		
ISCED 5a min. Father *(tertiary education)*	5.60 %	27.90 %
ISCED 5a min. Mother *(tertiary education)*	5.10 %	22.10 %
ISCED 2 max. Father *(secondary education first stage/ secondstep of basic education)*	27.20 %	9.40 %
ISCED 2 max. Mother *(secondary education first stage/ secondstep of basic education)*	25.60 %	8.30 %
EGP VI Father (*Skilled manual workers*)	32.60 %	11.60 %
EGP I Father *establishments; large proprietors*) (*Higher-grade professionals, administrators, and officials; managers in large industrial establishments; large proprietors*)	10.30 %	36.50 %
Father born in Germany	42.50 %	73.80 %
Mother born in Germany	46.80 %	75.40 %
Context Data		
Turnout of Voters at Residence (mean)	45.70 %	60.40 %
Unemployment Rate at Residence (mean)	8.60 %	4.90 %
Rate of Welfare Recipients at Residence (mean)	21.40 %	8.50 %

Appendix B: Lower triangle: Latent Correlations of the Leadership facets (Level1/Level2); Diagonal: Scale Reliability Estimates (Omega Level1/Omega Level2)

		Transformational			Instructional			Shared
		IIA	**IC**	**IM**	**SM**	**MA**	**SV**	
Transformational	Idealized Influence attrributed (IIA)	**.89/.99**						
	Individual Consideration (IC)	.84/.92	**.90/.98**					
	Inspirational Motivation (IM)	.66/.85	.67/.73	**.80/.95**				
Instructional	School Management (SM)	.68/.81	.62/.77	.66/.73	**.79/.93**			
	Instructional Management (MA)	.76/.84	.74/.85	.67/.68	.88/.91	**.78/.96**		
	Supervision (SV)	.55/.71	.51/.68	.52/.68	.75/.92	.79/.84	**.76/.84**	
Shared		.75/.87	.73/.91	.67/.65	.70/.71	.77/.80	.55/.65	**.81/.93**

Bernd Groot-Wilken, Kevin Isaac,
Jörg-Peter Schräpler (Hrsg.)

Sozialindices für Schulen

Hintergründe, Methoden und Anwendung

Waxmann 2016
Münster • New York

Klaudia Schulte, Johannes Hartig & Marcus Pietsch

Berechnung und Weiterentwicklung des Sozialindex für Hamburger Schulen[1]

In Hamburg gibt es seit 1996 einen Sozialindex für Grundschulen und weiterführende Schulen mit Sekundarstufe I. Auch einige weitere deutsche Bundesländer wie Bremen, Berlin und Nordrhein-Westfalen nutzen verschieden operationalisierte Indikatoren der sozialen Belastung, um gezielt schulische Ressourcen einzusetzen (Tillmann & Weishaupt, 2015; siehe auch den Beitrag von Weishaupt in diesem Band). Durch den Einsatz von Sozialindices sollen Schulen in schwierigen Lagen mit zusätzlichen Mitteln unterstützt werden, „um Effekte der Schülerzusammensetzung kompensieren und chancenausgleichend wirken zu können: Gleiche Bildungschancen sollen mit ungleichem Mitteleinsatz erreicht werden." (Tillmann und Weishaupt, 2015, S. 7). Auch in Hamburg beschreibt der Sozialindex, seit 2006 (Bos, Pietsch, Gröhlich & Janke, 2006) basierend u.a. auf der Kapitaltheorie von Bourdieu (1982; 1983), die sozialen Rahmenbedingungen der Schulen. Die auf dem Index basierende Zuordnung zu sechs abgestuften Belastungsgruppen hat Auswirkungen auf diversen Ebenen: Auf der einen Seite determiniert der Sozialindex unterschiedliche Ressourcenallokationen (z.B. kleinere Klassenfrequenzen oder höhere Sprachfördermaßnahmen für Schulen mit niedrigeren Indices). Auf der anderen Seite wird der Sozialindex in Hamburg auch in weiteren Zusammenhängen genutzt: bei der Bildung repräsentativer Stichproben im Rahmen von wissenschaftlichen Untersuchungen und Evaluationen (z.B. bei der Auswahl einer repräsentativen Kernstichprobe von Schulen pro Schuljahr für die Schulinspektion), bei der Berechnung und Rückmeldung von Vergleichswerten („fairer Vergleich") für die schulbezogenen Ergebnisrückmeldungen im Rahmen von KERMIT oder bei der Bildung von Vergleichsgruppen im Kontext der Bildungsberichterstattung. Hamburg reagiert damit bildungspolitisch auf den über die Jahre leicht entkoppelten, aber auch aktuell noch beschriebenen Zusammenhang zwischen der sozialen Herkunft und dem Kompetenzerwerb sowie damit verbundenen Bildungschancen in Deutschland, wie auch in den PISA-Ergebnissen gezeigt werden konnte (Klieme et al., 2010).

Im folgenden Artikel werden, nach einer theoretischen Einführung in das zugrunde liegende Konzept der sozialen Belastung, Durchführung und Methode der Berechnung des Sozialindex dargestellt sowie aktuelle Überlegungen zur Weiterentwicklung des Sozialindex aufgeführt, die sich aus den inzwischen langjährigen Erfahrungen mit der Verwendung eines Sozialindex für Ressourcenallokationen ergeben haben.

---

1 Überarbeiteter und ergänzter Nachdruck von Schulte, Hartig und Pietsch (2014).

1. Theoretische Fundierung des Sozialindex

In der Tradition der bisherigen Sozialindices wird die soziale Belastung Hamburger Schulen über ein theoretisches Modell beschrieben, welches verschiedene Aspekte der sozialen Belastung voneinander unterscheidet:

- soziales Kapital
- ökonomisches Kapital
- kulturelles Kapital
- Migrationshintergrund

Die drei erstgenannten Facetten orientieren sich an Bourdieus Ansatz der Kapitalarten (1982; 1983), welcher ressourcenorientiert Differenzen in bestehenden sozialen Ungleichheiten und deren Reproduktion in den Familien der Schülerinnen und Schüler aber auch im Bildungssystem (Bourdieu & Passeron, 1971) aufzeigt. Zentral bei Bourdieu ist die ungleiche Verteilung von Macht zur Verfolgung eigener Interessen und Ziele. Das soziale Kapital, welches auch von Coleman (1988) beschrieben wird, meint das Netzwerk sozialer Beziehungen, welches Personen über die Zugehörigkeit zu einer Gemeinschaft und die damit verbundenen Pflichten in ihrem Bildungserfolg unterstützt. Bourdieu beschreibt das „Sozialkapital" als „die Gesamtheit der aktuellen und potentiellen Ressourcen, die mit dem Besitz eines dauerhaften Netzes von mehr oder weniger institutionalisierten Beziehungen gegenseitigen Kennens oder Anerkennens verbunden sind (...)" (Bourdieu, 1992, S. 63). Das ökonomische Kapital bildet in Bourdieus Theorie die Bedeutung von Kapital im finanziellen Sinne ab, z.B. das Einkommen. Beim kulturellen Kapital unterscheidet Bourdieu (1992) drei Zustände:

1. Inkorporierter Zustand: Bildung der jeweiligen Person, die sie sich durch die Investition von Zeit angeeignet hat.
2. Objektivierter Zustand: kulturelle Güter, wie z.B. Bücher oder Kunstgegenstände.
3. Institutionalisierter Zustand: die Erlangung von Titeln im Bildungsverlauf, z.B. der Erwerb eines Schulabschlusses.

Des Weiteren werden bei der Konstruktion des Hamburger Sozialindex Migrationshinweise der Schülerinnen und Schüler genutzt (Bonsen et al., 2010), da auch in Bezug auf den Migrationshintergrund – selbst unter Kontrolle weiterer sozioökonomischer Hintergrundmerkmale – noch immer bedeutsame soziale Disparitäten nachgewiesen werden können (Klieme et al., 2010; siehe auch den Beitrag von Kemper in diesem Band).

Für eine breitere Fundierung des Sozialindex, über die hinsichtlich der verschiedenen theoretischen Facetten auf Schülerebene erfassten Indikatoren hinaus, wurden zusätzlich Daten des Statistikamts Nord herangezogen. Dabei handelt es sich um Sozialraumdaten (z.B. die Arbeitslosenquote), die auf Ebene der Statistischen Gebiete vorliegen: „Statistische Gebiete sind kleinräumige Gebietseinheiten, die nach städtebaulichen und sozialstrukturellen Homogenitätskriterien im Anschluss an die Volkszählung 1987 gebildet wurden." (Freie und Hansestadt Hamburg, 2012, S. 22). Diese

Daten können gerade auch bei hohen Datenausfällen oder möglichen Verzerrungen durch selektives Beantworten wertvolle Informationen liefern (Pietsch, Bonsen & Bos, 2006).

2. Methode

2.1 Stichprobenziehung

Mit Ausnahme der Sonder- und Förderschulen nahmen alle staatlichen Grundschulen und staatlichen weiterführenden Schulen an der Schüler- und Elternbefragung teil. Den nichtstaatlichen Schulen stand die Teilnahme an den Befragungen und damit auch die Berechnung eines schulischen Sozialindex frei. Zur Sicherung einer für jede Schule repräsentativen Auswahl von Schülerinnen und Schülern wurden pro Schule einfache Zufallsstichproben (simple random sample) intakter Klassen der Klassenstufen drei bis neun gezogen. Die Anzahl der Klassen war abhängig von der Schulgröße: Bei Schulen mit bis zu 100 Schülerinnen und Schülern gab es Vollerhebungen, dies war vor allem bei kleineren Grundschulsystemen der Fall. In Schulen mit 100 bis 400 Schülerinnen und Schülern wurden vier Klassen ausgewählt, bei mehr als 400 Schülerinnen und Schülern fünf Klassen. Bei Schulen, die sowohl einen Grundschul- als auch einen Sekundarschulzweig führen, wurden für beide Bereiche einzelne Teilstichproben gezogen. Insgesamt ergab sich so eine Stichprobengröße von N = 35.437.

2.2 Durchführung der Erhebung

Die Schulen erhielten wenige Wochen vor der Erhebung ein erstes Informationspaket. In diesem waren Informationsmaterialien für die Schulleitung, Manuale zur Durchführung der Erhebung für die jeweiligen Klassenlehrkräfte sowie Flyer und Anschreiben inklusive Einwilligungserklärungen für die Eltern enthalten. Flyer und auch Anschreiben lagen in fünf weiteren Sprachen vor (Türkisch, Russisch, Farsi, Englisch, Französisch). Die eigentlichen Befragungen fanden nur für die Schülerinnen und Schüler statt, deren Eltern ihr schriftliches Einverständnis zur Teilnahme gegeben hatten. Für die Durchführung der Schülerbefragung im Klassenverband waren die jeweiligen Klassenlehrkräfte verantwortlich. Des Weiteren erhielten die Eltern aller Schülerinnen und Schüler einen Fragebogen, der zu Hause ausgefüllt und im verschlossenen Umschlag wieder mit in die Schule gebracht werden sollte. Der achtseitige Fragebogen für die Eltern enthielt 20 Fragen, der Fragebogen für die Schülerinnen und Schüler umfasste 45 Fragen. Die Familienzugehörigkeit der Eltern- und Schülerfragebogen war durch einen Code, welcher auf die Fragebogen gedruckt war, gekennzeichnet.

2.3 Rücklauf und Stichprobe

Der durchschnittliche Rücklauf in den Schulen lag (unter Berücksichtigung der Vorlage eines Schüler und/oder Elternfragebogens) bei etwa 69% der Fälle. Insgesamt ergaben sich 24.452 ausgefüllte Schüler- und/oder Elternfragebogen an 332 Schulen. Die Verteilung der Schülerinnen und Schüler auf die befragten Klassenstufen im Vergleich zu den gezogenen Teilstichproben ist in Tabelle 1 aufgeführt.

Tab. 1: Anzahl der Fälle (Fragebogen von Schülern und / oder Eltern), in denen ein Fragebogen vorlag sowie Anzahl der ausgefüllten Fragebogen von Schülerinnen und Schülern und Eltern pro Jahrgangsstufe

	Klasse 3	Klasse 4	Klasse 5	Klasse 6	Klasse 7	Klasse 8	Klasse 9
Anzahl der ausgewählten Schülerinnen und Schüler	9.560	9.719	3.313	3.028	2.977	3.507	3.333
Anzahl der Fälle (Familien), in denen ein Fragebogen vorlag	7.016	6.876	2.348	2.054	1.993	2.094	2.071
Anzahl der ausgefüllten Schülerfragebogen	6.354	6.359	2.267	1.956	1.923	2.023	1.982
Anzahl der ausgefüllten Elternfragebogen	6.371	6.140	2.027	1.770	1.691	1.742	1.662

2.4 Methodisches Vorgehen

Im Gegensatz zu den bisherigen Berechnungen des Hamburger Sozialindex (z.B. Bos, Gröhlich & Bonsen, 2009) wurde keine Item-Response-Theorie-basierte Partial-Credit-Skalierung der Daten auf Schülerebene vorgenommen. Stattdessen wurden exploratorische und konfirmatorische Faktorenanalysen mit aggregierten Daten auf Schulebene durchgeführt. Für das gewählte Vorgehen sprachen aus unserer Sicht die folgenden Gründe:

- Die Interpretation der Belastungswerte erfolgt auf Schulebene, dementsprechend erscheint uns dies auch als die angemessene Ebene der Datenanalyse. Zudem erlaubt die Analyse auf Schulebene auch die Integration der Daten des Statistikamts Nord, die nicht auf Individualebene vorliegen.
- Die Spezifikation eines Messmodells im Rahmen von Strukturgleichungsmodellen erlaubt eine bessere Beurteilung der Modellanpassung als dies im Kontext der Item-Response-Theorie möglich ist, wo keine etablierten Maße für globale Modellgüte verfügbar sind.
- Das Vorgehen ermöglicht es, dass die verwendeten Variablen mit unterschiedlichen, frei geschätzten Gewichtungen in den Belastungswert eingehen. Hierdurch werden Merkmale, die stärker zur Differenzierung von Schulen beitragen, im Index höher gewichtet. Bei dem bisher berechneten Partial-Credit-Modell wurden alle Variablen gleich gewichtet.

Um einen direkten Vergleich der Belastungswerte von Grund- und Sekundarschulen zu ermöglichen, war das Ziel, ein Modell für alle Schulen zu berechnen, und nicht wie bisher zwei separate Skalierungen mit unterschiedlichen Variablensätzen durchzuführen. Dem zugrunde liegt die Forderung, dass das Konstrukt sozialer Belastung über Schulformen hinweg konsistent definiert wird, d.h. Belastungsmerkmale für Kinder aus Grundschulen grundsätzlich die gleichen sind wie solche für Schülerinnen und Schüler an weiterführenden Schulen.

3. Ergebnisse

Im Vorfeld der Auswertungen wurden die mithilfe der per Fragebogen erhobenen Individualdaten aufbereitet und auf Schulebene aggregiert. Auch für die Daten des Statistikamts Nord wurde in einem mehrstufigen Vorgehen ein Datensatz auf Schulebene erzeugt: Für jedes Statistische Gebiet in Hamburg existieren Durchschnittswerte der Belastung in Bezug auf verschiede Variablen. Zunächst wurden diese Daten aus dem Statistikamt Nord für alle Statistischen Gebiete in Hamburg in einen Datensatz exportiert. Dabei wurden alle grundsätzlich in Frage kommenden acht Variablen einbezogen (siehe Tabelle 2). Dieser aus den Daten des Statistikamts Nord erstellte Datensatz wurde mit einem Datensatz der Schulstatistik zusammengeführt. In diesem Datensatz der Schulstatistik gibt es für jede Schülerin und jeden Schüler in Hamburg eine Angabe über das Statistische Gebiet, in dem er oder sie lebt. Somit kann jedem Kind auch der durchschnittliche Wert für die ausgewählten acht Variablen für das Statistische Gebiet, in dem es lebt, zugeordnet werden. Darüber hinaus ist in dem Datensatz der Schulstatistik verzeichnet, auf welche Schule die Kinder gehen. In einem letzten Schritt konnte durch diese Schulzugehörigkeit der Kinder ein Mittelwert auf Schulebene für jede Variable aus den Daten des Statistikamts Nord gebildet werden. Der finale Datensatz ergab sich sodann aus der Zusammenführung des Befragungsdatensatzes auf Schulebene mit den Daten des Statistikamts Nord auf Schulebene.

Tab. 2: In die Variablenauswahl eingehende Daten des Statistikamts Nord

	Daten des Statistikamts Nord
1	Arbeitslosenquote bei Einwohnern zwischen 15 und 65 Jahren
2	Anteil der Empfängerinnen und Empfänger von Grundsicherung an der Bevölkerung über 65 Jahre
3	Anteil der Bevölkerung mit Migrationshintergrund an der Gesamtbevölkerung
4	SGBII-Empfängerinnen und -Empfänger
5	Ausländische erwerbsfähige SGBII-Empfängerinnen und Empfänger zwischen 15 und 65 Jahren
6	Erwerbsfähige hilfebedürftige Jugendliche zwischen 15 und 25 Jahren
7	Nichterwerbsfähige Hilfebedürftige unter 15 Jahren
8	Wahlbeteiligung

3.1 Variablenauswahl und Modellbildung

Das Vorgehen der Variablenauswahl war mehrstufig. Ziel war es, aus den Befragungsdaten und den Variablen des Statistikamts Nord ein sparsames Modell zu finden, welches nichtsdestotrotz jeweils mehrere Indikatoren aus allen abzubildenden inhaltlichen Bereichen beinhaltet. Für die ordinalen Variablen Bildungsabschlüsse der Eltern und EGP-Klassen wurden jeweils Dummy-Variablen für die niedrigsten und höchsten Kategorien gebildet (Hauptschulabschluss und Universitätsabschluss bzw. EGP-Klassen 1 und 6), da diese im Unterschied zu den mittleren Kategorien bezüglich des Sozialstatus eine eindeutige Wertigkeit haben. Die EGP-Klassen (Erikson, Goldthorpe & Portocarero, 1979; Erikson & Goldthorpe, 2002) beschreiben qualitative Unterschiede in Bezug auf die Berufe der Eltern und ordnen diese nach der Art der Tätigkeit, der Stellung im Beruf, der Weisungsbefugnis und der Qualifikation. Für die Analysen wurden die originären elf EGP-Klassen zu sechs Klassen zusammengefasst, wie auch in den Berechnungen zum Programme for International Student Assessment (PISA) realisiert (Prenzel et al. 2007) (siehe Tabelle 3). Darüber hinaus wurde die Varianz der Variablen zwischen Schulen (Intraklassenkorrelation) als Auswahlkriterium herangezogen, da Schülervariablen, hinsichtlich derer sich Schulen kaum unterscheiden, nicht zur Differenzierung auf Schulebene geeignet sind. Es wurden zehn Variablen ausgeschlossen, die Intraklassenkorrelationen unter 0.015 aufwiesen (Intraklassenkorrelationen der für das finale Modell ausgewählten Variablen siehe Tabelle 4).

Tab. 3: Beschreibung der EGP-Klassen

EGP-Klassen
Obere Dienstklasse (I)
Freie akademische Berufe, führende Angestellte, höhere Beamte, selbstständige Unternehmer mit mehr als 10 Mitarbeiter-/innen, Hochschul- und Gymnasiallehrkräfte
Untere Dienstklasse (II)
Angehörige von Semiprofessionen, mittleres Management, Beamte im mittleren und gehobenen Dienst, technische Angestellte mit nicht manueller Tätigkeit
Routinedienstleistungen Handel und Verwaltung (III)
Büro- und Verwaltungsberufe mit Routinetätigkeiten, Berufe mit niedrig qualifizierten, nicht-manuellen Tätigkeiten, die oftmals auch keine Berufsausbildung erfordern
Selbstständige (IV)
Selbstständige aus manuellen Berufen mit wenigen Mitarbeitern und ohne Mitarbeiter, Freiberufler, sofern sie keinen hoch qualifizierten Beruf haben
Facharbeiter und Arbeiter mit Leistungsfunktion (V)
Untere technische Berufe wie Vorarbeiter, Meister, Techniker, die in manuelle Arbeitsprozesse eingebunden sind; Aufsichtskräfte im manuellen Bereich
Un- und angelernte Arbeiter, Landarbeiter (VI)
Alle un- und angelernten Berufe aus dem manuellen Bereich, Dienstleistungstätigkeiten mit manuellem Charakter und geringem Anforderungsniveau, Arbeiter in der Land-, Forst- und Fischwirtschaft

Quelle: Prenzel et al., 2007, S. 313.

Mit den verbliebenen 53 Variablen wurde mit Mplus (Muthén & Muthén, 2012) eine explorative Faktorenanalyse mit obliquer Bi-Geomin-Rotation (Jennrich & Bentler, 2012) berechnet. Dieses Rotationsverfahren wurde gewählt, um einen varianzstarken Generalfaktor zu erhalten, der möglichst viele gemeinsame Unterschiede zwischen den Schulen abbildet. Da nicht erwartet wurde, dass eine einzige Dimension zur Beschreibung der Daten ausreicht, wurden zusätzlich korrelierte Residualfaktoren zugelassen. Diese können durch den Generalfaktor nicht repräsentierte, variablenspezifische Abhängigkeiten abbilden, sind jedoch für das hier verfolgte Ziel einer eindimensionalen Beschreibung der Belastung inhaltlich nicht von Interesse. Beim Vergleich von Modellen mit unterschiedlichen Faktorzahlen ergab sich für ein 8-Faktor-Modell eine nach gängigen Kriterien (vgl. Hu & Bentler, 1999) akzeptable Anpassung (χ^2 = 2319, df = 982; χ^2/df = 2.36; Root-Mean-Square-Error of Approximation (RMSEA) = .064; Comparative-Fit-Index (CFI) = 0.944; Tucker-Lewis Index (TLI) = 0.922). Auf Basis des aus der explorativen Faktorenanalyse gewonnen Modells wurde ein „konfirmatorisches" Strukturgleichungsmodell mit einem Generalfaktor und sieben korrelierten Residualfaktoren spezifiziert. Hierbei wurden für die Residualfaktoren die jeweils höchsten Ladungen aus dem exploratischen Modell freigesetzt, alle weiteren Ladungen wurden auf null fixiert. Das Modell wies eine akzeptable Passung auf (χ^2 = 3136, df = 1172; χ^2/df = 2.68; RMSEA = .071; CFI = 0.917; TLI = 0.906). Die Ladungen der Variablen auf dem Generalfaktor, der später zur Berechnung des Sozialindex herangezogen werden sollte, wurden als Selektionskriterium für die im finalen Modell behaltenen Variablen herangezogen.

Auf Grundlage dieses ersten Modells wurden nach den folgenden Kriterien 24 Variablen ausgewählt: Jeder der theoretisch angenommenen Bereiche sollte mit drei bis vier Indikatoren abgebildet sein. Ausgewählt wurden zunächst die Variablen, welche innerhalb jedes Bereichs die jeweils höchsten Ladungen auf dem Generalfaktor aufwiesen. Darüber hinaus sollen drei bis vier Variablen der Daten des Statistikamts Nord in die Berechnung eingehen. Aus diesen Daten wurden die Variablen ausgeschlossen, bei denen von einer starken Redundanz mit der für die Analyse ausgewählten Variable Arbeitslosenrate zu erwarten war (Anteil der Empfängerinnen und Empfänger von Grundsicherung an der Bevölkerung über 65 Jahre, SGBII-Empfängerinnen und -Empfänger, Ausländische erwerbsfähige SGBII-Empfängerinnen und Empfänger zwischen 15 und 65 Jahren, Erwerbsfähige hilfebedürftige Jugendliche zwischen 15 und 25 Jahren). Darüber hinaus sollte eine Überlappung mit den Befragungsdaten ausgeschlossen werden (Anteil der Bevölkerung mit Migrationshintergrund an der Gesamtbevölkerung). In die Analysen gingen daher die drei Variablen Arbeitslosenrate bei Einwohnern zwischen 15 und 65 Jahren, Nichterwerbsfähige Hilfebedürftige unter 15 Jahren sowie die Wahlbeteiligung ein. Die Variablen aus der Fragebogenerhebung sollten untereinander nicht redundant sein: war die Variable bei Eltern und Schülerinnen und Schülern vorhanden, wurde auf die Elternfrage zurückgegriffen. Wenn eine Variable aus den oben genannten Gründen die Eigenschaft eines Elternteils beschreibt, sollte immer auch der andere Elternteil herangezogen werden. Die ausgewählten Variablen inklusive der Generalfaktorladungen und Dimensionszuordnung sind in Tabelle 4 ersichtlich. Auch die sozialen Raumdaten wurden theoriegeleitet den Dimensionen

Tab. 4: Verwendete Variablen, differenziert nach theoretischen Dimensionen, Datenquellen, den Faktorladungen auf dem Generalfaktor sowie den Intraklassenkorrelationen

Variable	Datenquelle	Faktor-ladung	Intraklassen-korrelation
Dimension Kulturelles Kapital			
Anzahl der Bücher zu Hause	Elternfragebogen	-.986	.305
Häufigkeit des gemeinsamen Besuchs mit den Kindern im Museum	Elternfragebogen	-.848	.108
Bildungsabschluss Universität des Vaters	Elternfragebogen	-.819	.155
Bildungsabschluss Universität der Mutter	Elternfragebogen	-.800	.130
Bildungsabschluss Hauptschule des Vaters	Elternfragebogen	.859	.097
Bildungsabschluss Hauptschule der Mutter	Elternfragebogen	.893	.109
Dimension Ökonomisches Kapital			
Einkommen	Elternfragebogen	-.959	.325
EGP-Klasse 1 des Vaters	Elternfragebogen	-.847	.159
EGP-Klasse 1 der Mutter	Elternfragebogen	-.710	.066
EGP-Klasse 6 des Vaters	Elternfragebogen	.822	.109
EGP-Klasse 6 der Mutter	Elternfragebogen	.750	.092
Eigenes Zimmer für das Kind	Schülerfragebogen	-.875	.136
Anteil Arbeitslosigkeit	Soziale Raumdaten	.873	.630
Anteil Hilfebedürftige nicht-Erwerbsfähige	Soziale Raumdaten	.880	.634
Dimension Soziales Kapital			
Kind verbringt seine Freizeit mit Klassenkameraden	Schülerfragebogen	-.453	.024
Kind verbringt seine Freizeit mit den Eltern	Schülerfragebogen	-.416	.026
Die Eltern loben das Kind für eine gute Schulnote	Schülerfragebogen	-.396	.018
Die Eltern sind stolz auf das Kind	Schülerfragebogen	-.317	.020
Wahlbeteiligung	Soziale Raumdaten	-.766	.764
Dimension Migrationshinweise			
Geburtsland Vater	Elternfragebogen	-.883	.167
Geburtsland Mutter	Elternfragebogen	-.834	.149
Sprachhäufigkeit Deutsch mit der Mutter	Schülerfragebogen	-.874	.124
Sprachhäufigkeit Deutsch mit dem Vater	Schülerfragebogen	-.890	.109
Sprachhäufigkeit Deutsch mit den Geschwistern	Schülerfragebogen	-.730	.040

zugeordnet, so zählten Arbeitslosigkeit und Hilfebedürftige Kinder zu dem ökonomischen Kapital, die Wahlbeteiligung wäre als soziales Kapital zu verorten.

3.2 Überprüfung des finalen Modells

Zur Überprüfung und Spezifikation des finalen Modells wurde mit den ausgewählten 24 Variablen eine weitere explorative Faktorenanalyse berechnet, die als Zwischenschritt dazu diente, die Faktorenspezifikationen der abschließenden konfirmatorischen Faktorenanalyse über die höchsten Faktorladungen der Variablen auf den jeweiligen Faktoren herzuleiten. Nachdem die Modellgütemaße bei der Analyse der 24 verbliebenen Variablen für ein Modell mit sechs Faktoren sprachen, wurde ein Generalfaktormodell mit fünf unkorrelierten Residualfaktoren spezifiziert, welches einen akzeptablen Gesamtfit zeigt (χ^2 = 784, df = 216, χ^2/df = 3.63; RMSEA = .089, CFI = 0.951; TLI = 0.938). Auf Basis dieses Modells wurden Faktor-Scores für den Generalfaktor geschätzt, diese werden als geschätzter Belastungswert der Schulen verwendet (deskriptive Statistiken des Generalfaktors siehe Tabelle 5).

Tab. 5: Deskriptive Statistiken des Generalfaktors für das finale Modell

Mittelwert	0.01
Standardabweichung	1.00
Minimum	-1.58
Maximum	2.40
Spannweite	3.98

Anmerkung: Mittelwert und Standardabweichung ergeben sich als Restriktionen des berechneten Modells

3.3 Zusammenhänge mit bisherigen Sozialindices

Zur Überprüfung der Validität der Belastungswerte und um eine Aussage über die Stabilität der sozialen Belastung treffen zu können, wurden punkt-biseriale Korrelationen zwischen den neu errechneten Belastungswerten sowie den vorherigen Belastungswerten berechnet. Zwischen den im Rahmen von KESS 4 errechneten Belastungswerten der Grundschulen (Bos et al., 2006) und den neu berechneten Belastungsscores ergab sich eine Korrelation von -.914, zwischen denen im Rahmen von KESS 7 errechneten Belastungswerten (Bos et al., 2009) für die weiterführenden Schulen ergab sich eine Korrelation von -.927.

4. Diskussion und Überlegungen zur Weiterentwicklung der Erfassung sozialer Belastung

Im Rahmen der Aktualisierung des Sozialindex für Hamburger Schulen wurden auf Grundlage einer konfirmatorischen Faktorenanalyse Belastungswerte für jede Schule berechnet. Sowohl die Güte des geschätzten Modells erscheint zufriedenstellend als auch die Validität der Messung, welche durch die Zusammenhänge zu vorherigen Berechnungen abgebildet werden kann. Grundsätzlich scheint das Konstrukt sozialer Belastung, wie aus den Korrelationen in Abschnitt 3.3 erkennbar ist, eher zeitstabil zu sein. Im Hinblick auf die methodische Weiterentwicklung des Sozialindex ergeben sich verschiedene Überlegungen, die im Folgenden dargestellt werden.

1. Bei der Berechnung des Sozialindex wurde und wird sich an einer rein normorientierten Interpretation (z.B. Goldhammer & Hartig, 2012) der Belastungswerte orientiert. Für die Bildung der sechs Belastungsgruppen wird eine äquidistante Einteilung der Belastungsskala in sechs Abschnitte vorgenommen, und Schulen derjenigen Belastungsgruppe zugewiesen, in deren Abschnitt ihr schulspezifischer Sozialindex fällt. Die Bildung der Abschnitte und damit die Definition der Gruppen erfolgt hierbei ausschließlich basierend auf der bei der Berechnung des Sozialindex gebildeten Skala und der Verteilung der geschätzten Werte. Naturgemäß hängt die auf diese Weise vorgenommene Zuweisung der Schulen zu Belastungsgruppen immer von der empirischen Stichprobenverteilung zum Zeitpunkt der Analyse ab, zudem kann sie sich in Abhängigkeit der verwendeten Variablen und Skalierungsmodelle unterscheiden. Diese Stichproben- und Methodenabhängigkeit kann angesichts der hohen praktischen Bedeutung des Sozialindex als problematisch betrachtet werden. Demnach stellt die Definition von Belastungsstufen nicht das Ergebnis, sondern den Ausgangspunkt der Messung sozialer Belastung dar. Darüber hinaus ließen sich in jedem Fall beliebige Anzahlen von Belastungsstufen definieren, unabhängig davon, wie eng oder breit die Verteilung der sozialen Belastung sich darstellt. Die Grenzen sozialer Belastung sind somit nicht stabil, eine längsschnittliche Entwicklung lässt sich nicht abbilden. Parallele Problemstellungen sind aus der Armutsforschung bekannt (Christoph, 2015). Im Falle der sozialindexbasierten Ressourcenallokationen, die dem Prinzip der Verteilungsgerechtigkeit folgen, nach dem Ressourcen nach definierten Notwendigkeiten an die Akteure distribuiert werden, lässt sich eine verteilungsbasierte Einteilung durchaus rechtfertigen. Nichtsdestotrotz wurde eine alternative Möglichkeit der Definition von Belastungsstufen im Rahmen von ersten Ansätzen überprüft.

   Eine Möglichkeit, die der verteilungsbasierten Definition von Belastungsstufen inhärenten Probleme abzuschwächen, besteht darin, statt der normorientierten Interpretation der schulspezifischen Belastung eine kriteriumsorientierte Interpretation, die verteilungsunabhängig ist, vorzunehmen. Mit diesem Ziel wurde daher der Versuch einer Schwellensetzung der Gruppeneinteilung über eine Standardsetzung unternommen, eine in diesem Forschungsfeld bisher nicht angewandte, sondern eher aus dem Feld der Kompetenzstufendefinitionen bekannte Methodik (Cizek & Bunch, 2007; Harsch, Pant & Köller, 2010). Durch eine interdisziplinäre

Herangehensweise sollte der charmante Gedanke umgesetzt werden, mit Schulforscherinnen und Schulforschern, Psychometrikerinnen und Psychometriker sowie Personen aus der Schulpraxis schulform- und bundeslandübergreifend zu definieren, was belastete Schulen von weniger belasteten Schulen unterscheidet.

Bei der Durchführung eines Experten-Workshops mit dem Ziel der Standardsetzung ergaben sich jedoch Schwierigkeiten, insbesondere bei der Beurteilung von Aspekten des sozialen Kapitals auf Schulebene. Während der durchschnittliche Anteil von Kindern mit Migrationshinweis auf Schulebene gut eingeschätzt werden kann, stellt sich dies bei der durchschnittlichen Beurteilung des sozialen Kapitals auf Schulebene weitaus problematischer dar: Ab welchem Mittelwert zwischen 1 und 4 ist eine Schule im Hinblick auf das soziale Kapital ihrer Schülerschaft also belastet?

Die Berechnung latenter Klassenanalysen sollte im Anschluss an den im Sinne der Verwertbarkeit gescheiterten Versuch des Experten-Workshops Hinweise darauf liefern, ob die durch die äquidistante Einteilung der intervallskalierten Skala hierarchische Stufung von Aspekten sozialer Belastung sinnvoll ist, oder ob eher Profile sozialer Belastung an Schulen vorliegen. Erstgenanntes ist der Fall, d.h., die hierarchische Einteilung von Belastung aufgrund der erhobenen Variablen erscheint die den Daten angemessene Methodik zu sein. Jedoch ergeben sich bei der Berechnung latenter Klassenanalysen sowohl bei den Hamburger Daten als auch bei vergleichbaren Berechnungen für die IGLU/TIMMS Stichprobe (Vennemann, Eickelmann & Wendt, 2014) nur drei bis vier Gruppen der sozialen Belastung von Schulen. Bei einer höheren Anzahl an Gruppen ist mit niedrigeren Klassifikaktionsgenauigkeiten zu rechnen. Da das Risiko eines im weiten Sinne Fehlers zweiter Art (falsch negativ) bei Klassifikationsungenauigkeiten gerade im Rahmen von sozialindexbasierten schulischen Ressourcenallokationen weitreichende Konsequenzen hat, sollte die Empfehlung lauten, zukünftig eine geringere Anzahl von Belastungsstufen zu definieren.

2. In methodischer Hinsicht wurde mit dem Einsatz von Faktorenanalysen der ursprüngliche, im Rahmen von KESS 4 etablierte Ansatz, wieder aufgegriffen und die Partial-Credit-Skalierung der vergangenen Jahre verworfen. Insbesondere die Integration von auf Aggregatebene anfallenden Daten in die auf Individualebene erhobenen Informationen bietet hier Möglichkeiten der ökonomischen Kostenreduktion als auch für den Umgang mit einem potenziellen Datenausfall im Rahmen der Befragungen. Gleichwohl scheint es sinnvoll zukünftig zu prüfen, in wie weit die Schätzung der Modelle akkurat erfolgt und ob die unterstellten theoretischen Annahmen nach Bourdieu auf Aggregatebene überhaupt zutreffen, da z.B. bekannt ist, dass Individual- und Aggregatebenen häufig nicht über eine strukturelle Äquivalenz verfügen (Van de Vijver & Poortinga, 2002) aber auch die Annahme der faktoriellen Invarianz der Faktorstruktur beim Einsatz einer klassischen Faktorenanalyse fehlerhaft sein kann (Kim, Kwok & Yoon, 2012). Zukünftig sollte daher geprüft werden, inwieweit der Einsatz von Mehrebenenfaktoranalysen sinnvoll ist und zu möglicherweise akkurateren Ergebnissen führt als eine einfach-klassische Faktorenanalyse.

3. Insbesondere im Rahmen der Fragebogenbefragung werden von Seiten der betroffenen schulischen Akteure teilweise Möglichkeiten der Manipulation von Fragebögen, mit dem Ziel höherer Ressourcenzuweisungen, vermutet. Aus diesem Grund erscheint das hohe Gewicht der Daten aus Schüler- und Elternfragebögen bei der Berechnung des Sozialindex in Zukunft nicht haltbar. Stattdessen muss überprüft werden, welche zusätzlichen Daten der amtlichen Statistik für die Berechnung des Sozialindex verwendet werden können (vgl. den Beitrag von Schräpler in diesem Band). Dies kann mögliche Verzerrungstendenzen durch selektive Teilnahmen oder Manipulationen verringern. Darüber hinaus erscheint ein solches Verfahren aus den o. g. legitimatorischen Gründen empfehlenswert. Denkbar sind folgende amtliche Daten, deren empirische Zusammenhänge zu sozialer Belastung und Validität im Vorfeld einer Verwendung überprüft werden müssten:
   - Daten zu Migrationshinweisen
   - Daten zur Nutzung des Bildungs- und Teilhabepakets
   - Daten zum sonderpädagogischen Förderbedarf im Rahmen der Förderung der mit Defiziten in den Bereichen Lernen, Emotionen und Sprache diagnostizierten Schülerinnen und Schüler
   - Daten zu erreichten Bildungsabschlüssen
   - Daten zur Anzahl von Gewaltvorfällen an Schulen
   - Daten zum Übergang in die Oberstufe

4. Auch die Prüfung der möglichen künftigen Berücksichtigung der Veränderung der Schülerschaft durch Flüchtlingszuzüge ist notwendig, da diese Veränderungen zum Teil erheblich sein können. Die Flüchtlinge, die Hamburg im schulpflichtigen Alter erreichen, werden zunächst in den Zentralen Erstaufnahmen beschult und gehen mit dem Übergang in die Folgeunterbringung in reguläre Schulen über. An den Schulen werden die Flüchtlinge in sog. Internationalen Vorbereitungsklassen beschult, bis sie über ausreichende Deutschkenntnisse verfügen, um dem normalen Unterricht folgen zu können. In den Jahrgängen 1 und 2 werden die Kinder direkt in den Regelunterricht integriert und erhalten zusätzliche Sprachförderung. D.h., bedingt durch die stark erhöhten Flüchtlingszuzüge zum Jahr 2015 (vgl. Tab. 6) wird erst im Verlauf des Jahres 2016 deutlich werden, welche einzelschulischen Veränderungen der Schülerschaft durch die Flüchtlingszuströme auftreten und wie bzw. ob diese Veränderungen der Ausgangslagen im Rahmen einer sozialindexbasierten Ressourcenzuweisung berücksichtigt werden können.

Tab. 6: Entwicklung der Zuwanderungszahlen von 2009–2015

Jahr	Schutzsuchende mit Verbleib in Hamburg
2009	770
2010	1.378
2011	1.546
2012	2.091
2013	3.619
2014	6.638
2015 (Jan – Sep)	13.179

Quelle: www.hamburg.de/fluechtlinge-daten-fakten

Bei allen methodischen Entscheidungen zur Erhebung und Berechnung des Hamburger Sozialindex ist zu beachten, dass die Ergebnisse und die Berechnungsweisen an betroffene Akteure (Schulen, Eltern, Politik) vermittelt werden müssen. Die durch die finanziellen Konsequenzen für die Schulen bedingte hohe öffentliche Aufmerksamkeit führt dazu, dass alle Teilschritte der Berechnung und Erhebung des Sozialindex in Hamburg thematisiert werden, z.B. im Rahmen parlamentarischer Anfragen.

Literatur

Bonsen, M., Bos, W., Gröhlich, C., Harney, B., Imhäuser, K., Makles, A., Schräpler, J.-P., Terpoorten, T., Weishaupt, H. & Wendt, H. (2010). *Zur Konstruktion von Sozialindizes. Ein Beitrag zur Analyse sozialräumlicher Benachteiligung von Schulen als Voraussetzung für qualitative Schulentwicklung.* Bildungsforschung, Bd. 31. Bonn: Bundesministerium für Bildung und Forschung.

Bos, W., Gröhlich, C. & Bonsen, M. (2009). Der Belastungsindex für die Schulen der Sekundarstufe I in Hamburg. In W. Bos, M. Bonsen & C. Gröhlich (Hrsg.), *KESS 7: Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (S. 87–93). Münster: Waxmann.

Bos, W., Pietsch, M., Gröhlich, C. & Janke, N. (2006). Ein Belastungsindex für Schulen als Grundlage der Ressourcenzuweisung am Beispiel von KESS 4. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G, Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung, Bd. 14, Daten, Beispiele und Perspektiven* (S. 149–159). Weinheim: Juventa.

Bourdieu, P. (1982). *Die feinen Unterschiede – Kritik der gesellschaftlichen Urteilskraft.* Frankfurt am Main: Suhrkamp.

Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In R. Kreckel (Hrsg.), *Soziale Ungleichheiten*, Sonderband 2 der Zeitschrift „Soziale Welt" (S. 183–198). Göttingen: Schwartz.

Bourdieu, P. (1992) (Hrsg.). Ökonomisches, soziales und kulturelles Kapital. In *Die verborgenen Mechanismen der Macht* (S. 49–75). Hamburg: VSA Verlag.

Bourdieu, P. & Passeron, J.-C. (1971). *Die Illusion der Chancengleichheit: Untersuchungen zur Soziologie des Bildungswesens am Beispiel Frankreichs.* Stuttgart: Klett.

Christoph, B. (2015). *Empirische Maße zur Erfassung von Armut und materiellen Lebensbedingungen – Ansätze und Konzepte im Überblick.* IAB-Discussion Paper, 25/2015. Nürnberg.

Cizek, G. J. & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications.

Coleman, J. S. (1988). Social capital in the creation of human capital. In *American Journal of Sociology, 94*(1), 95–120.

Erikson, R. & Goldthorpe, J. H. (2002). Intergenerational inequality: a sociological perspective. *Journal of Economic Perspectives, 16*(3), 31–44.

Erikson, R., Goldthorpe, J. H. & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *British Journal of Sociology, 30*, 341–415.

Freie und Hansestadt Hamburg (2012). *Sozialmonitoring Integrierte Stadtteilentwicklung – Bericht 2012.* Hamburg.

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Test- und Fragebogenkonstruktion* (2. Auflage, S. 173–201). Berlin: Springer.

Harsch, C., Pant, H. A. & Köller, O. (Hrsg.) (2010). *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany* (Vol. 2). Münster: Waxmann.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jennrich, R. I. & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika, 77*, 442–454.

Kim, E. S., Kwok, O.-M. & Yoon, M. (2012). Testing Factorial Invariance on Multilevel Data: A Monte Carlo Study. *Structural Equation Modeling, 19*(2), 250–267.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W. & Stanat, P. (Hrsg.) (2010). *PISA 2009. Bilanz nach einem Jahrzehnt.* Münster: Waxmann.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide.* Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Pietsch, M., Bonsen, M. & Bos, W. (2006). Ein Index sozialer Belastung als Grundlage für Rückmeldungen von Leistungsergebnissen an Schulen und Klassen und für ‚faire Vergleiche' von Grundschulen in Hamburg. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen.* Hamburger Schriften zur Qualität im Bildungswesen, Band 1 (S. 225–245). Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (2007) (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie.* Münster: Waxmann.

Schulte, K., Hartig, J. & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann & N. Maritzen (Hrsg.), *Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ).* Hanse – Hamburger Schriften zur Qualität im Bildungswesen, Band 13 (S. 67–80). Münster: Waxmann.

Tillmann, K. & Weishaupt, H. (2015). Ansätze bedarfsorientierter Ressourcenausstattung von sozial belasteten Schulen in Deutschland. Eine Situationsanalyse. *Zeitschrift für Bildungsverwaltung, 31*(2), 3–26.

Van de Vijver, F. J. R. & Poortinga, Y. H. (2002). Structural Equivalence in Multilevel Research. *Journal of Cross-Cultural Psychology, 33*(2), 141–156.

Vennemann, M., Eickelmann, B. & Wendt, H. (2014). Heterogenität auf Schulebene? Kompositionsmuster an Grundschulen in Deutschland und ihr Zusammenspiel mit schulischen Ressourcen. In K. Drossel, R. Strietholt & W. Bos (Hrsg), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (S. 65–86). Münster: Waxmann.

Belgeo
Revue belge de géographie

2-3 | 2017
Une géographie sociale de l'enseignement/A social geography of education

Bordering the area of spatial relevance for schools: a stochastic network approach using the example of Hamburg, Germany

Eingrenzung von Räumen schulischen Marktgeschehens: Stochastische Netzwerkmodellierung anhand des Beispiels Hamburg

Sebastian Leist and Marcus Pietsch

Electronic version
URL: http://journals.openedition.org/belgeo/20332
DOI: 10.4000/belgeo.20332
ISSN: 2294-9135

Publisher:
National Committee of Geography of Belgium, Société Royale Belge de Géographie

Electronic reference
Sebastian Leist and Marcus Pietsch, « Bordering the area of spatial relevance for schools: a stochastic network approach using the example of Hamburg, Germany », *Belgeo* [Online], 2-3 | 2017, Online since 30 June 2017, connection on 15 April 2018. URL : http://journals.openedition.org/belgeo/20332 ; DOI : 10.4000/belgeo.20332

Bordering the area of spatial relevance for schools: a stochastic network approach using the example of Hamburg, Germany

Eingrenzung von Räumen schulischen Marktgeschehens: Stochastische Netzwerkmodellierung anhand des Beispiels Hamburg

Sebastian Leist and Marcus Pietsch

Introduction

1 In Germany, the assignment of students to schools was for a long period of time determined by rigid governmental specifications and the clash of interests between parents and schools (Weiß and Steinert, 1996). In those days a more or less strong assertiveness of the protagonists was necessary (Gomolla and Radtke, 2002). Since the turn of the millennium, a sequential movement from input-steering to output-steering was established. This is considered to involve the establishment of quasi-markets. On these quasi-markets, the schools have to perform under the conditions of competition while funding, control and supervision are assigned to the public authority (Bellmann, 2007; Weiß, 2001).

2 Quasi-markets in education are, according to Weiß (2001) characterised by free choice of school, funding in relation to performance, autonomy of the schools and internal possibilities for further development of the human resources and organisational structures. Thereby, the economic regulation system needs mechanisms which increase transparency on the market by providing information which enables the participants of the market to compare the institutions, e.g. results of evaluations or systems of quality management. Recent reforms in education aim both at more efficiency and an as

transparent as possible use of the resources. Meanwhile, student achievement is aimed to increase.

3 Two premises indicate a perfect schooling-market: competition between the suppliers and freedom of choice of the consumers. Gewirtz, Ball and Bowe (1995, p. 2) focus in this context the following:

> "The education market (like all other markets) is intended to be driven by self-interest: first, the self-interest of parents, as consumers, choosing schools that will provide the maximum advantage to their children; second, the self-interest of schools or their senior managers, as producers, in making policy decisions that are based upon ensuring that their institutions thrive, or at least will survive, in the marketplace. The demand for school places is inelastic; that is the number of potential students is fixed. Where there surplus places, the result is meant to be competition, emulation and rivalry: survival can only be assured by attracting consumers away from other schools."

4 The idea of competition between schools based on market structures is comparatively simple: If students and their families can choose from a range of schools, then schools can't take their clientele as granted and must ensure to improve output and achievement to meet the preferences of students and their parents and to persist on the market. Enhanced choice options for students and parents thus still contain the promise of a quality improvement in education by assuming a positive causal relationship of choice options, competition and quality in education: choice options for students and their families generate competition between schools and competition between schools generates quality in education.

5 Nonetheless, Altrichter and Rürup (2010, p. 143) notice in their summary of the discourse on school autonomy an increasing awareness about potentially undesired side-effects of the realised configurations of school autonomy. It seems as if occurs, besides the desired differentiation and pluralism in education, the introduction of a hierarchy with the formation of residue classes and schools. Accordingly, they postulate to investigate the criticism that autonomy promotes the disintegration of an educational system more consequent than before (Altrichter and Rürup, 2010, p. 143).

6 The following part concludes the current state of research about social segregation on educational markets and turns afterwards towards the difficulty of an adequate bordering of educational markets.

Traditional bordering of schooling-markets

7 The present literature about the quantification of effects of competition on public-sector markets mainly deals with two approaches: Either the competitiveness of a market is defined by an index of market concentration or, based on theories about spatial competition, defined by the number of suppliers which are accessible within a given travel time, distance or within given travel costs (Hotelling, 1929). The method used in this article refers to the first approach. Thus, the situation of rivalry within a market is going to be captured by the Herfindahl-Index, which quantifies the supplier concentration on a specific market (Belfield and Levin, 2002). To avoid arbitrary results, it is very important to tackle the difficulty of market definition, i.e. to properly border the area of a market. Mostly, the spatial extent of a market is ambiguous. The scientific literature therefore uses auxiliary approaches which means that market borders are

assumed to be along areas of responsibility of municipal authorities (Bradley and Crouchley; Millington and Taylor, 2000) or even take whole agglomerations as a base (Hoxby, 2000). The validity of analysis based on these approaches anyhow is in dispute; critics mention that at least the consideration of spatial barriers, e.g. rivers, motorways or railways, to classify markets is a reasonable strategy (Rothstein, 2007; Hoxby, 2007). Nevertheless markets bordered via spatial barriers may remain with a too big extension, i.e. that a low market concentration in these cases may not be equal to more choice and competition. The reason is that not all suppliers (= schools) are equally accessible for all customers (= students), because the accessibility of schools, for example by public transport, is likely to vary.

8 The disadvantages of these auxiliary approaches may be resolved if an endogenous criterion, which is part of the available data to localise the markets, would be taken into consideration, replacing above mentioned exogenous criteria (area of responsibility etc.). Accessibility of the suppliers is assured and the danger of arbitrary values of indicators is avoided.

Data

9 The performed analysis to border the schooling markets is based upon data of transitions between primary and secondary schools after year 4. Primary school in Hamburg finishes after year 4, so every student who completes year 4 leaves towards a secondary school. In considering all transitions made by students as paths (or as lines on a map, to have a visual approach) between primary and secondary schools, a network might be spanned that covers Hamburg. The utilized dataset is generated from the Individual Student Database of the Educational Department of the Federal State of Hamburg which contains all students that attend a school within the jurisdiction, no matter which school type, school maintaining body, year level or place of residence (Freie und Hansestadt Hamburg, 2012a). It contains students who were in year 5 during school year 2011/12 and were the year before in grade 4 and who swapped schools within the federal state of Hamburg. The dataset represents 400 schools and 14,032 students who made transitions between primary and secondary schools on 2,446 different paths (or lines on a map, to stay with the notion as introduced above), often by more than one student. A transition is defined either as a switchover to a different school, or the continuance in a school if year 5 is attended, which is possible in a small number of the comprehensive schools.

10 The described data is regarded as “relational” in the sense that schools are interconnected via the transitions of students (paths, resp. lines on a map) and is transformed to a social network by the use of various techniques (Butts, 2008). In technical terms: The social network is directed, because students are only permitted to swap from primary school to secondary school and not the opposite direction, and the network contains loops which means that a student may remain on the same school if she/he stays and attends year 5.

Stochastic modelling of schooling markets

11 In opposition to the traditional approaches to border schooling markets, the relational data at hand enables access to an endogenous criterion to define the extension of a

market. Structures of close connection and agile exchange become recognised as cluster (= market) and generate the chance to quantify spatial-temporal phenomena like competition or social segregation. The algorithm projects all schools in a so-called latent social space. Hence, Primary schools are located close together, if they "serve" similar Secondary Schools, and the Primary schools are located close to the Secondary Schools they serve. Considering the perspective of Secondary schools in turn, they are placed in close neighbourhood in the latent social space, if they receive their students in Year 5 from similar Primary Schools. Again, the algorithm assigns coordinates to these Secondary schools which locate them close to these Primary Schools. This is how the assignment to positions in the latent social space takes place for all schools. The cluster procedure subsequently assigns the schools to clusters in considering gaps in the distribution in the latent social space.

12 The stochastic modelling of school networks (= educational markets) was carried out by the package "latentnet" for the open-source software "R" (Krivitsky and Handcock, 2008; Krivitsky and Handcock, 2014). Latentnet evaluates "latent position and cluster models for statistical networks" according to Hoff, Raftery and Handcock (2002) and Handcock, Raftery and Tantrum (2007). These are extensions of Generalized Bilinear Mixed-Effects Models (GBME) by a Finite Mixture Model to reveal group structures. Finite Mixture Models are stochastic models which specify the likelihood of observed data as a function of multiple groups (Templin, 2008, p. 325). The probability of a dyad is expressed via a function of distances between two vertices in a latent space as well as with functions of observed dyadic covariates. The probability of a network g for a set of nodes is a product of dyad probabilities. Each is a Generalized Linear Model with following linear component:

$$\eta_{i,j} = \sum_{k=1}^{p} \beta_k X_{i,j,k} + d(Z_i, Z_j) + \delta_i + \gamma_j$$

$X_{i,j,k}$ is an Array of dyadic covariates, β_K is a vector of coefficients of the covariates. Z_i and Z_j are the positions of the vertices i and j in the latent space. d is a function of 2 positions; either negative Euclidian $(-\|Z_i - Z_j\|)$ or bilinear $(Z_i * Z_j)$. δ_i and γ_j are vectors of sender- and receiver-effects (Krivitsky; Handcock, 2014).

13 The stochastic modelling was conducted in two steps. At first, latent clusters were modelled for the entire network. Therefore, all transitions had to be integrated unweighted into the model, i.e. the number of students actually executing each of the transitions between all Primary and Secondary Schools had to remain unconsidered and only the presence or absence of a connection (= path or line on a map) was considered. The reason for this proceeding is the weak density of the entire network[1]. The magnitude of a connection was integrated into the model in the subsequent step. It comprises the subdivision of the regional clusters, identified in the first step of the analysis, into local Subcluster. The density of the regional networks is sufficient enough to model the local networks considered as weighted.

14 The proceeding for the calculations for both spatial levels was to set up an increasing number of groups, starting with one group. The procedure is similar to latent class analysis (Lazarsfeld and Henry, 1968). The number of two latent dimensions to model latent cluster is adequate for descriptive purposes according to Hoff (2005). The Markov Chain Monte Carlo runs contained a burnin of 10,000 iterations which were discarded and 40,000 iterations of which one in ten was used for the modelling (Raftery and Lewis, 1995). The Bayes Information Criterion was considered for model selection whose lowest value

indicates which model fits best to the data and therefore should be chosen (Schwarz, 1976).

Results

15 The elucidated procedure unveiled regional and, in a subsequent step, local structures of the schooling landscape in the federal state of Hamburg.

Regional schooling markets

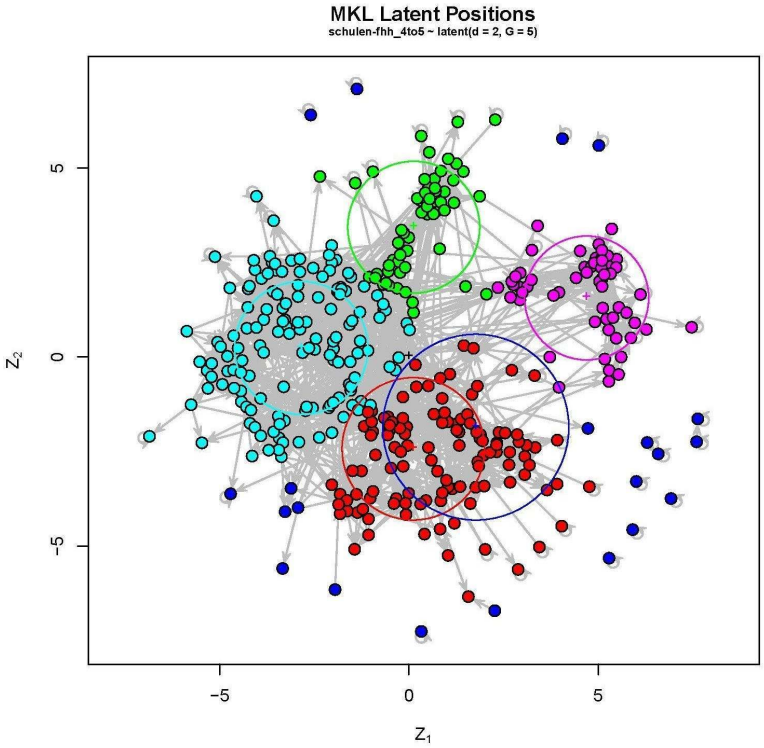
16 The stochastic modelling of the entire and, due to the fact of low density, unweighted network resulted in multiple cluster solutions of which the one with the lowest BIC-value was selected.

Table 1. Model selection according to BIC-value for the regional schooling markets.

Cluster-solution	Overall BIC
1	25861,45
2	25869,74
3	26271,3
4	26280,89
5	25771,32
6	25842,18
7	25900,31

17 Table 1 suggests that the solution which fits best to the data reveals five regional schooling markets in Hamburg. These markets are subdivided into local markets in the following subsequent step. The distribution of the schools in the latent space is visualized in figure 1.

Figure 1. Distribution of schools in the latent space and assignment to regional schooling markets.



18 Both axes represent the latent dimensions. Each circle represents a school. An arrow between two schools indicates a transition of one or more students. An arrow which ends at the same school indicates that one or more students continue attending the same school (in technical terms: loop). Schools which share the same color belong to the same group. Within these groups centroids are placed, which are surrounded by circles in the same color. Their size represents the magnitude of the distribution of each group in latent space. Next to central and close to each other placed schools there are clear gaps between school groups. Furthermore, in the periphery, schools are placed which lack connection to the network. All these schools are united in one outlier-group.

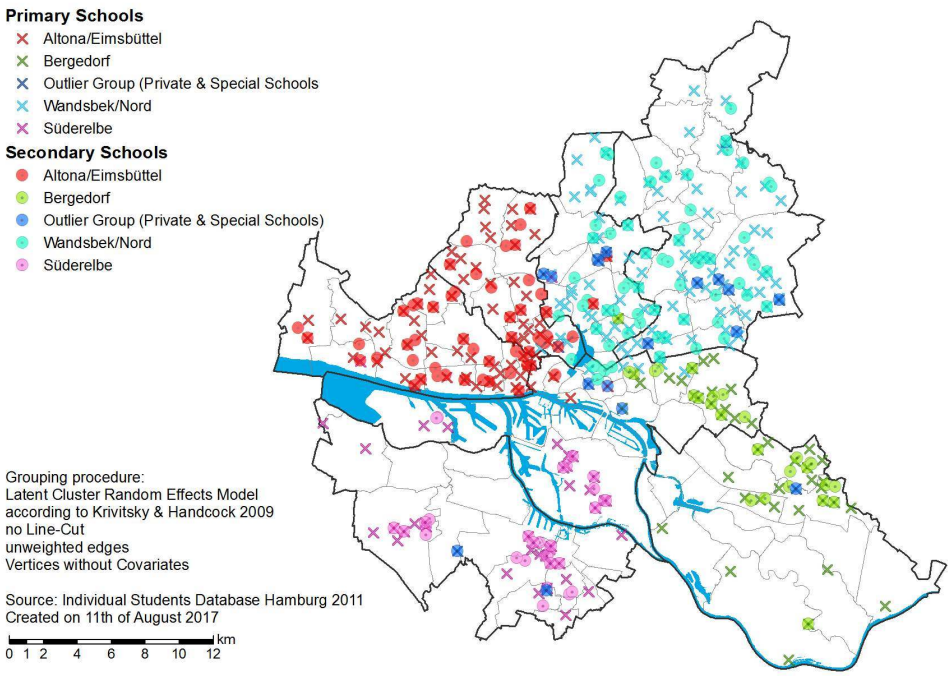
Table 2. Composition of the regional schooling markets.

	primary schools	comprehensive schools	grammar schools	special schools	sum
Facilities in					
Cluster 1 "Altona/ Eimsbüttel"	67	20	22	10	119
therefrom public	*61*	*14*	*19*	*8*	*102*
Cluster 2 "Bergedorf"	33	12	8	7	60
therefrom public	*31*	*9*	*7*	*7*	*54*

Cluster 3 "Outlier-Group"	3	4	2	12	21
therefrom public	*0*	*0*	*0*	*10*	*10*
Cluster 4 "Wandsbek/ Nord"	82	26	30	5	143
therefrom public	*74*	*21*	*26*	*5*	*126*
Cluster 5 "Süderelbe"	29	12	8	8	57
therefrom public	*27*	*9*	*7*	*7*	*50*

19 This group (cluster 3) consists basically of small private schools and special schools as table 2 shows. The column which contains the sums shows that the group sizes vary. The largest group consists of 143 schools and all groups contain public schools as well as private schools.

Figure 2. Localization of the regional schooling markets in the administrative area of Hamburg, grouped by transitions from primary to secondary schools in Summer 2011.



20 Figure 2 points out the spatial positions of the five groups in Hamburg. For the bordering of these groups, spatial barriers apparently are most important. The pink group is clearly bordered by the Elbe River and the harbor. Most of the border between the bright blue group and the green group is a motorway which is difficult to cross as well. The red group and the bright blue group are partly divided by the airport in the north. The group represented by the dark blue color is the outlier-group, as mentioned above.

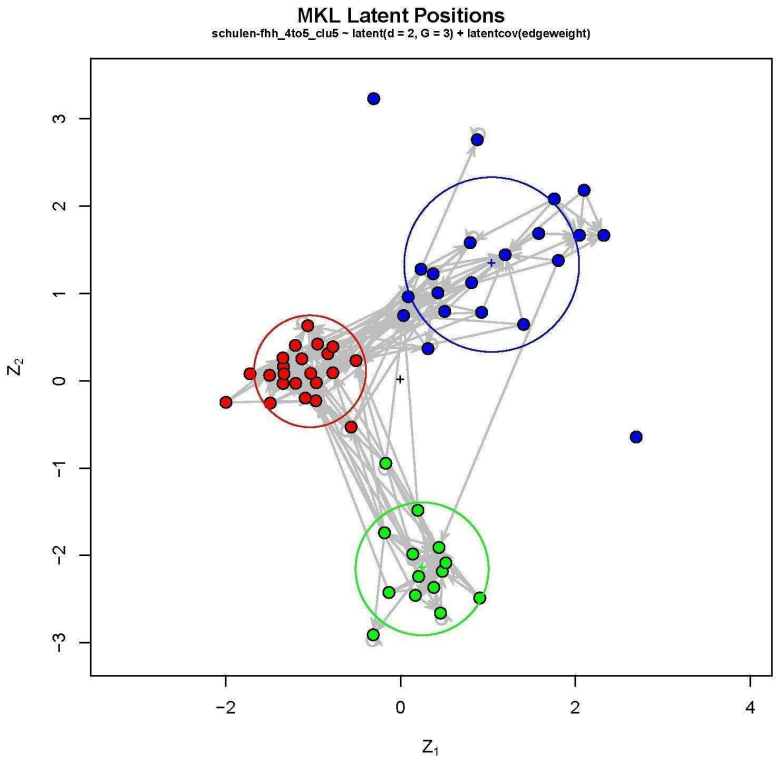
Local schooling markets

21 Until now it is not possible to unveil multilevel latent group structures in social networks. This is the reason why the subsequent step is necessary to sub-divide the regional schooling markets. Within the regional schooling markets the connectivity is adequate enough to model the local sub-groups by considering the number of students who swap from primary to secondary school. These function as weights for the transitions, giving transitions of severe students more importance than transitions of only one student. The model is exemplified within one regional cluster which is localized in the south-west of the administrative boundaries of Hamburg and indicated by the pink color (Figure 3). The rationale behind the choice of this particular cluster is firstly, that it appears to be persistent over time due to the clear spatial borders that divide this cluster from the other parts of the city. Borders of clusters within densely populated areas could vary or even disappear due to the transitions made in one particular year, but not taking place the year after. Additionally, the cluster combines sparsely populated, rural parts of Hamburg with densely populated urban parts as well as wealthy neighborhoods whose inhabitants live mostly in single houses with deprived areas characterized by huge multi-storey social housing building clusters. The settlements in this cluster stretch along a railway line that runs south from the city centre and turns westwards after crossing the Elbe river. A highway runs parallel. These are the main transport axes in this part of the city. This cluster seems to be appropriate to show the benefits of the approach at hand.

Table 3. model selection according to BIC-value for the local schooling markets.

Cluster-solution	Overall BIC
1	843,25
2	848,80
3	789,54
4	790,68
5	792,02
6	794,73

Figure 3. Distribution of schools in the latent space and assignment to local schooling markets.



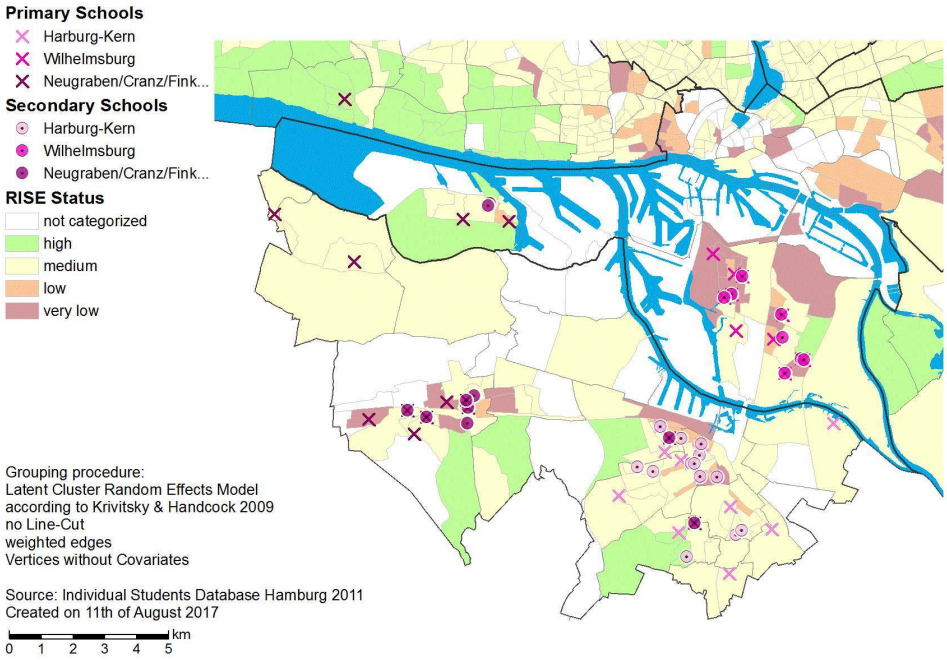
22 The distribution of the three sub-groups in the latent space is visualized in Figure 3. The schools are divided in three more-or-less clearly differentiated groups. Additionally, two schools belong to the clusters which don't possess connectivity to the other schools. They are positioned in the periphery[2]. In contrast, the schools in the red cluster are positioned close together, i.e. they seem to have a considerable amount of transitions within the group.

Table 4. Composition of the local schooling markets.

	primary schools	comprehensive schools	grammar schools	special schools	sum
Facilities in					
Sub-cluster "Harburg-Kern"	11	4	5	2	22
therefrom public	*11*	*3*	*4*	*2*	*20*
Sub-cluster "Wilhelmsburg"	6	4	1	3	14
therefrom public	*6*	*3*	*1*	*3*	*13*
Sub-cluster "Neugraben/ Cranz/Finkenwerder"	12	4	2	3	21
therefrom public	*10*	*3*	*2*	*2*	*17*

23 Table 4 provides a summary of the local sub-clusters which were revealed within the regional schooling market in the south-western part of Hamburg and shows the amount of schools according to school type and school maintaining body. Sub-cluster 1 and 3 contain a similar amount of schools while sub-cluster 2 contains one third less schools. In all sub-clusters approximately half of the schools are primary schools. Comprehensive schools are spread equally throughout the three sub-clusters as well as the special schools. Only grammar schools are bunching in sub-cluster 1 which contains five out of eight grammar schools in the region.

Figure 4. Local schooling markets in the south-western part of Hamburg, grouped by transitions from primary to secondary school in Summer 2011.



24 The localization of the schools belonging to the sub-clusters is shown in Figure 4. The primary school in the north, across the Elbe River, is one of the mentioned schools without connectivity. As already pointed out in relation to the formation of the regional clusters, the local clusters are obviously bordered along spatial barriers as well. The islands Veddel and Wilhelmsburg in the Elbe River form one sub-cluster. In the south, the local cluster Harburg-Kern is localized and encompasses the core settlement in this area. To the west there is a spatial gap which marks the border to the sub-cluster Neugraben/ Cranz/Finkenwerder, which in turn covers some rural areas. Additionally, the RISE-Index is shown as signature for the areas in which the students live. Brown indicates areas where the risk that a student lives in an environment of multiple discriminatory factors (unemployment, living on welfare, low educational degrees, high rate of migrants, and high rate of single parents) is highest. Orange indicates similar areas where the risk is lower. A significant number of brown and orange areas are on the islands Veddel and Wilhelmsburg and in Neugraben/Cranz/Finkenwerder. Yellow colored areas have medium risk to live in an environment of multiple discriminatory factors. These areas are common and well-spread throughout the considered part of Hamburg. Green colors indicate low risk to live in a deprived environment. The population of these areas is

wealthy by a high chance. Each of the local schooling markets is surrounded by at least one of these areas. It is important here to point out that these areas are sparsely populated because in most cases these areas are covered with single family detached houses and residences, i.e. that not many students live in these areas. By contrast, the brown and orange areas often contain multi-storey buildings in densely populated hotspots with accumulated social disadvantages (Freie und Hansestadt Hamburg, 2011). By far more students live in these areas.

Application of schooling markets on competition: an example

25 At first, it will be pointed out, how the exact definition of the schooling markets outperforms a wide-spread traditional, auxiliary approach of market definition. Afterwards, it is shown, how competition may be differentiated depending on the social status of the students.

Herfindahl-index

26 The turn towards more autonomy for schools fostered processes of profiling of schools which result in increasing competition between schools (Altrichter and Rürup, 2010). A commonly used measure for competition is the Herfindahl-Index which is applied to the competition of schools for students:

$$HI = \sum_{i=1}^{N} s_i^2$$

The Herfindahl-Index quantifies competition based on supplier-concentration (Belfield; Levin, 2002). s_i corresponds to the market share of supplier i on the market and N represents the total number of suppliers. The index may have values between 1/N and 1 and is interpreted in accordance with the following guidelines:

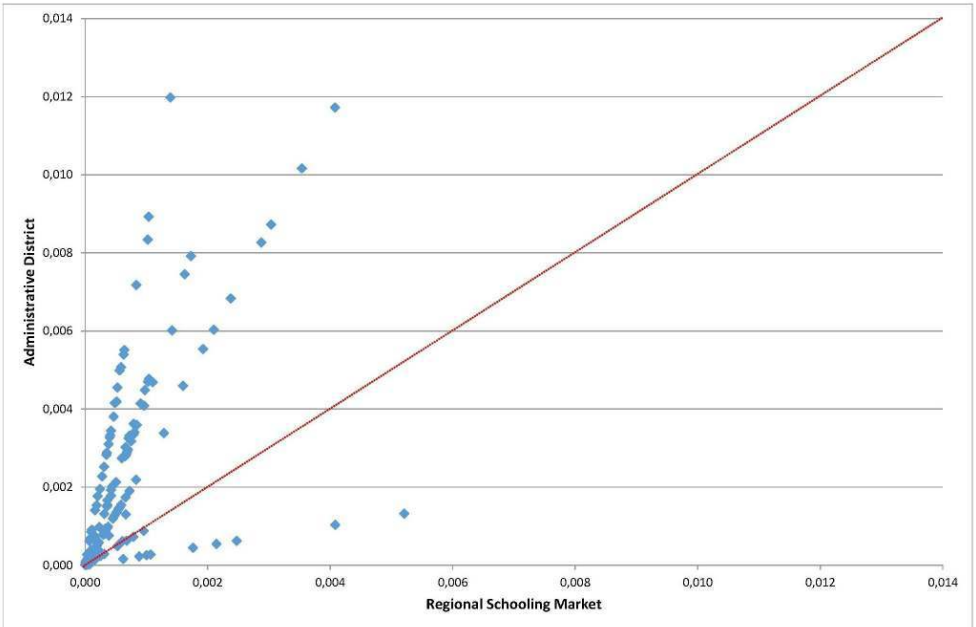
- Values below 0.15 indicate a low market concentration, i.e. a perfect market,
- Values between 0.15 and 0.25 indicate a moderate concentration and indicate an oligopoly,
- Values above 0.25 indicate a strong market concentration, i.e. a monopoly (U.S. Department of Justice and the Federal Trade Commission, 2010).

Regional schooling markets vs. administrative districts

27 As mentioned above, traditional approaches are usually used to define the extensions of a market. This might be done via a given travel time, distance or by given travel costs (Hotelling, 1929) and leads to individual markets for each particular school. Another often applied approach is to border markets along areas of responsibility of municipal authorities (Bradley; Crouchley; Millington; Taylor, 2000) or even take whole agglomerations as a basis (Hoxby, 2000). The Federal State of Hamburg consists of seven jurisdictions (“Bezirk”) which represent the communal, i.e. local level within the federal hierarchy in Germany. The responsibilities within the School Supervisory Board of Hamburg are spatially shaped along the borders of these jurisdictions. Therefore, competition between individual schools is compared by looking at the values on the market in the sense of an administrative district and by looking at the values for

competition on the Regional Schooling markets (except the outlier-group). Figure 5 shows, how competition between schools is different, depending on the approach to border the market.

Figure 5. Comparison of market shares of schools between administrative districts and regional schooling markets.



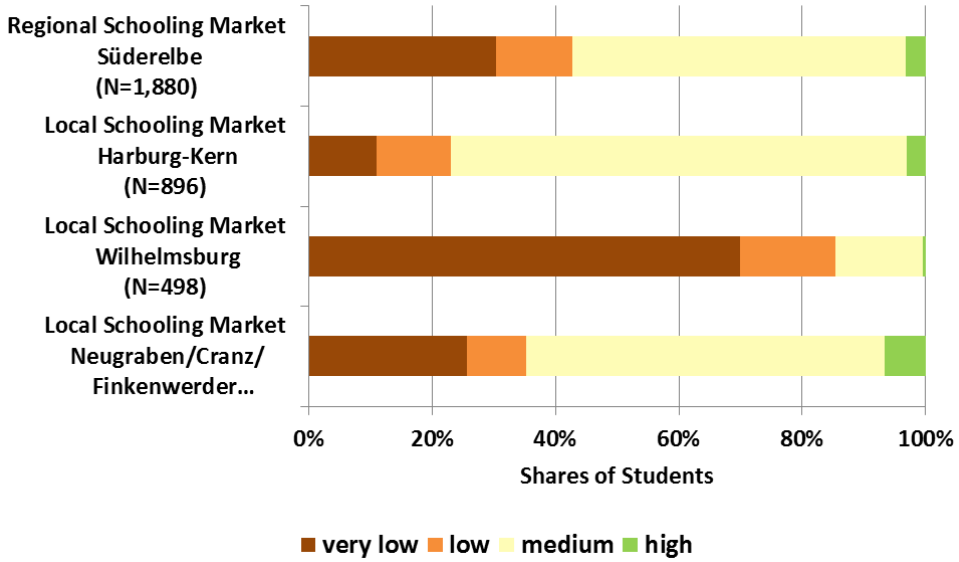
28 The red diagonal line indicates perfect correlation of the values for the Administrative districts and the Regional Schooling Markets. Only a handful of schools is located on this line. All the other rhombi indicate that the use of the Administrative boundaries induces incorrect values for competition which are mostly overestimated. The correlation (Pearson's *r*) between the values is 0.732. Summarized, only the use of as accurate as possible market borders guarantees valid results, otherwise the values depend highly on the fit between administrative boundaries and the unknown market boundaries in real. The risk of arbitrary results then is immense.

Differentiated competition on local schooling markets

29 For the reason of quantifying competition, the categorisation of students according to the RISE-Status of the place of residence is made. The categories of the RISE-Status are "very low", "low", "medium" and "high".[3] The presumption is, that students who reside in wealthy areas are more likely to have parents with high educational attainment and may be regarded as probably high-performing whereas students out of areas characterised by multiple discriminatory factors are considered to be probably low-performing (Kuhl, Siegle and Lenski, 2013, p. 275ff). A high share of high-performing students is regarded to be an advantage in competition between schools, because this might be seen by the parents, amongst others, as an obvious and easy-to-understand indicator for school quality compared to more abstract indicators like "system-performance" or "gains in student achievement" (Altrichter and Rürup, 2010, p. 140). This suggests that schools strive especially to students from high status residential areas and hence competition takes place preferentially to these students. Figure 6 shows the composition of the

students within the south-west region of Hamburg and its local schooling markets according to the RISE-Index of the students of grade 5 at their place of residence.

Figure 6. Student composition on the regional schooling market Süderelbe and its local schooling markets according to RISE-status (year 5).



30 Figure 6 shows discrepancies within the region. Whereas the composition of the local schooling market "Neugraben/Cranz/Finkenwerder" is more or less similar to the one of the whole region, the two remaining markets differ from that. The local schooling market "Harburg-Kern" has a heterogeneous composition and a broad share of medium classified residential areas which indicates a less problematic situation. The situation on the schooling market "Wilhelmsburg" then again is totally different. The students are comparatively homogeneous and 70 per cent reside in areas which are characterised by a high risk of discriminative factors. An additional 15 per cent lives in areas of a bit fewer risks of discriminative factors. In other words, 17 out of 20 students live in deprived circumstances by a high chance. Only two out of all students in year 5 reside in a wealthy area. In relation to the evidence of a high correlation between student achievement and social origin, it can be stated that this local schooling market is cut off from high achieving students in Hamburg.

Table 5. Competition on the regional schooling market and its local schooling markets.

				Herfindahl-Index			
	N (No. of secondary schools)	*Hi_{min} (1/n)*	*all students*	*only RISE-Index "very low"*	*only RISE-Index "low"*	*only RISE-Index "medium"*	*only RISE-Index "high"*
Regional Schooling Market							

Süderelbe	30	0,03	0,05	0,08	0,07	0,06	0,12
Local Schooling Markets							
Harburg-Kern	13	0,08	0,10	0,12	0,11	0,11	0,28
Wilhelmsburg	8	0,13	0,17	0,17	0,31	0,21	0,50
Neugraben/ Cranz/ Finkenwerder	9	0,11	0,17	0,25	0,20	0,17	0,24

31 Table 5 points out the competition both globally for the whole south-west region of Hamburg and the three local markets. The whole region suggests to be a perfect market (HI< 0.15). A closer look at the local schooling markets shows the situation more differentiated. The schooling market "Harburg-Kern" may be regarded as a market with low concentration of suppliers, but looking at the student groups according to RISE-Index, there is a remarkable high concentration of students in some schools who live in wealthy areas (HI = 0.28). The local schooling market "Wilhelmsburg", where probably the highest achieving students are absent, has a high concentration of suppliers, especially by considering the social composition of the students. The Herfindahl-Index has values above 0.15, i.e. a moderate or even high social concentration of the students is a characteristic of this market. The HI of 0.5 for the students residing in the wealthy areas has to be interpreted with care due to the fact that only two students in year 5 live in these areas. The local schooling market "Neugraben/Fischbek/Finkenwerder" is again a remarkably concentrated market (HI = 0.17). The social differentiation shows tendencies of concentration, especially for the students living in wealthy areas as well as in areas of high risk of cumulative factors of deprivation.

Conclusion, prospects and discussion

32 It is possible to tackle the difficulties of bordering regional and local structures. In case of stochastic network approaches, it is a requirement to possess relational data. Then social network analysis tools are able to unveil latent structures and clusters of strong relationships between the actors. In this example, the schooling landscape of Hamburg can be divided into regional and local schooling markets. Mostly, the borders are along spatial gaps in settlement structures, linear barriers (rivers, motorways, railways) or point-shaped barriers (lakes, airport). Hence, if no relative data is available, the use of spatial barriers is a reasonable approach to border markets. Nonetheless, difficulties in bordering persist in areas without spatial barriers. The presented stochastic network approach then shows one of its strengths and still assigns the schools to clusters. The comparison of a traditional approach and the presented approach shows the evidence that the calculated values for socio-spatial phenomena like competition vary between both approaches and shows the risk of arbitrary results. On a small spatial scale, the sub-division of the Regional Schooling Markets into Local Schooling Markets creates options to understand small-scale processes of competition and social segregation. Therefore, the

further analysis of qualitative and quantitative data content is a way to understand the processes which explain the reasons for the attractiveness of a certain school for certain social groups. Further research should also take the questions of persistence of the Regional and Local Markets over time into consideration. Markets may collapse or show continuity over time.

BIBLIOGRAPHY

ALLEN R., VIGNOLES A. (2007), "What Should an Index of School Segregation Measure?", *Oxford Review of Education*, *33,* 5, Oxford, Taylor & Francis, pp. 643-668.

ALTRICHTER H., RÜRUP M. (2010), "Schulautonomie und die Folgen" ["School autonomy and its consequences"], *in* ALTRICHTER H., MAAG MERKI K. (eds.), *Handbuch Neue Steuerung im Schulsystem,* Wiesbaden, VS Verlag für Sozialwissenschaften, pp. 111-142.

BELFIELD C., LEVIN H. (2002), "The Effects of Competition between Schools on Educational Outcomes: A Review for the United States", *Review of Educational Research, 72,* Thousand Oaks, SAGE Publications, pp. 279-341.

BELLMANN J. (2007), "Das Monopol des Marktes. Wettbewerbssteuerung im Schulsystem" [The monopoly of the market. Competition control in the school system], *Berliner Debatte Initial, 18*, Berlin, WeltTrends, pp. 58-71.

BELLMANN J., WEIß M. (2009), "Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. Theoretische Konzeptualisierung und Erklärungsmodelle" [Risks and side effects of the New control in the school system. Theoretical conceptualization and explanatory models], *Zeitschrift für Pädagogik , 55,* 2, Weinheim, Beltz, pp. 286-308.

BRADLEY S., TAYLOR J. (2002), "The Effect of the Quasi-Market on the Efficiency- Equity Trade-Off in the Secondary School Sector", *Bulletin of Economic Research, 54,* 3, Chichester, Wiley-Blackwell, pp. 295-314.

BRADLEY S., CROUCHLEY R., MILLINGTON J. & TAYLOR J. (2000), "Testing for Quasi-Market Forces in Secondary Education", *Oxford Bulletin of Economics and Statistics 62*, Oxford, Wiley-Blackwell, pp. 357-390.

BUCKLEY J., SCHNEIDER M. (2003), "Shopping for Schools: How do Marginal Consumers gather Information about Schools?", *The Policy Studies Journal, 31, 2,* Chichester, Wiley-Blackwell, pp. 121-145.

BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (2010), *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten [The transition from primary to secondary school. Power justice and regional, social and ethnic-cultural disparities],* Bildungsforschung Band 34, Berlin.

BUTLER T., HAMNETT C. (2007), "The Geography of Education: Introduction", *Urban Studies 44,* 7, Thousand Oaks, SAGE Publications, pp. 1161-1174.

BUTTS C. (2008), "Social Network Analysis with sna", *Journal of Statistical Software, 24,* 6, Foundation for Open Access Statistics, pp. 1-51.

DUNCAN O., DUNCAN B. (1955), "A methodical analysis of segregation indexes", *American Sociological Review, 20*, Thousand Oaks, SAGE Publications, pp. 210-217.

FREIE UND HANSESTADT HAMBURG (2010), *Pilotbericht "Sozialmonitoring im Rahmenprogramm Integrierte Stadtteilentwicklung" (RISE) [Pilot report "Social Monitoring Programme in the Framework of Integrated Urban District Development" (RISE)],* Hamburg. http://www.hamburg.de/contentblob/2673088/data/pilotbericht-rise.pdf, accessed 22. Sep. 2015)

FREIE UND HANSESTADT HAMBURG (2012a), *Schulstatistiken der Freien und Hansestadt Hamburg [School statistics of the Free and Hanseatic City of Hamburg],* Hamburg. http://www.hamburg.de/schulstatistiken, accessed 18 September 2015.

FREIE UND HANSESTADT HAMBURG (2012b), *Sozialmonitoring Integrierte Stadtteilentwicklung. Bericht 2011 [Social Monitoring Programme in the Framework of Integrated Urban District Development. Annual Report 2011],* Hamburg. http://www.hamburg.de/contentblob/3335496/data/d-sozialmonitoring-bericht-2011.pdf, accessed 22 September 2015.

FRIEDMAN M. (1955), "The Role of Government in Education", *in* SOLO R.A., (eds.), *Economics and the Public Interest,* New Brunswick, Rutgers University Press, pp. 123-144.

GEWIRTZ S., BALL S. & BOWE R. (1995), *Markets, Choice and Equity in Education*, Buckingham, Open University Press.

GIBBONS S., SILVA O. (2006), "Competition and accessibility in school markets: empirical analysis using boundary discontinuities", *in* GRONBERG T., JANSEN D. (eds.), *Improving school accountability: check-ups or choice,* Emerald Group Publishing Limited, pp. 157-187.

GOMOLLA M., RADTKE F. (2002), *Institutionelle Diskriminierung: Die Herstellung ethnischer Differenz in der Schule [Institutional discrimination: The preparation of ethnic difference in school],* Wiesbaden, Leske + Budrich.

GORARD S., TAYLOR C. (2002), "What is segregation? A comparison of measures in terms of 'strong' and 'weak' compositional invariance", *Sociology, 36,* 4, Thousand Oaks, SAGE publications, pp. 875-895.

GORARD S., TAYLOR C. & FITZ J. (2003), *Schools, Markets and Choice Policies*, London, Routledge Falmer.

HANDCOCK M., RAFTERY A. & TANTRUM J. (2007), "Model-based clustering for social networks", *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170,* 2, London, Wiley-Blackwell, pp. 301-354.

HANNEMAN R., RIDDLE M. (2005), *Introduction to social network methods*, Riverside, on-line textbook of the Department of sociology at the University of California. http://faculty.ucr.edu/~hanneman/nettext/Introduction_to_Social_Network_Methods.pdf, accessed 18 September 2015.

HERBERT D. (2000), "School Choice in the Local Environment: headteachers as gatekeepers in an uneven playing field", *School Leadership & Management, 20,* 1, Oxford, Taylor & Francis, pp. 29-97.

HOFF P., RAFTERY, A. & HANDCOCK M. (2002), "Latent Space Approaches to Social Network Analysis", *Journal of the American Statistical Association, 97,* Taylor & Francis, pp. 1090-1098.

HOFF P. (2005), "Bilinear mixed-effects models for dyadic data", *Journal of the American Statistical Association, 100*, Taylor & Francis, pp. 286-295.

HOTELLING H. (1929), "Stability in Competition", *Economic Journal, 39*, London, Wiley-Blackwell, pp. 41-57.

HOXBY C. (2000), "Does Competition Among Public Schools Benefit Students and Taxpayers?", *American Economic Review, 90*, Nashville, American Economic Association, pp. 1209-1238.

HOXBY C. (2007), "Competition Among Public Schools: A reply to Rothstein (2004)", *American Economic Review, 97,* 5, Nashville, American Economic Association, pp. 2038-2055.

KRIVITSKY, P., HANDCOCK, M. (2008), "Fitting position latent cluster models for social networks with latentnet", *Journal of Statistical Software, 24,* 5, Foundation for Open Access Statistics, pp. 1-23.

KRIVITSKY P., HANDCOCK M. (2014), "Latentnet: Latent Position and Cluster Models for Statistical Networks. The Statnet Project", http://www.statnet.org., R package version 2.4.2, https://CRAN.R-project.org/package=latentnet, accessed 18 September 2015.

KUHL P., SIEGLE T. & LENSKI A. (2013), "Soziale Disparitäten" [Social disparities], *in* PANT H.A., STANAT P., SCHROEDERS U., ROPPELT A., SIEGLE T. & PÖHLMANN C. (eds.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I,* Münster, Waxmann, pp. 275-296.

LAZARSFELD P., HENRY N. (1968), *Latent Structure Analysis,* Boston, Houghton Mifflin.

LEIST S., PIETSCH M. (2013), *Modellierung latent-regionaler Schulmärkte [Modelling latent regional education markets]*, Presentation on 1st GEBF in Kiel, 11th to 13th march 2013.

RAFTERY A., LEWIS S. (1995), "The number of iterations, convergence diagnostics and generic metropolis algorithms", *in* GILKS W., SPIEGELHALTER D. & RICHARDSON S. (eds.), *Markov chain Monte Carlo in Practice,* London, Chapman and Hall, pp. 115-130.

ROBERTSON D., SYMONS J. (2003), "Self-selection in the state school system", *Education Economics, 11,* 3, pp. 259-272.

ROTHSTEIN J. (2007), "Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)", *American Economic Review 97,* 5, pp. 2026-2037.

SCHWARZ G. (1976), "Estimating the dimension of a model", *Annals of Statistics,* 6, pp. 461-464.

TAYLOR C. (2001), "Hierarchies and 'local' Markets: the geography of the 'lived' market place in secondary education provision", *Journal of Education Policy, 16,* 3, pp. 197-214.

TEMPLIN J. (2008), "Methods for Detecting Subgroups in Social Networks", *in* CARD N.A., SELIG J.P. & LITTLE T.D. (eds.), *Modeling Dyadic and Interdependent Data in the Developmental and Behavioral Sciences*, Routledge, New York, pp. 309-334.

TESKE P., FITZPATRICK J. & KAPLAN G. (2006), "The Information Gap?", *Review of Policy Research 23,* 5, pp. 969-981.

U.S. DEPARTMENT OF JUSTICE AND THE FEDERAL TRADE COMMISSION (2010), *Horizontal Merger Guidelines*, http://www.justice.gov/atr/horizontal-merger-guidelines-08192010#5c, accessed 21 September 2015.

WEIß M., STEINERT B. (1996), "Germany: Competitive Inequality in Educational Quasi-markets", *in* WALFORD G. (ed.), *School Choice and the Quasi-market,* Wallingford, Triangle Books, pp. 77-94.

WEIß M. (2001), "Quasi-Märkte im Schulbereich. Eine ökonomische Analyse" [Quasi-markets in education. An economic analysis], *Zeitschrift für Pädagogik, 43*, pp. 6-85.

NOTES

1. The density of an unweighted network is calculated by the share of realized connections in relation to the sum of all possible connections. For weighted networks, the density is defined as the sum of the magnitude of all realized connections divided by the number of all possible connections (Hanneman, Riddle, 2005). The density of this (weighted) network is 0.088.
2. The assignment of these two schools is artificial. Both are small private schools whose students made one transition after grade 4. Each transition is fulfilled to the same secondary school which belongs to cluster 1. The structure of the transition of both schools has no similarity in relation to the other schools which "deliver" students to this secondary school. Therefore the assignment wasn't to cluster 1, but instead to another cluster. For the analysis on local level, both schools are considered, but can't tamper the results.
3. The RISE-Status is a social index and calculated on a small spatial scale by the authority responsible for urban development in Hamburg. Used indicators for the index are proportions of: Elderly on Welfare, Children on Welfare, Total Population on Welfare, Unemployed, Single Parents, Students with high Graduation (inverted), Migrants (Freie und Hansestadt Hamburg, 2010).

ABSTRACTS

Most approaches to spatial definitions of schooling markets are based on assumptions which may cause incorrect estimates. This paper presents stochastic network analysis as an alternative approach. Based upon individual student data of the metropolis Hamburg, the results are compared to those of traditional approaches. First, this article gives a short introduction to the current setting of Germany's educational system, the national efforts for school improvement and the relevance of an adequate spatial definition of a market in this context. Subsequently, the applied method and the essential data structure are described. Following the identification of local and regional schooling markets in Hamburg by applying stochastic network analysis, the authors quantify meso- and small-scale competition amongst schools involved in the transition of students from primary to secondary schools in context of social composition.

Herkömmliche Ansätze zur räumlichen Definition schulischer Märkte fussen auf Annahmen, die fehlerhafte Einschätzungen räumlicher Phänomene verursachen können. Dieser Beitrag stellt die Stochastische Netzwerkanalyse als Alternative zur Eingrenzung des Marktgeschehens vor. Die identifizierten Räume werden auf Grundlage vollständiger Schülerindividualdaten der Freien und Hansestadt Hamburg mit Raumdefinitionen traditioneller Herangehensweisen verglichen. Einleitend werden knapp Deutschlands Schulsystem und rezente Maßnahmen zur Qualitätssteigerung vorgestellt, um die gestiegene Relevanz adäquater räumlicher Definitionen zu verdeutlichen. Nachfolgend wird auf die vorgestellte Methode und die dafür notwendige Datenstruktur eingegangen. Im Anschluss an die Identifizierung lokaler und regionaler Marktgeschehen quantifizieren die Autoren mesoskalig und kleinskalig Wettbewerb zwischen Schulen beim Übergang von der Grundschule zur weiterführenden Schule vor dem Hintergrund sozialer Schülerzusammensetzungen.

INDEX

Keywords: competition, geography of education, school market, segregation, stochastic network analysis
Schlüsselwörter: Bildungsgeographie, Schulmarkt, Segregation, stochastische Netzwerkanalyse, Wettbewerb

AUTHORS

SEBASTIAN LEIST

Institut für Bildungsmonitoring und Qualitätsentwicklung Hamburg, Sebastian.leist@hotmail.com.au

MARCUS PIETSCH

Leuphana Universität Lüneburg, pietsch@leuphana.de

Inspektionsbasierte Unterrichtsentwicklung an Schulen in schwieriger Lage

Marcus Pietsch, Stephanie Graw und Klaudia Schulte[1]

Keywords: Schulinspektion, Schulleitung, Sozialindex, Unterrichtsentwicklung

Abstract

Schulinspektionen in Deutschland sollen zu einer evidenzbasierten und zielgerichteten Entwicklung von Schule und Unterricht beitragen. Gleichwohl findet inspektionsbasierte Schul- und Unterrichtsentwicklung immer auch im Kontext der jeweiligen Einzelschule statt. Ob, und falls ja, wie eine entsprechende Entwicklung initiiert wird und gelingt, ist daher stets auch abhängig von den jeweiligen Rahmenbedingungen, unter denen die schulischen Akteure agieren. Im vorliegenden Beitrag wird daher anhand von n=49 Schulen untersucht, ob der soziale Kontext einer Schule einen Effekt auf die inspektionsbasierte Unterrichtsentwicklung hat. Hierfür werden Mehrebenenstrukturgleichungsmodelle genutzt und der direkte und indirekte Einfluss schulischer Rahmenbedingungen auf Unterrichtentwicklungsmaßnahmen analysiert. Die Befunde zeigen, dass der soziale Kontext einer Schule sowohl einen negativen Einfluss darauf hat, ob eine Schule infolge einer Schulinspektion Unterrichtentwicklungsmaßnahmen ergreift als auch auf die Anzahl der ergriffenen Maßnahmen. Diese Effekte sind direkt und werden nicht durch Schulleitungen, die in der Regel als zentrale Ansprechpartner von Inspektionen gelten, moderiert. Für Schulen in schwieriger Lage sollten daher kompensatorische Maßnahmen sowie passgenaue Rückmeldeformate und externe Unterstützungsmaßnahmen bereit gestellt werden, damit eine inspektionsbasierte Entwicklung des Unterrichts erfolgen kann.

---

1 Autor | Leuphana Universität Lüneburg | marcus.pietsch@leuphana.de
Autorin | IfBQ Hamburg | stephanie.graw@ifbq.hamburg.de
Autorin | IfBQ Hamburg | klaudia.schulte@ifbq.hamburg.de

Inhalt

1 Einführung

Schulinspektionen in Deutschland sollen zu einer evidenzbasierten und zielgerichteten Entwicklung von Schule und Unterricht beitragen. Die Grundannahme lautet: Schulinspektionen sammeln systematisch Informationen zu relevanten schulischen Stärken und Schwächen und stellen diese innerschulischen Akteuren bereit, die diese zielorientiert nutzen um eine inspektions- bzw. evidenzbasierte Schul- und Unterrichtsentwicklung zu betreiben.

Gleichwohl findet inspektionsbasierte Schul- und Unterrichtsentwicklung immer auch im Kontext der jeweiligen Einzelschule statt. Ob, und falls ja, wie eine solche Entwicklung initiiert wird und gelingt, ist daher stets auch abhängig von den jeweiligen Rahmenbedingungen, unter denen die schulischen Akteure agieren.

Dies gilt insbesondere für das Schulleitungshandeln. Etliche Untersuchungen haben in den vergangenen Jahren deutlich gemacht, dass dieses von dem spezifischen Setting abhängt, in dem sie tätig sind. Insbesondere der kulturelle, der ökonomische und der soziale Kontext bedingen dabei das Handeln von Schulleitungen (Hallinger 2016).

Da Schulleitungen die zentralen Adressaten und Ansprechpartner für Schulinspektionen in Deutschland sind, ist anzunehmen, dass der soziale Kontext einer Schule einerseits einen direkten Einfluss darauf hat, ob und wie eine inspektionsbasierte Schul- und Unterrichtsentwicklung stattfindet, andererseits aber auch indirekte, über die Schulleitung vermittelte, Effekte des sozialen Kontexts zu beobachten sind.

Im folgenden Beitrag wird daher untersucht, ob, und falls ja welche, Effekte der soziale Kontext einer Schule für die inspektionsbasierte Unterrichtsentwicklung hat. Nach einer kurzen Einführung zum Thema inspektionsbasierte Schul- und Unterrichtswicklung an Schulen im sozialen Kontext, wird mithilfe von Mehrebenenstrukturgleichungsmodellen untersucht, welche Rolle der soziale Kontext einer Schule für die Unterrichtsentwicklung spielt. Abschließend werden die Befunde mit Blick auf die Konsequenzen für Wissenschaft und Praxis diskutiert.

2 Erfolgreiche Unterrichtsentwicklung an Schulen in schwieriger Lage

Schulen in schwieriger Lage, die in sozialen Brennpunkten mit einer Schülerklientel aus belasteten und bildungsfernen häuslichen Milieus arbeiten und die sich vergleichsweise stark verbessern resp. entwickeln, zeichnen sich vor allem dadurch aus, dass sie relevante Expertise und Beratung im Umgang mit Daten aus (externen) Evaluation mobilisieren konnten (Thomas, Walker &

Webb 1998), dass an diesen Schulen diesbezüglich besonders engagierte und kompetente Schulleitungen agieren (Leithwood, Harris & Strauss 2010) und dass sich alle Schulbeteiligten auf das Lernen und Lehren an der Schule fokussieren (Potter, Reynolds & Chapman 2001).

Wie Hallinger und Murphy (1986, s. auch Gärtner 2015 und Tab. 1) verdeutlichen, unterscheiden sich Schulen, denen es trotz sozial herausfordernder Voraussetzungen gelingt, hohe Lernzuwächse aufseiten der Schülerinnen und Schüler zu erzielen, von weniger erfolgreichen Schulen vor allem dadurch, dass Grundlagen geschaffen, Ziele vergleichsweise eng gesteckt und Monitoring sowie Qualitätskontrollen, insbesondere mit Blick auf die Gestaltung des Unterrichts, ausgebaut werden. Darüber hinaus kommt der Schule als Lernort eine besondere Bedeutung zu.

Mujis et al. fassen entsprechend zusammen (2004, S. 170): „*In all cases, teaching and learning should be at the heart of the school, driving its daily efforts. Effective, instructional leadership, data richness, and having high expectations of achievement among staff, pupils, and parents, are likewise elements that would appear to characterise all improving and effective schools.*"

Entwicklungen an Schulen in schwieriger sozialer Lage erfolgen dabei in der Regel in zwei Phasen (Teddlie, Stringfield & Reynolds 2003): 1. Eine kompensatorische Phase, in der die Voraussetzungen geschaffen werden müssen, um hierauf aufbauend Entwicklungsprozesse zu etablieren und 2. eine langfristige Phase, die darauf ausgerichtet ist, Prozesse auf Ebene der Schule und des Unterrichts nachhaltig zu verändern.

Grundvoraussetzungen für eine erfolgreiche Schulentwicklung an Schulen in sozial schwierigen Lagen sind daher primär (Lasky et al. 2007): a) Die Schaffung eines geregelten, geordneten und lernförderlichen Schulklimas, b) die Befähigung von Schulleitungen den Veränderungsprozess aktiv zu gestalten, c) die Etablierung angemessen hoher Erwartungen an alle Schulbeteiligten, d) die Kultivierung einer innerschulischen Feedbackkultur, in der die Beteiligten Rückmeldungen als Entwicklungsimpuls verstehen sowie e) die Befähigung aller Schulbeteiligten sich kontinuierlich weiter zu entwickeln.

Mit Blick auf eine inspektionsbasierte Unterrichtsentwicklung kommt dabei wiederum der Schulleitung eine besondere Bedeutung zu, da Schulleitungen zentrale Ansprechpartner und Adressaten von Schulinspektionen in Deutschland sind (Pietsch & Hosenfeld 2017). Entsprechend hat sich die Rolle von Schulleiterinnen und Schulleitern an Schulen in Deutschland in den vergangenen Jahren infolge der Einführung von Inspektionsverfahren grundlegend verändert.

So hat sich ihre Rolle einerseits vom „*Primus inter Pares zum Manager und Leader*" (Wissinger, 2016 S. 261) gewandelt. Andererseits wurden Schulleitungen durch die Einführung neuer Steuerungsmechanismen gegenüber Lehrkräften strukturell stark aufgewertet (Preuß, Wissinger & Brüsemeister 2015).

Ob die Entwicklung von Schule und Unterricht an einer Schule, auch infolge einer Schulinspektion, gelingt, hängt daher mittlerweile zunehmend von den Einstellungen, Fähigkeiten und Kompetenzen ab, über die Schulleitungen verfügen (Sowada & Terhart 2015).

Schulmerkmale	**Schule mit sozial schwacher Schülerschaft**	**Schule mit sozial starker Schülerschaft**
Schulinternes Curriculum		
Breite	Eng	Breit
Anforderungen	Grundlegend	Akademisch
Verknüpfung mit Unterricht	Moderat	Eng
Lerngelegenheiten		
Schwerpunktsetzungen	Basiskompetenzen	Akademisch
Erwartungen an Hausaufgaben	gering bis moderat	Hoch
Leitbild		
Art des Leitbildes	Beherrschen von Basiskompetenzen	Beherrschen akademischer Leistungen
geteilte Überzeugung im Kollegium	Hoch	Hoch
Unterrichsbezogene Führung		
Koordination des Curriculums	Hoch	Hoch
Überwachung der Unterrichts	Hoch	gering bis moderat
Aufgabenorientierung	Hoch	Moderat
Beziehungsorientierung	gering bis moderat	moderat bis hoch
Kooperation Schule-Eltern		
Nähe zum Elternhaus	Gering	Hoch
Beteiligung der Eltern	Gering	Allgegenwärtig
Rolle der Schulleitung	Puffer	Realisierung von Schnittstellen
Belohnung		
Häufigkeit von Belohnungen	Hoch	Gering
Art der Belohnungen	extrinsisch, öffentlich	intrinsisch, privat
Hohe Erwartungen		
Quelle der Erwartungen	Schule	Schule und Elternhaus
aktuelle Erwartungen	Hoch	Hoch
zukünftige Erwartungen	Moderat	Hoch

Tab. 1: Profile effektiver Schulen mit sozial schwacher bzw. starker Schülerschaft

Gleichwohl agieren auch Schulleitungen nicht unabhängig vom sozialen Kontext einer Schule (Hallinger 2016). Während Befunde aus Deutschland hierzu kaum vorliegen, zeigen internationale Befunde (zur Übersicht: Klein 2018) einerseits, dass an Schulen in sozial schwieriger Lage häufiger Schulleitungswechsel stattfinden, was eine zielorientierte und nachhaltige Entwicklung von Schule und Unterricht behindern kann, und dass auch das konkrete Führungshandeln von Schulleitungen durch die sozialen Rahmenbedingungen geprägt sein kann, da häufig akute Problemlösungsbemühungen im Fokus von Schulleitungen stehen, die unter solchen Bedingungen arbeiten.

3 Methodisches Vorgehen

3.1 Forschungsfragen

Vor diesem Hintergrund sollen die folgenden Forschungsfragen beantwortet werden:

1. Lassen sich direkte und indirekte, über das Handeln der Schulleitung vermittelte, Effekte des sozialen Kontexts auf die inspektionsbasierte Unterrichtsentwicklung nachweisen?
2. Lässt sich ein Effekt des sozialen Kontexts einer Schule darauf finden, ob eine Schule infolge einer Inspektion Unterrichtentwicklungsmaßnahmen initiiert?
3. Lässt sich ein Effekt des sozialen Kontexts einer Schule darauf finden, wie viele Unterrichtentwicklungsmaßnahmen eine Schule infolge einer Inspektion initiiert?

3.2 Stichprobe

Grundlage für die nachfolgende Untersuchung sind Informationen von 49 Schulen, die zwischen 2012 und 2014 zum zweiten Mal durch die Schulinspektion Hamburg extern evaluiert wurden, wobei die ersten Inspektionen der Schulen in den Jahren 2007 und 2010 stattfanden. Die Stichprobe setzt sich aus 30 Grundschulen, 12 Gymnasien, sechs Stadtteilschulen und einer Sonderschule zusammen. Einige Grundschulen wurden im ersten Zyklus als Grund- Haupt und Realschulen geführt, sind jedoch infolge von Reformmaßnahmen in reine Grundschulen überführt worden, mussten also auch mit strukturellen Veränderungen umgehen.

Um zu untersuchen, welchen Einfluss der soziale Kontext von Schulen auf die inspektionsbasierte Unterrichtsentwicklung an den Schulen hat, werden Daten aus mehreren voneinander unabhängigen Datenquellen kombiniert. Ers-

tens werden Daten aus den regelhaften Erhebungen der Schulinspektion Hamburg genutzt, um das Führungshandeln von Schulleitungen auf Basis von Angaben durch Lehrkräfte zu beschreiben. Diese Informationen werden zweitens mit Schulentwicklungsberichten, die wiederum Schulleitungen ausgefüllt haben, in Verbindung gebracht. Und drittens werden Informationen des hiervon unabhängig ermittelten Hamburger Sozialindex berücksichtigt, um den sozialen Kontext der Schulen zu beschreiben.

3.3 Datengrundlage

Die *inspektionsbasierte Unterrichtsentwicklung* wird anhand von Entwicklungsberichten dargestellt, die durch Schulleitungen zum Zeitpunkt der zweiten Inspektion abgegeben wurden. Hierin werden die Schulleitungen darum gebeten, anzugeben, welche für die Schulen bedeutenden Entwicklungsmaßnahmen infolge der ersten Inspektion ergriffen sowie ob und wie diese Maßnahmen umgesetzt wurden. Die Berichte sollen auf den zwischen Schule und Schulaufsichten getroffenen Ziel-Leistungs-Vereinbarungen basieren. Der Schulaufsicht obliegt es auch die Umsetzung der Maßnahmen zu kontrollieren. Für die vorliegende Studie wird auf Informationen aus einer Inhaltsanalyse dieser Informationen zurückgegriffen (Pietsch, Feldhoff & Petersen 2016). Als Maßnahmen der Unterrichtsentwicklung wurden dabei z.B. folgende Angaben kodiert: „Individualisierung des Unterrichts voran treiben“, „Förderung der Arbeit mit Smartboards im Unterricht“ und „kompetenzorientierte Arbeitsmaterialien entwickeln“. Insgesamt wurden nach Aussage der Schulleitungen der 49 Schulen insgesamt 119 solcher Unterrichtentwicklungsmaßnahmen durchgeführt, im Schnitt rund 2,4 Maßnahmen je Schule, wobei die Spannbreite von null bis sieben Maßnahmen reichte.

Das *Schulleitungshandeln* wird anhand von Daten modelliert, die zum Zeitpunkt der zweiten Inspektion mithilfe von Fragebögen an Lehrkräfte erhoben wurden. Dabei knüpft die Inspektion an die Idee des lernzentrierten Führungshandelns (*Leadership for Learning, Boys & Bowers 2018; Hallinger 2011*) an und erfasst daher u.a. die Dimensionen instruktionale, transformationale und geteilte Führung mithilfe etablierte Instrumente: a) Transformationale Führung: Skalen aus dem Multifactor Leadership Questionnaire (MLQ, vgl. Bass & Avolio 1995), b) Instruktionale Führung: Skalen aus dem Teaching and Learning International Survey (TALIS, vgl. Schmich & Schreiner 2008; OECD 2009) und c) Geteilte Führung: Skala aus einem Instrument von Wahlstrom und Louis (2008) zur Erhebung von *Shared Leadership Among Principal and Others.* Die Stichprobe der Lehrkräfte an den 49 Schulen umfasst n=1140 Personen. Die Beschreibung der genutzten Facetten sowie die zugehörigen Skalenkennwerte

finden sich in Tabelle 2, eine kurze Beschreibung der Facetten sowie Beispielitems in Anhang A.

Der *soziale Kontext der Schulen* wiederum wird anhand des offiziellen Hamburger Sozialindex für Schulen aus dem Jahr 2013 beschrieben (Schulte, Hartig & Pietsch 2014). Der Sozialindex liegt auf Schulebene vor und berücksichtigt Informationen zum sozialen Hintergrund der Schülerfamilien sowie sozialräumliche Daten der jeweiligen Wohnorte. Die Skala ist eindimensional und wo möglich wurden standardisierte Maße, z. B. EGP-Klassen (Erikson & Goldthorpe 1992), als Indikatorvariablen berücksichtigt. Genutzt werden in der vorliegenden Studie Faktorscores. Ein niedriger Wert zeigt dabei eine geringe soziale Belastung der Schule, ein hoher Wert eine hohe soziale Belastung an. Für die genutzte Schulstichprobe reichen die Werte von -1,53 bis zu 1,90. Der Mittelwert liegt bei -0,07, die Standardabweichung liegt bei 0,94.

Als *Kontrollvariablen* werden darüber hinaus berücksichtigt: a) Die Schulform, b) ob zwischen der ersten und der zweiten Inspektion an der Schule ein Schulleiterwechsel stattgefunden hat (Dummy-kodiert) und c) ob sich die Struktur der Schule in diesem Zeitraum (Überführung von Grund-, Haupt- und Realschule in reine Grundschule) verändert hat (Dummy-kodiert).

3.4 Statistische Modellierung

Die nachfolgenden Analysen beschreiben den Zusammenhang des sozialen Kontexts sowie der inspektionsbasierten Unterrichtsentwicklung grundsätzlich auf Ebene der Schule, da Entwicklungsberichte nur auf dieser Ebene vorliegen. Gleichwohl erfolgte die Erhebung der Daten zum Schulleitungshandeln durch Befragung von Lehrkräften. Entsprechend ist bei der Datenanalyse zu beachten, dass Schulleitungshandeln auf mehreren Ebenen eine Schule angesiedelt ist. So kann sich eine Schulleitung dem gesamten Mitarbeiterstab widmen, einzelnen Gruppen innerhalb des Kollegiums sowie einzelnen Lehrkräften.

Um dies in den Analysen angemessen zu berücksichtigen, wurden mit MPLUS 7.0 (Muthén & Muthén 2012) so genannte doppelt-latente (doubly-latent) Strukturgleichungsmodelle geschätzt (Marsh et al. 2009), die die Führungskonstrukte durch die Spezifizierung analoger Messmodelle auf Lehrkraft- und Schulebene stichproben- und messfehlerbereinigt erfassen. Fehlende Werte wurden mithilfe der in MPLUS 7.0 implementierten Maximum-Likelihood-Methode behandelt.

Als Gütekriterien für die Modellanpassung werden der Comparative Fit Index (CFI), der Root Mean Square Error of Approximation (RMSEA) sowie der Standardized Root Mean Square Residual (SRMR, dieser für beide Ebenen) berichtet. Als akzeptable Modellfits gelten dabei folgende Werte (Hu & Bentler

1999; Marsh, Hau & Wen 2004): CFI-Werte ≥ 0,90, RMSEA-Werte ≤ 0,08 und SRMR ≤ 0,08.

3.5 Vorläufige Prüfung der Mehrebenenannahmen

Um zu prüfen, ob das intendierte mehrebenenanalytische Vorgehen angemessen ist, wurden in einem ersten Schritt Intra-Klassen-Korrelationen (ICC1: Varianz zwischen Schulen, ICC2: Inter-Lehrer-Reliabilität auf Schulebene) für alle Führungsskalen mithilfe des R-Paketes `multilevel` (Bliese 2016) berechnet. In der Literatur (Bliese 2000; Lüdtke et al. 2008) wird dabei davon ausgegangen, dass Daten mithilfe eines Mehrebenenansatzes modelliert werden müssen, wenn die ICC1≥ 0,05 und die ICC2≥ 0,70 ist.

	MW	SD	ICC1	ICC2	ω_{Ebene1}	ω_{Ebene2}
Transformationale Führung						
Einfluss durch Vorbildlichkeit und Glaubwürdigkeit	2,98	0,84	0.271	0.898	0.883	0.945
Individuelle Unterstützung und Förderung	3,13	0,71	0.322	0.918	0.904	0.969
Motivation durch begeisternde Visionen	2,99	0,87	0.202	0.857	0.772	0.885
Instruktionale Führung						
Schulzielmanagement	2,83	0,79	0.237	0.880	0.793	0.895
Unterrichtsmanagement	2,71	0,81	0.281	0.903	0.759	0.976
Direkte Supervision des Unterrichts	2,09	0,76	0.195	0.852	0.732	0.790
Geteilte Führung	3,06	0,70	0.245	0.885	0.773	0.884

Tab. 2: Kennwerte der eingesetzten Führungsskalen

Wie Tabelle 2 zeigt, überschreiten alle Führungsskalen diese Cut-Off-Werte deutlich, weisen also eine eindeutige Mehrebenenstruktur auf. 20 bis 32 Prozent der beobachteten Varianz (ICC1) liegt zwischen Schulen, das Führungshandeln unterscheidet sich demnach klar von Schule zu Schule. Auch die Konsistenz der Einschätzungen durch die Lehrkräfte (ICC2) ist mit Werten von 0.852 bis 0.918

sehr hoch und bestätigt, dass im Weiteren der Einsatz eines mehrebenenanalytischen Designs notwendig und angemessen ist.

4 Befunde

Um den Einfluss des sozialen Kontexts einer Schule auf die inspektionsbasierte Unterrichtsentwicklung zu prüfen, wurden in einem ersten Schritt jeweils zwei Modellvarianten überprüft und bezüglich der Modellpassung miteinander verglichen. Einerseits wurde ein indirektes Pfadmodell spezifiziert, bei dem angenommen wird, dass der Kontext einer Schule sowohl direkt auf die Entwicklungsmaßnahmen als auch indirekt, vermittelt über das Schulleitungshandeln und mögliche Schulleiterwechsel zwischen dem ersten und zweiten Messzeitpunkt, einen Einfluss ausübt. Andererseits wurde jeweils ein Modell ohne entsprechende Moderationen geschätzt.

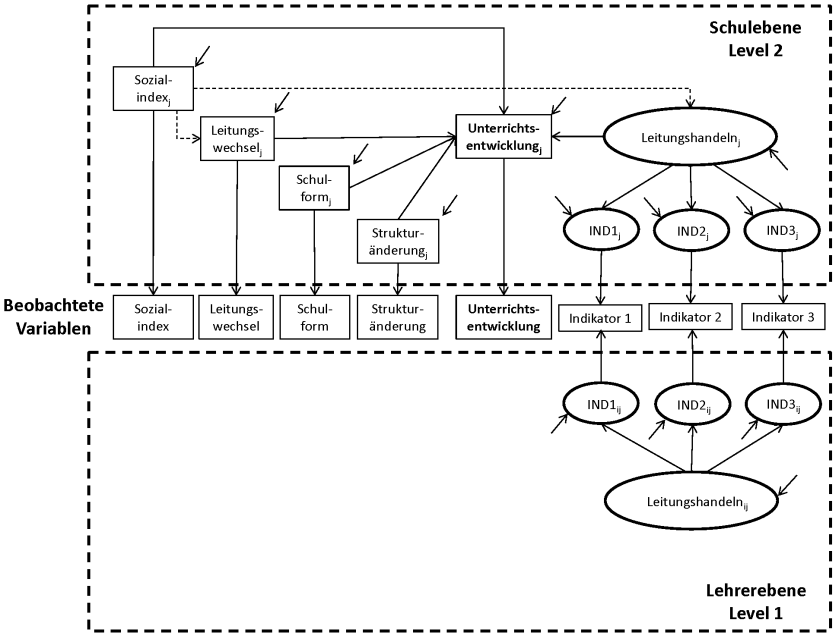


Abbildung 1: Schematische Darstellung der spezifizierten doppelt-latenten Strukturgleichungsmodelle

Eine schematische Darstellung der Modelle findet sich in Abbildung 1. Beide Modelle wurden anschließend mithilfe des Aikake-Schwarz- (AIC) sowie des Bayesianischen Informationskriteriums (BIC) verglichen, wobei ein niedrigerer Wert eine bessere Passung auf die vorhandenen Daten indiziert (West, Taylor & Wu 2012).

	Maßnahmen eingeleitet ja/nein		**Anzahl eingeleiteter Maßnahmen**	
	B	**p**	**β**	**p**
Transformationale Führung				
Einfluss durch Vorbildlichkeit und Glaubwürdigkeit	n.s.		n.s.	
Individuelle Unterstützung und Förderung	n.s.		n.s.	
Motivation durch begeisternde Visionen	n.s.		n.s.	
Instruktionale Führung				
Schulzielmanagement	n.s.		n.s.	
Unterrichtsmanagement	n.s.		n.s.	
Direkte Supervision des Unterrichts	n.s.		n.s.	
Geteilte Führung	n.s.		n.s.	
Sozialindex	-.251	.064	-.409	.011
Schulleiterwechsel	n.s.		n.s.	
Schulform	n.s.		n.s.	
Strukturänderung	n.s.		n.s.	
R^2	**.679**		**.676**	
Fit-Werte	CFI: .943, RMSEA: .043, $SRMR_w$: .034, $SRMR_b$: .118		CFI: .943, RMSEA: .043, $SRMR_w$: .034, $SRMR_b$: .117	

Tab. 3: Effekte schulischer Bedingungen auf die inspektionsbasierte Unterrichtsentwicklung, Standardisierte Regressionskoeffizienten auf Ebene der Schule (Level 2), R^2, Fit-Indices; n.s.= nicht (1- oder 2-seitig) signifikant (*p>.100)*.

Diese Werte weisen auf eine deutlich bessere Passung derjenigen Modelle hin, in denen keine indirekten Pfade berücksichtigt wurden (Maßnahmen eingeleitet: ΔBIC=199,745, ΔAIC=130,136; Anzahl eingeleiteter Maßnahmen: ΔBIC=133,146, ΔAIC=202,737). Auch lassen sich im indirekten Pfadmodell weder statisch signifikante Effekte des Sozialindex auf das Schulleitungshandeln noch auf Schulleiterwechsel nachweisen. Ob an einer Schule ein Leitungswechsel zwischen den Messzeitpunkten stattfand und wie eine Schulleitung im Schulalltag agiert, ist demnach in der vorliegenden Stichprobe unabhängig von den sozialen Rahmenbedingungen einer Schule.

Mit Blick auf die inspektionsbasierte Unterrichtsentwicklung zeigt sich hingegen ein klarer Zusammenhang mit den sozialen Rahmenbedingungen einer Schule (s. Tabelle 3). Einerseits haben die sozialen Rahmenbedingungen tendenziell einen Effekt darauf, ob an einer Schule auf Basis einer Schulinspektion überhaupt Maßnahmen für die Unterrichtsentwicklung ergriffen werden ($\beta_{\text{Maßnahmen eingeleitet,L2}} = -.251$). Andererseits haben diese Bedingungen dann wiederum einen noch größeren Effekt darauf, wie viele Maßnahmen in der Folge durchgeführt werden ($\beta_{\text{Anzahl Maßnamen,L2}} = -.409$). Für alle weiteren im Modell berücksichtigten Variablen (Führungspraktiken, Schulleiterwechsel usw.) konnten hingegen keine signifikanten Zusammenhänge mit der inspektionsbasierten Unterrichtsentwicklung festgestellt werden.

5 Diskussion und Fazit

Auf Basis der Ergebnisse des vorherigen Kapitels lassen sich mit Blick auf die erste Forschungsfrage direkte Effekte des sozialen Kontexts einer Schule, jedoch keine indirekten Effekte vermittelt über die Schulleitung, feststellen. Darüber hinaus zeigt sich in Bezug auf die zweite und dritte Forschungsfrage sowohl ein Effekt auf die Frage, ob infolge einer Inspektion Unterrichtsentwicklungsmaßnahmen initiiert wurden, als auch auf die Frage, wie viele inspektionsbasierte Maßnahmen eine Schule initiiert hat.

An Schulen mit einer hohen sozialen Belastung werden demnach im Vergleich zu Schulen mit einer geringen sozialen Belastung seltener und weniger inspektionsbasierte Maßnahmen der Unterrichtsentwicklung durchgeführt. Das Schulleitungshandeln, die Schulform sowie ein möglicher Schulleitungs- oder Schulformwechsel spielen bei der Initiierung von Maßnahmen hingegen keine Rolle. Nicht berücksichtigt werden konnte bei den Analysen die Qualität sowie das Ausmaß der ergriffenen Maßnahmen. Hier bietet sich ein Anknüpfungspunkt für zukünftige Studien.

Entsprechend ist die Anzahl der eingeleiteten Unterrichtsentwicklungsmaßnahmen auch kein besonders guter Indikator für eine gute, gelingende Unterrichtsentwicklung. Dass jedoch die Wahrscheinlichkeit, ob eine Schule infolge einer Schulinspektion überhaupt Maßnahmen zur Entwicklung des Unterrichts ergreift, systematisch mit den sozialen Rahmenbedingungen kovariiert, ist ein Befund, der vermuten lässt, dass Schulen in schwieriger Lage sich häufig noch in einer kompensatorischen Phase befinden, in der überhaupt erst die Voraussetzungen (z.B. die Schaffung eines geregelten, geordneten und lernförderlichen Schulklimas und der Aufbau von Vertrauen zwischen allen Schulbeteiligten) für weitere Entwicklungen geschaffen werden müssen. Auch hier ist im Weiteren empirisch zu klären, ob diese Annahme zutrifft.

Besonders bemerkenswert ist darüber hinaus, dass in den empirischen Analysen, anders als international häufig berichtet, keine empirischen Zusammenhänge zwischen Schulleitungsmerkmalen bzw. –handeln und den sozialen Kontextbedingungen einer Schule festgestellt werden konnten. Zumindest in der hier vorliegenden Stichprobe unterscheidet sich somit weder das Leitungshandeln, noch die Wahrscheinlichkeit, dass eine Schulleitung eine Schule in schwieriger Lage verlässt, substanziell von anderen Schulen mit weniger herausfordernden Kontextbedingungen. Diesbezüglich stellt sich die Frage, inwieweit Annahmen und Befunde der Schulleitungsforschung aus dem anglo-amerikanischen Raum auf Deutschland übertragbar sind.

Mit Blick auf die inspektionsbasierte Unterrichtsentwicklung an Schulen in schwieriger sozialer Lage machen die Befunde jedoch deutlich, dass die sozialen Rahmenbedingungen einer Schule beachtet werden müssen, wenn Schulinspektionen entsprechende Wirkungen nach sich ziehen sollen. Dabei implizieren die Ergebnisse der vorliegenden Untersuchung für die Durchführung von Schulinspektionen bzw. die inspektionsbasierte Unterrichtsentwicklung vor allem die besondere Herausforderung, das Inspektionsverfahren so zu gestalten, dass auch Schulen mit einer hohen sozialen Belastung in die Lage versetzt werden, infolge einer Inspektion Unterrichtentwicklungsmaßnahmen einzuleiten.

Da Schulleitungen einer der wichtigsten Treiber für die Schul- und Unterrichtsentwicklung und darüber hinaus die zentralen Ansprechpartner und Adressaten von Schulinspektionen sind, stehen sie bei diesen Überlegungen grundsätzlich im Fokus. Die vorgelegten Analysen machen jedoch deutlich, dass Schulleitungen anscheinend weniger Einfluss auf die inspektionsbasierte Unterrichtsentwicklung haben, als angenommen. Zumindest die Tatsache, ob überhaupt (und wie viele) entsprechende Entwicklungsmaßnahmen infolge einer Inspektion ergriffen werden (können), hängt nicht von der Führung einer Schulleitung ab. Entsprechend kommen den entwicklungsförderlichen Rahmenbedingungen einer Schule bzw. externen Unterstützungsmaßnahmen eine besondere Bedeutung zu (Bryk et al. 2010).

Einerseits kann dabei die Gestaltung der Rückmeldung der Inspektionsergebnisse die Initiierung von Maßnahmen von Schulen mit hoher sozialer Belastung unterstützen. Die Rückmeldung sollte so gestaltet werden, dass sie im Rahmen der kompensatorischen Phase der Entwicklung die Voraussetzungen für die langfristigen Entwicklungsprozesse schafft (Harris 2010). Um den Schulen die Weiterarbeit mit den Inspektionsergebnissen zu erleichtern, scheint es sinnvoll, dass die Schulinspektion passgenaue Formen der Rückmeldung anbietet. Durch die Formulierung konkreter Bereiche für Entwicklungen an der Schule sowie Blaupausen für klar definierte Entwicklungsschritte ergeben sich direkte Anknüpfungspunkte für die zielorientierte Weiterarbeit (Slavin 2005).

Die vorgelegten Befunde lassen jedoch vermuten, dass die innerschulischen Ressourcen an Schulen in schwieriger Lage für sich genommen nicht ausreichen, um aus eigener Kraft eine inspektionsbasierte Unterrichtsentwicklung zu betreiben. Externe Unterstützungsleistungen sollten daher nicht nur von der Schulinspektion, im Sinne eines Entwicklungsimpulses, sondern auch von weiteren externen Akteuren übernommen werden, um eine inspektionsbasierte Unterrichtsentwicklung zu ermöglichen (Leithwood, Harris & Strauss 2010). Diese Akteure könnten die Schule über einen längeren Zeitraum, auch nach Abschluss der Inspektion – und ggf. gemeinsam mit dieser (Johnson et al. 2009) –, begleiten. Dadurch könnte die Unterstützung auf die langfristige Phase der Entwicklung ausgeweitet werden und es gelingen, den für die Schul- und Unterrichtsentwicklung hinderlichen Kontextbedingungen von Schulen in schwieriger Lage nachhaltig entgegen zu wirken.

Literatur

Bass, B. M., & Avolio, B. J. (1995). MLQ Multifactor Leadership Questionnaire. Technical Report. Redwood City: Mind Garden.

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

Bliese, P. D. (2016). Multilevel Modeling in R (2.6)—A Brief Introduction to R, the multilevel package and the nlme package.

Boyce, J., & Bowers, A. J. (2018). Toward an evolving conceptualization of instructional leadership as leadership for learning: Meta-narrative review of 109 quantitative studies across 25 years. Journal of Educational Administration, 56 (2), https://doi.org/10.1108/JEA-06-2016-0064.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). Organizing schools for improvement: Lessons from Chicago. Chicago: University of Chicago Press.

Erikson, R., & Goldthorpe, J. H. (1992). The Constant Flux: A Study of Class Mobility in Industrial Societies. Oxford: Oxford University Press.

Gärtner, H. (2015). Zusammenhang von Schul- und Unterrichtsqualität und schulischen Rahmenbedingungen. In M. Pietsch, B. Scholand & K. Schulte (Hrsg.), Schulinspektion in Hamburg. Der erste Zyklus der 2007 – 2013: Grundlagen, Befunde, Perspektiven (S. 273-294). Münster: Waxmann.

Hallinger, P. (2011). Leadership for learning: lessons from 40 years of empirical research. *Journal of Educational Administration*, *49*(2), 125–142.

Hallinger, P. (2016). Bringing context out of the shadows of leadership. *Educational Management Administration & Leadership*, *46*(1), 5–24.

Hallinger, P., & Murphy, J. (1986). The social context of effective schools. American Journal of Education, 94 (3), 328-355.

Harris, A. (2010). Improving schools in challenging contexts. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Hrsg.), Second international handbook of educational change (S. 693-706). Dordrecht: Springer.

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. Structural Equation Modeling, 6 (1), 1-55.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. American Journal of Evaluation, 30(3), 377-410.

Klein, E. D. (2018). Erfolgreiches Schulleitungshandeln an Schulen in sozial deprivierter Lage. Eine Zusammenschau zentraler Grundlagen und Befunde aus der nationalen und internationalen Bildungsforschung. Expertise im Auftrag der Wübben Stiftung. SHIP Working Paper Reihe, No. 02. Essen: Universität Duisburg-Essen.

Lasky, S., Datnow, A., Stringfield, S., & Sundell, K. (2007). Diverse populations and school effectiveness and improvement in the USA. In T. Townsend (Ed.), International Handbook of School Effectiveness and Improvement (S. 557-579). Dordrecht: Springer.

Leithwood, K., Harris, H., Strauss, T. (2010), Leading school turnaround: how successful leaders transform low-performing schools. London: Jossey-Bass.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. Psychological Methods, 13(3), 203–229.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. Structural Equation Modeling: A Multidisciplinary Journal, 11(3), 320–341.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error. Multivariate Behavioral Research, 44(6), 764–802.

Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socioeconomically disadvantaged areas – A review of research evidence. School Effectiveness and School Improvement, 15(2), 149-175

Muthén, L. K., & Muthén, B. O. (2012). Mplus software (Version 7). Los Angeles: Muthén & Muthén.

OECD (2009). Creating Effective Teaching and Learning Environments. First Results from TALIS. Paris: OECD.

Pietsch, M., & Hosenfeld, I. (2017). Von der Schulinspektion zur Unterrichtsentwicklung: Welche Rolle spielt die Schulleitung. Empirische Pädagogik, 31(2), 202-220.

Pietsch, M., Feldhoff, T., & Petersen, L. S. (2016). Schulentwicklung durch Inspektion? Welche Rolle spielen innerschulische Verarbeitungskapazitäten? In AG Schulinspektion (Hrsg.),

Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten (S. 227-262). Wiesbaden: VS.

Potter, D., Reynolds, D., & Chapman, C. (2001). School Improvement for Schools Facing Challenging Circumstances: A review of research and practice, School Leadership & Management, 22 (3), 243–256.

Preuß, B., Wissinger, J., & Brüsemeister, T. (2015). Einführung der Schulinspektion: Struktur und Wandel regionaler Governance im Schulsystem. In H. J. Abs, T. Brüsemeister, M. Schemmann & J. Wissinger (Hrsg.), Governance im Bildungssystem (S. 117-142). Wiesbaden: Springer Fachmedien Wiesbaden.

Schmich, J., & Schreiner, C. (2008). TALIS 2008. Schule als Lernumfeld und Arbeitsplatz. Graz: Leykam.

Schulte, K., Hartig, J., & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann, D. & N. Maritzen (Hrsg.), Grundlagen für eine daten- und theoriegestützte Schulentwicklung – Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ) (S. 67-80). Münster: Waxmann.

Slavin, R. E. (2005). Sand, bricks, and seeds: School change strategies and readiness for reform. In Hopkins, D. (Hrsg.), The Practice and Theory of School Improvement (S. 265-279). Dordrecht: Springer.

Sowada, M. G., & Terhart, E. (2015). Schulinspektion und Unterrichtsentwicklung. In H.-G. Rolff (Hrsg.), Handbuch Unterrichtsentwicklung (S. 195-208). Weinheim und Basel: Beltz.

Teddlie, C., Stringfield, S., & Reynolds, D. (2000). Context issues within school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), The International Handbook of School Effectiveness Research (S. 160–185), London: Falmer.

Thomas, G., Walker, D., & Webb, J. (1998) The Making of the Inclusive School. London: Routledge.

Wahlstrom, K. L. & Louis, K. S. (2008). How Teachers experiences Principal Leadership. The Roles of Professional Community, Trust, Efficiency, and Shared Responsibility. Educational Administration Quarterly, 44, 458-495.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling (S. 209-231). New York: Guilford.

Wissinger, J. (2016). Schulleitung im Fokus des Schulqualitätsdiskurses. In T. Steffens & T. Bargel (Hrsg.), Schulqualität - Bilanz und Perspektiven: Grundlagen der Qualität von Schule 1 (S. 257-276) (1. Auflage). Münster: Waxmann.

Anhang A

Skala	Beschreibung	Beispielitem
Transformationale Führung	1 = nie; 2 = selten; 3 = häufig; 4 = (fast) immer	
Einfluss durch Vorbildlichkeit und Glaubwürdigkeit	Die Schulleitung ist Vorbild für die Mitarbeiter und beeinflusst diese nachhaltig.	Die Schulleiterin/Der Schulleiter strahlt Stärke und Vertrauen aus.
Individuelle Unterstützung und Förderung	Die Schulleitung versteht sich als Coach/Mentor ihrer Mitarbeiter.	Die Schulleiterin/Der Schulleiter berücksichtigt meine Individualität und behandelt mich nicht nur als irgendeine Mitarbeiterin/irgendeinen Mitarbeiter.
Motivation durch begeisternde Visionen	Die Schulleitung begeistert mit attraktiven Zukunftsversionen und steht hinter diesen.	Die Schulleiterin/Der Schulleiter formuliert eine überzeugende Zukunftsvision.
Instruktionale Führung	1 = selten oder nie; 2 = selten; 3 = oft; 4 = sehr oft	
Schulzielmanagement	Die Schulleitung bemüht sich die Bildungsziele umzusetzen. Zur Erreichung der benannten Lernziele werden Unterrichts- und Fortbildungsaktivitäten der Lehrkräfte aufeinander abgestimmt.	Die Schulleiterin/Der Schulleiter stellt sicher, dass die Arbeit der Pädagoginnen und Pädagogen mit den Lehrzielen der Schule übereinstimmt.
Unterrichtsmanagement	Im Fokus steht die stetige Optimierung des Unterrichts. Die Schulleitung steht den Lehrkräften als Ansprechpartner bei Problemen zur Verfügung.	Die Schulleiterin/Der Schulleiter kümmert sich um Probleme in Bezug auf störendes Verhalten in den Klassen.
Direkte Supervision des Unterrichts	Die Schulleitung versucht Unterrichtsaktivitäten und Bildungsziele in Übereinstimmung zu bringen. Die Lehrkräfte erhalten Anstöße zur Verbesserung des Unterrichts.	Die Schulleiterin/Der Schulleiter oder jemand anderes aus dem Leitungsteam hospitiert im Unterricht.
Geteilte Führung	1 = trifft nicht zu; 2 = trifft eher nicht zu; 3 = trifft eher zu; 4 = trifft zu	
	Das gesamte Kollegium wird bei steuerungsrelevanten Entscheidungen durch die Schulleitung aktiv beteiligt.	Die Schulleiterin/der Schulleiter sorgt für eine umfassende Beteiligung, wenn Entscheidungen zur Schulentwicklung anstehen.

Zeitschrift für Bildungsforschung

The Effects of Competition in Schooling Markets on Leadership for Learning

--Manuscript Draft--

Manuscript Number:	
Full Title:	The Effects of Competition in Schooling Markets on Leadership for Learning
Article Type:	Original Article
Corresponding Author:	Marcus Pietsch Leuphana Universitat Luneburg GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Leuphana Universitat Luneburg
Corresponding Author's Secondary Institution:	
First Author:	Marcus Pietsch
First Author Secondary Information:	
Order of Authors:	Marcus Pietsch
	Sebastian Leist
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	In the last decades German governments introduced market mechanisms into the education system by enhancing parental choice, abolishing schools' catchment areas and obliging schools to find, define and serve niches. At the core of schooling markets are choice policies, which give families the freedom of choice between individual schools and thus, create market incentive mechanisms. A basic assumption of school choice policies is that expanded school choice will generate more competition for schools, thereby increasing accountability and eventually school quality. Hence, upholding their school's competitive capacity has become a continuing concern of school leaders in German speaking countries and thus, competition is expected to have verifiable effects on the leadership behavior of principals. This study examines the associations of competition between schools and the perceived leadership behavior of principals in the federal state of Hamburg, Germany using data from n=3,950 teachers within n=74 secondary schools. Based upon latent network analysis, local schooling markets were estimated and related through multi-level structural equation models to teacher survey data. Findings suggest that strong competition between secondary schools in the federal state of Hamburg exists and that competition has major impacts on the leadership behavior of principals, even under control for social context.

The Effects of Competition in Schooling Markets on Leadership for Learning

Corresponding Author:

Dr. Marcus Pietsch, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany
Fon: 04131.677-203, e-mail: pietsch@leuphana.de, ORCID: http://orcid.org/0000-0002-9836-6793

Co-Author:

Sebastian Leist, Department of Education and Training Victoria, 2 Treasury Place, East Melbourne VIC 3002, Australia, Fon: 0475 072 912, e-mail: leist.sebastian.a@edumail.vic.gov.au,
ORCID: https://orcid.org/0000-0002-0885-3054

Abstract

In the last decades German governments introduced market mechanisms into the education system by enhancing parental choice, abolishing schools' catchment areas and obliging schools to find, define and serve niches. At the core of schooling markets are choice policies, which give families the freedom of choice between individual schools and thus, create market incentive mechanisms. A basic assumption of school choice policies is that expanded school choice will generate more competition for schools, thereby increasing accountability and eventually school quality. Hence, upholding their school's competitive capacity has become a continuing concern of school leaders in German speaking countries and thus, competition is expected to have verifiable effects on the leadership behavior of principals. This study examines the associations of competition between schools and the perceived leadership behavior of principals in the federal state of Hamburg, Germany using data from n=3,950 teachers within n=74 secondary schools. Based upon latent network analysis, local schooling markets were estimated and related through multi-level structural equation models to teacher survey data. Findings suggest that strong competition between secondary schools in the federal state of Hamburg exists and that competition has major impacts on the leadership behavior of principals, even under control for social context.

Keywords: Competition, Leadership, Learning, Markets, Schools

Die Auswirkungen schulischen Wettbewerbs auf lern-zentriertes Schulleitungshandeln (Leadership for Learning)

Zusammenfassung

In den vergangenen Jahrzehnten haben verschiedene Regierungen in Deutschland Wettbewerbsmechanismen in die Bildungssysteme der Länder implementiert, und beispielsweise Schulwahlmöglichkeiten erweitert, Schuleinzugsgebiete abgeschafft und Schulen dazu verpflichtet wurden, sich inhaltlich zu profilieren. Den Kern schulischer Märkte bilden ‚choice policies', die Familien freie Auswahl von Schulen ermöglichen und somit Anreize im Stile ökonomischer Märkte schaffen. Grundannahme erweiterter Schulwahlmöglichkeiten ist dabei, dass diese Wettbewerb zwischen Schulen fördern, Transparenz und Rechenschaftsdruck erhöhen und letztendlich die schulische Qualität anheben. Folglich ist es für Schulleitungen in deutschsprachigen Ländern ein

fortwährendes Anliegen, die Wettbewerbsfähigkeit ihrer Schule aufrechtzuerhalten. Entsprechend wird davon ausgegangen, dass Wettbewerb nachweisbare Auswirkungen auf Schulleitungshandeln hat. Die vorliegende Studie untersucht daher die Zusammenhänge des Wettbewerbs zwischen Schulen und dem von Lehrkräften wahrgenommenen Schulleitungshandeln in der Bundesland Hamburg, basierend auf Daten von n=3.950 Lehrern an n=74 Sekundarschulen. Mittels latenter Netzwerkanalysen werden in einem ersten Schritt schulische Märkte ermittelt, die in einem zweiten Schritt mithilfe von Mehrebenenstrukturgleichungsmodellen mit den Fragebogendaten zum durch Lehrkräfte wahrgenommenen Schulleitungshandeln in Beziehung gesetzt werden. Die Ergebnisse weisen auf ausgeprägten Wettbewerb zwischen den weiterführenden Schulen in Hamburg und einen bedeutenden Einfluss des schulischen Wettbewerb auf das Schulleitungshandeln hin, dies selbst unter Kontrolle sozialer Rahmenbedingungen.

Schlüsselwörter: Wettbewerb, Leadership, Lernen, Märkte, Schulen

The Effects of Competition in Schooling Markets on Leadership for Learning

1. Introduction

In the last two decades German governments increasingly introduced market mechanisms into the education system by enhancing parental choice, abolishing schools' catchment areas, publishing information on school quality and obliging schools to find, define and serve niches with regard to educational content and student characteristics (Thiel, Cortina, & Pant, 2014). The soaring of schools' market-orientation was attended by the implementation of policies for evidence-based governance, a push for greater school autonomy and for increased means regarding school-based management and leadership (Altrichter, Rürup, & Schuchart, 2016).

At the core of schooling markets in Germany are choice policies, which give students and parents the freedom of choice between individual schools and thus, to create a market incentive mechanism. Hence, a basic assumption of school choice policies is that expanded school choice will generate more competition for schools, thereby increasing accountability and eventually school quality (Unger, 2015). Thus, school choice is intended not only to serve families who actively choose; it also is anticipated to introduce (possibly as a side-effect) market pressure into unresponsive schools and thereby to improve the quality of education for all students (Hoxby, 2000, 2003).

With regard to school improvement and change, principals are often described as key actors in the education policy rhetoric. School leadership is regarded as crucial for improving school performance in terms of student learning and achievement, equity and the capacities of teachers (Leithwood, Harris, & Hopkins, 2008). Thus, following the recent educational reforms, upholding their school's competitive capacity has become a continuing concern of school leaders in German speaking countries (Kanape-Willingshofer, Altrichter, & Kemethofer, 2016). More precisely, Leithwood (2001, p. 221) summarizes the expectations of principals operating in the context of educational accountability policy regimes as follows:

"These leaders are able to market their schools effectively, develop good customer/client relations, and monitor 'customer' (student and parent) satisfaction. To prosper in such contexts, school leaders continuously redesign their organizations in response to fast changing market conditions. They collect data about competitors' services and prices, and find niches for their schools. They have exceptional levels of clarity about their missions because these missions are viewed as a central criterion in parent and student choices."

Thus, in terms of leadership behavior, practices or styles it is expected that increased competition between schools is associated with a more active leadership behavior, namely transformational and instructional leadership and a stronger focus on distributed or shared orientations to leadership (Leithwood, 2001), a triad of leadership practices to which educational scholars for quite some time now refer to as *Leadership for Learning* (LFL, Hallinger, 2011).

According to our best knowledge few studies to date have investigated the interrelationship of competition and the whole set of LFL practices of school principals. Neither for Germany nor on the international level does reliable empirical evidence yet exists with regard to the associations of both (sets of) variables. Thus, the aim of the following study is to examine the responsiveness of LFL to competition between schools for the first time based on a large data set. In the following sections we first provide a theoretical framework clarifying the concepts and potential associations of the

research variables. Second, we present the data used and our framework for research; a latent network and a latent multi-level technique. Third, we hypothesize that the LFL climate of schools will be contingent to the competitive setting and test this assumption by analyzing student tracking data ($n_{students}$=14,481) as well as teacher survey data ($n_{teachers}$=3,950, $n_{schools}$=74) collected in the federal state of Hamburg, Germany. The results are discussed in the context of current empirical findings as well as with regard to their implications for future research.

2. Leadership for Learning in the Context of Accountability and Competition

2.1 Competition in the Schooling Market

Within the discussion on school quality and student achievement school choice is regularly understood as a key driver for improvement and change. The core idea is that expanded school choice will generate more competition between schools leading to increased accountability and eventually school quality (Hoxby, 2003; Unger, 2015). Advocates of choice policies refer to economical theorists like Friedman (1962) who made the assertion that choice produces competition and competition in turn produces quality. Not an unjustifiable assumption, given that school choice in principle puts a new dynamism into education: Schools no longer can take their clientele for granted when parents and students can choose between different institutions and thus, have to compete for customers by providing (custom-fit) attractive high-quality offers. Despite greater freedom for parents and students and better matches between families and schools, this policy is anticipated to create a pressure to produce good education through a better allocation of resources, a cost-efficient education valued by students' families (Unger, 2015).

This voting by feet or Tiebout sorting (Hoxby, 2000) regularly takes place within local schooling markets, areas which are usually bordered by spatial barriers like rivers, motorways or railways (Author, blinded for Review) or the responsibilities of municipal authorities (Bradley, Crouchley, Millington, & Taylor, 2001). Thus, on the one hand it is only possible for families to choose between a limited number of schools as potential choices "are blocked for the vast majority of children by (…) physical space and the operations of the educational marketplace" (Raey, & Lucey 2003, p. 138). On the other hand the "demand for school places is inelastic; that is the number of potential students is fixed" (Gewirtz, Ball, & Bowe, 1995, p. 2). Hence, a small number of schools compete for a fixed number of students. Consequently, competition between schools therefore arises through attracting students away from other schools within the local schooling market.

Research has shown that within the schooling market oligopoly is the dominant form of competition (Adnett, & Davies, 2003). In this case the expected behavior of one school influences the behavior of all other local schools. Thus, the market-leading schools will aim to build upon their advantage and stay ahead of the competition. MacDonald (2006, p. 204) gives the following example: "For instance, if one school in an area added new specialized facilities, another school might follow suit to avoid losing students to the other school." Perfect competition between schools in turn is nearly non-existent and what comes closest to a perfect schooling market is monopolistic competition in second-best markets (Lubienski, 2006). Here schools provide differentiated facilities and services for students' families. Thus, if a school for example fills a specific educational niche, i.e. providing serious

sport or classical music facilities, the other schools within the local market will try to fill different niches.

2.2 Leadership for Learning (LFL)

Learning-oriented leadership has been identified as one of the most relevant correlates for school improvement and educational change (Boyce & Bowers, 2017; Hallinger, 2011). This is not surprising as, like Boyce and Bowers (2018, p. 198) point out, "The overarching goal of leadership for learning (…) is (…) to ensure that schools are collectively focused on what is necessary to ensure the success of their students". Thus, LFL is anticipated to stimulate, guide and monitor the learning on all levels of a school: the organizational learning, the professional learning of employees and the individual learning of students.

Indeed, learning-oriented principals focus on "school-wide alignment of all aspects of a school with instructional-centered leadership at its core" (Boyce, & Bowers 2018, p. 197). Hence, with regard to the content, the LFL approach is closely related to the idea of instructional leadership, but goes beyond its limitations, namely principal- and teaching-centering, by incorporating "a wider spectrum of leadership action to support learning and learning outcomes" (Bush, & Glover, 2014, p. 556) Consequently, the LFL model subsumes features of "instructional leadership, transformational leadership, and shared leadership" (Hallinger, 2011, p. 126) practices.

Further it is assumed that effective LFL is contingent to the setting (Hallinger, & Heck, 2010) and thus, principals behaviorally respond to the institutional (i.e., the structure of an education system), the community (i.e., the socio economic status (SES)), the cultural (i.e., the values shared within a cultural area), the economic (i.e., the economic development of a country), the political (i.e., the political aims of a society) and the school improvement context (i.e., the stage of school improvement, Hallinger, 2016). Hence, there is no single best leadership technique, but instead, different leadership skills can be an appropriate way to effectively handle certain contextual issues or situations.

2.3. Associations of Leadership for Learning and Competition

Not much is known about how principals respond to competition in the schooling market. In theory, competition should alter the behavior of all stakeholders involved in the education of students, including principals. Competition between schools should lead to more active and effective school leadership, because principals have greater control and more options for designing and managing the school (Cravens, Goldring, & Penazola, 2012).

Thus, with regard to LFL it is expected that increased competition between schools is positively associated with a more active leadership behavior, namely transformational, instructional leadership and a stronger focus on distributed or shared orientations to leadership (Leithwood, 2001). This holds especially true for instructional leadership practices, as, like Cravens, Goldring and Penazola (2012, p. 454) point out, "The autonomy and reduced bureaucracy of choice schools suggest that choice school leaders, compared with traditional public school principals, may be more likely to

attend to instructional leadership because they would be freed from administrative, compliance, and management tasks that often are required in complex, centralized organizations". Positive interrelations, thus, should be expected particularly with regard to the framing and communicating the schools' goals and its mission, as this is considered to be "the starting point for creating a learner-centered school" (Hallinger, & Wang 2015, p. 29).

Generally, the empirical evidence with regard to the effects of competition on the quality of schooling and teaching is mixed and vague at best. For instance, in a review of US studies Belfield and Levin (2002) summarized that the majority of studies showed a modest positive effect of competition on a range of outcomes, like educational efficiency and teaching quality, but that more than half of estimated effects were not statistically significant. The findings further suggest that competition may need to exceed a certain threshold in order to bring about an effect. Lubienski (2009) in contrast, synthesizing the evidence on educational innovations in competition-oriented markets from over 20 OECD and non-OECD countries, concludes that on the one hand market reforms appear to be more successful in creating innovations in the marketing and management of schools than in generating new classroom practices and that on the other hand there seems not to be a direct causal relationship between competition between schools and educational innovation and change. And Sahlgren (2013), by assessing the global evidence on the effects of school choice and competition on school quality, found that firstly, choice programmes have mixed impacts (i.e., positive effects in Sweden, no effects in the Netherlands), that secondly, even when positive effects are observable the effect size is small to moderate and that thirdly, the design of choice programmes is important for outcomes. Indeed, "very few methodologically persuasive studies find negative effects" (Sahlgren, 2013, p. 97).

With regard to the associations of competition and principal leadership the evidence (which is mainly available from the US) is even more sparsely and much vaguer. Whereas Cravens, Goldring and Panazola (2012) report that competition between schools is associated with a stronger focus by principals on instructional leadership practices and a broader involvement of teachers in decision-making processes and school management, Lubienski (2009) states, that principals of competing schools mainly try to shape the public image of their schools through marketing activities with less focus on process innovations. Further, Jennings (2010) reports that principals in a competitive market environment often try to actively manage the (de-)selection of students through cream-skimming and counseling out potential problem students. Thus, school leaders in fact tend to develop particular operational and organizational strategies based on the actions of their perceived rivals, not with regard to effectiveness but instead to equity.

3. Design & Method

3.1. Purpose

Against this background, the purpose of this research is twofold: First, the study aims to describe the education market for secondary schools in the federal state of Hamburg, Germany, and to develop competition measures based upon those analyses. Second, the study seeks to analyze the associations of competition between schools and the leadership climate of schools, namely LFL. With these aims outlined, the following research questions will be examined:

1) What do the education markets, the market areas, for public secondary schools in Hamburg, Germany look like?

2) How strong is the competition between public secondary schools in Hamburg, Germany?

3) Does the competition between schools affect LFL, namely the dimensions instructional, transformational and shared leadership, of secondary schools in Hamburg, Germany?

3.2. Sample

This study is a secondary analysis based on official student register data gathered by the Hamburg education authority for the school year 2014/2015 and teacher survey data collected by the Hamburg school inspection during the first five years of its second inspection cycle (2012-2019).

In the German federal state of Hamburg children are entitled to attend nursery education from one until the age of six and to attend pre-schooling at the age of five. Afterwards, primary education provides schooling from the age of six to ten. Following this, students are enrolled in the secondary schools at the lower secondary level (fifth grade) at an age of about eleven. Since 2010 two types of secondary schools exist in Hamburg[1]: grammar schools (Gymnasium) and comprehensive schools (Stadtteilschule). At the point of transition from primary to secondary schools students have a free choice to attend any secondary school they want within the borders of the federal state. Thus, a free school choice policy is in place for secondary schools which is attended by typical market incentives, like the legally mandated image management and marketing of schools (i.e., development and publication of school profiles and school programs) and the publication of school inspection reports. League tables of exam results and student achievement measures, however, are not publicly reported.

For modelling education markets and competition among schools this study uses student register data, which captures student transition from primary to secondary school in school year 2014/2015. The register contains only information about students attending a school within the area of responsibility of the Hamburg education authority, i.e. students that transition from a school or into a school beyond the administrative border are not part of the dataset. The data set represents transitions of 14,481 students between all sector primary and secondary schools (N=403) located in the federal state of Hamburg that year.

With regard to LFL and control variables this study uses school inspection data from n=74 secondary schools (n=43 upper secondary schools, “Gymnasium” in German and n=31 comprehensive schools, “Stadtteilschule” in German), which have been inspected between 2012 and 2017. The school sample in this study represents a proportion of 47 percent (N=156) of all secondary schools within the Hamburg school system. The teacher inquiry is conducted as a full population survey on the level of each school, thus, no sampling takes place and all teachers of a school are requested to participate in the online survey. Detailed in-depth information concerning the used measures and instruments can

[1] Special schools are an additional school type, which has minor importance in competition. Special schools serve students with special needs and show declining proportions of children since Germany’s commitment to the UN Convention on the Rights of Persons with Disabilities in 2009 (“Inclusion”).

be found in the methodological documentation of the Hamburg school inspection (Pietsch, Scholand, Graw, Hengstmann, & Kulin, 2013).

The teaching staff within a single secondary school ranges in size from 35 to 191 (mean=84). The response rate of the teachers in the online survey was 62.3 percent (n=3,950, N=6,352)

3.3. Measures and Instruments

Leadership for Learning (LFL), its respective leadership dimensions and facets are surveyed with three instruments:

Transformational leadership is surveyed with three scales from the Multifactor Leadership Questionnaire (MLQ, Bass, & Avolio, 1995). Firstly, teachers answer three items measuring idealized influence attributed ($\omega = .79$)[2], a leadership behavior in which teachers view their leader as charismatic and powerful (e.g. "The principal acts in ways that foster my respect."). Secondly, they answer three items indicating individual consideration ($\omega = .85$), a behavior focusing on the needs, wants and emotions of every single teacher while supporting and leading them individually (e.g. "The principal considers me as having different needs, abilities, and aspirations from others."). And thirdly, teachers responded to another three items capturing the facet inspirational motivation ($\omega = .80$), a leadership behavior which aims to motivate and inspire teachers by providing meaning and challenge to their work (e.g. "The principal articulates a compelling vision of the future.").

Instructional leadership is measured by applying three scales form the Teaching and Learning International Survey (TALIS, OECD, 2009), which are derivatives of the Principal Instructional Management Rating Scale (PIMRS, Hallinger, & Wang, 2015). Thus, the first facet, called management for school goals ($\omega = .73$), is similar to the first dimension of PIMRS in that it focusses on the framing and communication of school goals and the school's mission (Lee, Walker, & Chui, 2012, p. 589, e.g. "The principal ensures that teachers work according to the school's educational goals."). The other facets are overlapping with the second dimension of the PIMRS – Manages the Instructional Program – and partitions this dimension by proposing instructional management and direct supervision of instruction (Lee et al., 2012). Thus, the items of the second facet, which is called instructional management ($\omega = .68$), focus on principal's actions to improve teachers' instruction by encountering and solving instructional challenges and problems (e.g. "When a teacher has problems in his/her classroom, the principal takes the initiative to discuss the matter."). And the third facet, which is called direct supervision of instruction in the school ($\omega = .64$), refers to "actions to directly supervise teachers' instruction and learning outcomes" (OECD 2009, p. 194, e.g. "The principal or someone else in the management team observes teaching in classes.").

Shared leadership ($\omega = .73$) is measured by a three item scale which is a German translation by the Hamburg school inspection of items from a scale capturing shared leadership among principal and others which has been developed by Wahlstrom and Louis (2008). Thus, the construct follows the

[2] As this study works within a SEM-framework it reports McDonalds Omega instead of Cronbachs Alpha as the reliability coefficient. Applying this coefficient to parallel or tau-equivalent measurement will obtain an estimate of reliability equal to the coefficient alpha. If this coefficient is applied to congeneric measurement the estimated reliability will be higher than the coefficient alpha (Widhiarso & Ravand 2014). Thus, unlike coefficient alpha, coefficient omega is not a lower-bound estimate of internal consistency reliability.

idea that principals use less of their controlling power and are more willing to share their positional power with others through a participatory form of decision-making, eventually resulting in the empowerment of teachers (e.g. “Teachers have an effective role in school-wide decision making.”).

Schooling markets and competition are measured and modelled with the aforementioned data from the student register of the federal state of Hamburg. Specifically, this study used the variables of current school and year level, and previous school and year level to model schooling markets. Data used to calculate the Herfindahl index (Bellfield, & Levin 2002; Hoxby, 2000), a competition measure, draws from the same source and is based upon student enrolment data, i.e. the index is based on market share and thus, measures the concentration of students by schools in a specific schooling market.

Context control variables include the social index of a school, school type and school size. Those variables were chosen as other authors have shown that there “is almost nothing left for marketisation to explain” (Gorard, Taylor & Fitz, 2003, p. 186) when variables like the diversity of the local population, local levels of residential segregation, and school organization factors, such as the nature and the number of local schools, are considered as potential confounders in multivariate analyses.

Social index (Schulte, Hartig, & Pietsch, 2014) of a Hamburg school, which is unidimensional and covers facets like the socio-economic and the socio-cultural background of the students’ families and information on social indicators of the neighborhood at their families’ places of residence, i.e. the occupational classes of parents as defined within the Erikson–Goldthorpe–Portocarero (EGP, Erikson, & Goldthorpe, 1992) scheme and the local welfare rate. A lower index score indicates a high proportion of socially disadvantaged students within a single school.

School type is indicated by the two types of schools in our sample. The variable was dummy-coded.

School size is measured by two variables: first the total number of students enrolled in a school. Second the number of teachers working within a single school.

3.4. Analytic Strategy

3.4.1. Modelling Schooling Markets

Usually schooling markets are a priori defined within a geographical area (Hoxby, 2000). However, this approach depends entirely on prior assumptions about the markets and the appropriate units of choice (Gibbons, Machin, & Silva, 2008). To overcome this problem and potential fallacies, this research applied an empirical model which takes into consideration the observed market behavior of students and their families (Author, blinded for Review). The method we applied is a stochastic network analysis, in particular a *latent position and cluster model for statistical networks*, according to Hoff, Raftery and Handcock (2002) and Handcock, Raftery and Tantrum (2007). The modelling was carried out by the *R* package `latentnet` (Krivitsky, & Handcock, 2008, 2014).

The approach features generalized bi-linear mixed-effects models (GBME) that are extended by a finite mixture model, which in turn specifies the likelihood of observed data as a function of multiple groups (Templin, 2008). The expressed probability of a network g for a set of nodes (=schools) is a

product of dyad (=transition) probabilities. Each is a generalized linear model with this linear component:

$$\eta_{i,j} = \sum_{k=1}^{p} \beta_k X_{i,j,k} + d(Z_i, Z_j) + \delta_i + \gamma_j$$

$X_{i,j,k}$ is an array of dyadic covariates, β_K is a vector of coefficients of the covariates. Z_i and Z_j are the positions of the vertices i and j in the latent space. d is a function of 2 positions; either negative Euclidian $(-\|Z_i - Z_j\|)$ or bilinear $(Z_i * Z_j)$. δ_i and γ_j are vectors of sender- and receiver-effects (Krivitsky, & Handcock 2014).

The stochastic modelling was conducted in two subsequent steps. At first, latent clusters on the basis of all schools in the federal state of Hamburg were modelled, i.e. the entire network. Then the clusters were subdivided further. All transitions from primary to secondary school were integrated weighted by the numbers of students into the model, but a so-called *line-cut* (Templin, 2008) was applied beforehand to reduce the noise that is produced by the large number of *1-student transitions*, which account for 1,136 of the 2,644 transitions (data from 13,345 of the 14,481 students is retained). These *1-student transitions* are volatile from year to year, and neglected for this reason. The applied line-cut implies a so-called vertex-cut, because neglecting *1-student-lines*, transitions of only one student from a primary school to a secondary school, disconnects schools from the network of transitions.

Finally, we compared the different latent cluster market models using the Bayes Information Criterion (BIC), preferring models with lower values (Schwarz, 1978).

Based upon this information we calculated the market share and the Herfindahl index (HI) for each market, a measure that quantifies supplier concentration (Belfield, & Levin, 2002):

$$HI = \sum_{i=1}^{N} s_i^2$$

Here s_i represents the market share (=students enrolled) of supplier i on the market and N corresponds to the total number of suppliers (=schools). Possible Values fall between 1/N (lower boundary) and 1. Thus, the measure includes the market share as well as the density of competition. The closer the index is to one, the more concentrated is the market and the less competition occurs.

3.4.2. Multilevel Structural Equation Models

Leadership is a multi-level phenomenon per se (Bliese, Halverson, & Schriesheim, 2002). Thus, regarding questionnaire data on perceived leadership behavior of principals a proper disaggregation of the data is necessary as the individual level (level 1) reflects the effects of inter-individual differences in perceptions of a principal's leadership behavior whereas the school level (level 2) represent a climate or a context construct (Marsh et al., 2009, 2012). When the referent is the principal, the evaluation of the effects of leadership ideally should be based on the school level construct formed by the aggregation of perceptions by teachers, i.e. leadership climate, not on individual teacher-level responses (Bureau et al., 2017).

Thus, for estimating the effects of competition on LFL, the data was analyzed by applying multi-level structural equation models (ML-SEM) in MPLUS 7 (Muthén, & Muthén, 2012). More precisely a so called doubly latent structural equation approach was utilized (Marsh et al., 2009) which accounts for measurement error influences on the individual (within) as well as for measurement and sampling error influences on the school (between) level. As the association of interest is located at the school level, the modelling used the leadership facets on both levels and estimated latent regression on the school level to predict the effects of competition between schools on the leadership climate of schools while controlling for potential confounding context variables. An exemplary model of our approach is presented in figure 1.

Fig. 1: Example of the doubly-latent structural equation model

*** INSERT FIGURE 1 HERE ***

To assess the fit of the models the classic fit indexes comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) as provided by MPLUS are reported. Acceptable fit would be indicated by a CFI over .90, a RMSEA less than .08 and a SRMRw / SRMRb less than .08 (e.g. Hu, & Bentler, 1999; Marsh, Hau, & Wen, 2004).

3.4.3. Preliminary Verification of multi-level Assumptions

For checking if the chosen multi-level strategy is appropriate the intra-class correlations in terms of ICC(1) (i.e., the proportion of between-group variance to the total variance) and ICC(2) (i.e., the reliability of a schools mean) for all leadership measures were computed first. The estimation was done using the *R* package `multilevel` by Bliese (2016). Following the recommendations by Bliese (2000), ICC(1)-values greater than .05 and ICC(2)-values greater than .70 indicate that the given construct is a group-level construct, and thus, a multilevel modeling approach should be used.

Tab. 1: Estimates of reliability for leadership scales

	Mean	**SD**	**ICC1**	**ICC2**	**ω_{Level1}**	**ω_{Level2}**
Transformational Leadership						
Idealized Influence attributed	2.96	0.86	.281	.939	.785	.996
Individual Consideration	2.94	0.86	.134	.859	.847	.984
Inspirational Motivation	3.19	0.71	.284	.939	.797	.971
Instructional Leadership						
School Management	2.74	0.77	.143	.868	.725	.923
Instructional Management	2.62	0.80	.170	.889	.683	.906
Direct Supervision	2.02	0.72	.125	.848	.635	.889
Shared Leadership	2.98	0.69	.146	.870	.730	.877

Table 1 presents the means, standard deviations and ω-coefficients on the individual and the school level as well as ICC(1) and ICC(2) of the leadership scales. The results show that all leadership

dimensions are multilevel in nature as the ICC(1)- and ICC(2)-values exceed the cutoff criteria by a large margin. Between 12.5 and 28.4 percent of the variance of the leadership measures was attributable to school differences. Further the ICC(2) for the leadership measures ranged from .848 for the instructional leadership facet direct supervision to .939 for the transformational leadership facets idealized influence attributed and inspirational motivation, justifying all leadership constructs as school-level variables.

4. Results

4.1. Schooling Markets in the Federal State of Hamburg

To answer the first research question, the regional schooling markets of the federal state of Hamburg were estimated by applying the aforementioned stochastic network analysis. Due to vertex-cutting, five schools were disconnected from the network of transitions.[3]

The analysis of the comprehensive network unveiled two best fitting solutions: two and five regional markets ($BIC_{1\text{-}Group}$=9,388.232, $BIC_{2\text{-}Groups}$=9,209.964, $BIC_{3\text{-}Groups}$=9,308.676, $BIC_{4\text{-}Group}$s=9,255.886, $BIC_{5\text{-}Groups}$=9,219,397, $BIC_{6\text{-}Groups}$=9,282.293, the network density is reported in Appendix A). However, the BIC of the 5-group-solution was slightly higher, than the BIC of the 2-group-solution, but the strength of the 2-group-solution against the 5-group-solution is nearly non-existent. Thus, the group structure with the second-lowest BIC to align the number of groups to a previous model was selected, which was been estimated with data from school year 2011/12 (Author, blinded for Review) as the basis for further analyses.

In the next step the regional schooling markets were subdivided into local markets by applying the same method (as it is impossible to conduct integrative multi-level stochastic network analyses and group detection at the moment with R). This resulted in a model with 14 local schooling markets. These markets are shown in figure 3, the BICs for the models are reported in Appendix B. The regional schooling market of Altona/Eimsbütte*l*, is north of the Elbe river and in the western half of this area contains four local schooling markets (Eimsbüttel (Süd), Eimsbüttel (Nord), Elbvororte/Osdorf/Lurup, Altona (Ost)). The second largest regional schooling market, Wandsbek/Nord, is situated in the north-eastern areas of Hamburg and consists of four local schooling markets as well, namely Dulsberg/Bramfeld/Steilshoop, Rahlstedt/Jenfeld/Farmsen, Walddörfer and Langenhorn/Hummelsbüttel/Fuhlsbüttel. The third largest regional schooling market (Süderelbe), located in the southwest of the federal state has three local schooling markets, called Elbinseln, Harburg-Kern and Neugraben/Finkenwerder. The fourth largest regional schooling market (östl. Alster) is located close to the Alster Lake and consists of two local schooling markets named östl. Alster (Süd) and östl. Alster (Nord). It was not possible to subdivide the regional schooling market Bergedorf/Billstedt further.[4]

[3] Four of these schools are rather small private schools, where only few student transitions from primary to secondary school, and the fifth school is a governmental comprehensive school, which has been closed down.
[4] The regional markets Wandsbek/Nord and Altona/Eimsbüttel each contain one additional market, which were neglected. In both markets, schools that show poor connectivity to other schools accumulated (i.e. Special schools and small non-government schools).

Fig. 2: Map of local schooling markets in the federal state of Hamburg

*** INSERT FIGURE 2 HERE ***

These schooling markets are partially interconnected, but the student population is mostly retained within the borders of the specific schooling market. The retention rates for the local schooling markets range between 66 (östl. Alster (süd)) and 94 (Eimsbüttel Nord and Harburg Kern) percent. These retention rates appear to be influenced by the accessibility of adjacent schooling markets, i.e. that spatial barriers like motorways, airports, train tracks and waterways reduce accessibility, whereas spatial permeability, i.e. public transport routes, enhances accessibility. Thus, the lowest retention rates are present in schooling markets located in central areas of Hamburg, where few spatial barriers exacerbate mobility, i.e. schooling markets östl. Alster (Süd) and östl. Alster (Nord). In contrast, the highest retention rates are present in Eimsbüttel (Nord), an area bordered by an administrative border, train tracks, the Hamburg airport and a 4-lane ring road (Ring 2), and Harburg-Kern, which is a settlement core that is bordered again by the administrative border, train tracks and westwards, by a gap in the settlement structure.

It is noteworthy, that low retention rates are a sign of competition between schools overlapping borders of local schooling markets.

4.2. Competition between Schools in the Federal State of Hamburg

In considering the second research question, Herfindahl indexes and market shares for measuring the competition between schools in the federal state of Hamburg were calculated. The results are presented in table 2. The analyses show, that the market concentration is low in most of the local schooling markets in Hamburg, indicating that there is reasonable competition between schools. The Herfindahl indexes range from .07 to .22 (mean=.12), which is much lower than reported for US schools, for which the approximate average of competition has been reported with a value of .37 (Belfield, & Levin 2003).

Tab. 2: Estimates of competition for Hamburgs' local schooling markets

Local Schooling Market	**1/N**	**HI**	**Market Share of top 3 market-leading Schools**	**Student Retention Rate**
östl. Alster (Süd)	0.07	0.08	34%	66%
östl. Alster (Nord)	0.08	0.10	39%	69%
Bergedorf/Billstedt	0.07	0.08	29%	91%
Dulsberg/Bramfeld/Steilshoop	0.09	0.13	51%	75%
Rahlstedt/Jenfeld/Farmsen	0.06	0.07	27%	86%
Walddörfer	0.10	0.12	45%	80%
Langenhorn/Hummelbüttel/Fuhlsbüttel	0.10	0.13	49%	85%
Eimsbüttel Süd	0.08	0.10	36%	75%
Eimsbüttel Nord	0.10	0.13	47%	94%
Elbvororte	0.06	0.07	28%	90%
Altona-Ost	0.10	0.15	59%	87%

Elbinseln	0.13	0.16	62%	92%
Harburg-Kern	0.08	0.10	43%	94%
Neugraben/Finkenwerder	0.14	0.22	73%	92%

Nonetheless, three markets show moderate concentration, which points out to a potential oligopoly. Thus, the schooling market Altona-Ost (HI=.15), a rather small market in a densely populated area, Elbinseln (HI=.16), a market spatially enclosed by the Elbe river, and Neugraben/Finkenwerder (HI=.22), a schooling market that consists of two settlement cores in the south-west of Hamburg, show high market concentrations. These Schooling Markets share the characteristics the market shares of the market-leading schools (59%, 62%, 73%) and student retention rates (87%, 92%, 92%) are comparatively high. Thus, the potential oligopolies in those three schooling markets are only marginally affected by competition between schools that overlap the borders of the local schooling markets.

Overall these results indicate that there is reasonable competition for students between secondary schools in the federal state of Hamburg but in some regions oligopolistic tendencies, due to territorial exclusivities and restrictions, are observable.

4.3. Effects of Competition on Leadership Climate

In investigating the last research question multi-level structural equation models (ML-SEM) or more precisely doubly-latent models that control measurement error at the individual (Level 1) and the school level (Level 2) as well as sampling error in the aggregation of individual level 1 ratings to form level 2 leadership climate constructs (Marsh et al., 2012) were scrutinized. At the school level the Herfindahl index on the dependent variable was regressed and controlled for social composition (social index), school size and school type. For each of the seven leadership facets we estimated a separate model. As no regressions were estimated on the individual level, we only report school level regressions.

4.3.1. Effects of Competition on Instructional Leadership

With regard to the three instructional leadership facets a tendentially significant effect of competition on the management of school goals ($\beta_{\mathrm{Herfindahl,L2}} = -.196$) and a highly significant effect on instructional management ($\beta_{\mathrm{Herfindahl,L2}} = -.270$) (see tab. 3) was found. Thus, as a higher value of the index indicates increased market power, a higher concentration in the schooling market negatively correlates with the framing of a school's goals and the communication of a school's pedagogical goals as well as with active instructional management by a principal. Hence, higher competition among secondary schools aligns with a clearer academic mission and a stronger orientation toward the improvement of teachers' instruction. For all control variables no significant ($p<.10$) effects were found.

Tab. 3: Effects of competition on instructional leadership; standardized regression estimates on level 2, coefficients of determination on level 2, model fit indexes

	Management for School Goals		Instructional Management		Direct Supervision of Instruction	
	β	p	β	p	β	p
Herfindahl Index	-0.196	0.052	-0.270	0.015	-0.083	0.513
Social Index	-0.161	0.432	-0.164	0.435	-0.239	0.219
$N_{teachers}$	-0.170	0.449	-0.047	0.831	-0.284	0.282
$N_{students}$	0.191	0.451	0.034	0.881	0.254	0.328
School Type+	0.095	0.656	-0.072	0.728	-0.136	0.281
R^2	***0.054***		***0.084***		***0.045***	
x2; df; p	27,169; 11; <.05		15.537; 11; >.05		3.835; 5; >.10	
CFI	0.993		0.998		1.000	
RMSEA	0.021		0.011		0.000	
$SRMR_w$	0.001		0.000		0.000	
$SRMR_b$	0.005		0.034		0.025	

+ Reference Group = Upper Secondary School (Gymnasium)

In total between six and eight percent of the instructional leadership climate between schools could be explained by competition and the other chosen context variables as the coefficient of multiple determination (R^2) reveals.

4.3.2. Effects of Competition on Transformational Leadership

Regarding the effects of competition between secondary schools in Hamburg and transformational leadership practices (see tab. 4) a very small but significant effect on idealized influence attributed ($\beta_{Herfindahl,L2} = -.003$) only was found. Thus, principals working in more competitive environments are perceived as being slightly more confident and powerful by their employees than their colleagues in less competitive markets.

Tab. 3: Effects of competition on transformational leadership; standardized regression estimates on level 2, coefficients of determination on level 2, model fit indexes

	Individualized Influence attributed		Individual Consideration		Inspirational Motivation	
	β	p	β	p	β	p
Herfindahl Index	-0.003	0.025	-0.176	0.135	-0.170	0.116
Social Index	-0.079	0.146	-0.134	0.476	-0.285	0.096
$N_{teachers}$	0.000	0.417	-0.013	0.960	-0.253	0.280
$N_{students}$	0.000	0.896	0.035	0.894	0.339	0.145
School Type+	-0.072	0.227	-0.116	0.573	-0.038	0.847
R^2	***0.093***		***0.042***		***0.078***	

x2; df; p	35.488; 11; <.05	33.002; 11; <.05	35.611; 11; <.05
CFI	0.995	0.996	0.990
RMSEA	0.026	0.024	0.026
$SRMR_w$	0.001	0.001	0.001
$SRMR_b$	0.017	0.025	0.039

+ Reference Group = Upper Secondary School (Gymnasium)

The coefficients of multiple determination (R^2) of the models ranged between .042 (individual consideration) and .093 (idealized influence attributed). For all control variables no significant (p<.10) effects were found

4.3.3. Effects of Competition on Shared Leadership

Table 5 reports the results of the ML-SEM that was estimated for analyzing the relationship of school competition and shared leadership behavior of principals. As for instructional and transformational leadership behavior, a positive relation between competition and shared leadership ($\beta_{\mathrm{Herfindahl,L2}} = -.268$) was found. Thus, the higher the competition among schools, the more principals are willing to share their positional power with others through a participatory form of decision-making, eventually resulting in the greater empowerment of teachers.

Tab. 4: Effects of competition on shared leadership; standardized regression estimates on level 2, coefficients of determination on level 2, model fit indexes

	Shared Leadership	
	β	p
Herfindahl Index	-0.268	0.020
Social Index	-0.216	0.285
$N_{teachers}$	-0.227	0.301
$N_{students}$	0.282	0.203
School Type+	-0.135	0.151
R^2	***0.102***	
x2; df; p	18.438; 11; >.05	
CFI	0.997	
RMSEA	0.014	
$SRMR_w$	0.000	
$SRMR_b$	0.049	

+ Reference Group = Upper Secondary School (Gymnasium)

About 10 percent (R^2) of the variation of the shared leadership behavior could be explained by competition and further context control variables. For all control variables no significant (p<.10) effects were found.

5. Discussion

Public schools in Germany have been reformed to ensure less hierarchical educational accountability, moving away from a mainly centralized system of control. In such a context increasing choice in the market for schools is expected to lead to better school and teaching quality for all students – a rising tide that lifts all boats (Hoxby, 2003). The findings of our study show that the competition for students between secondary schools in the German federal state of Hamburg is very strong. In about 80 percent of the local schooling markets we found low Herfindahl indexes, indicating low market concentration and high competition between schools. Hamburgs' local schooling markets are often bordered by spatial barriers like motorways, airports, train tracks and waterways which reduce accessibility, whereas spatial permeability, i.e. public transport routes, enhances accessibility. However, with a mean of 84 percent the retention rate for the local schooling markets is considerably high – nearly eight out of ten students retain within the borders of the local schooling market after the transition from primary to secondary school.

Further we found that competition has a major impact on the leadership behavior of principals in German secondary schools: the stronger the competition between schools, the higher the leadership activities of principals. In doing so, the nature of school leadership varies directly with the level of competition, even when controlled for other potential contextual confounders like the socio economic status of students' families and school organization factors. What is striking is that all facets of LFL are positively associated with competition. Thus, in a competitive market environment principals more often frame and communicate school goals, provide better and more focused support for improving teachers' instruction and tend to share their decision-making power more frequently with their staff. Further, the stronger the competition, the more charismatic a principal is perceived by his or her employees. Thus, the LFL climate of a school as indicated by a principal's leadership behavior directly reflects a school's competitive context. Further, the social composition of students as well as characteristics of the local population do not play a vital role with regard to leadership in competitive schooling markets – a result that contradicts the well-established theoretical assumptions with regard to successful principal leadership in schools in challenging circumstances (Klein, 2018) and prior empirical findings from the UK on competition between schools (Gorard, Taylor, & Fitz, 2003), which pointed to the exact opposite.

Thus, at first glance, the findings regarding the competition in the local schooling markets are in contrast to the international empirical knowledge base. Often only moderate correlations between school variables and the competition between schools, which vanish when controlled for other contextual variables, are reported. In contrast this analysis reveals a strong competition between secondary schools for students in the German federal state of Hamburg and strong correlations of competition with educational leadership even when social contextual characteristics are considered. This warrants some further research.

One possibility is that these differences may be a result of the methodological approach. Researchers to-date have generally used geographical or administrative boundaries for defining local schooling markets, which entirely depend on the researchers' prior assumptions about the markets and the appropriate units of choice. This study, in contrast, applied a behavioral approach, which is based on observational data, and thus, does not rely on such assumptions. In that respect, an interesting topic for further research could be the comparison of behavioral and traditional geographical approaches

in the modeling of schooling markets. Another possible explanation may be that the observed differences may be a result of varying regional, national, cultural and/or administrative political contexts. As the findings within this research indicate that boundaries play a significant role for defining schooling markets – even though not exclusively – it stands to reason that these factors could explain, at least partly, the differing results. Hence, another prospect for further research is to investigate any differential effects by conducting cross-contextual studies.

In relation to the associations of LFL and competition, the results of this research align with the theoretical assumptions and the sparse empirical evidence available, in that the findings support the hypothesis that stronger competition leads to better school quality in terms of principal leadership. Hence, a schools' LFL climate is highly responsive to the competitive context. This is particularly remarkable as it has been demonstrated for the federal state of Hamburg that LFL in turn is positively associated with other learning-relevant within-school variables and outcomes, i.e. teaching quality and the innovation capacity of teachers (Author, blinded for Review), and student achievement (Author, blinded for Review). Thus, strong indication exists, that competition in local schooling markets may positively affect the effectiveness of schools through LFL – this assumption needs to be investigated further.

Conversely it is as yet unclear whether the positive effects of competition in the local schooling markets outweigh negative ones. As some scholars have noted (Bradley, & Taylor, 2002), competition between schools often is attended by negative side-effects, i.e. social segregation of students. This too is an effect that has been found and reported for the federal state on Hamburg (Author, blinded for Review). Thus, it has yet to be determined the complex interactions among competition between schools and the trade-off of the two poles of high-quality schooling: effectiveness and equity.

6. References

Adnett, N., & Davies, P. (2003). Schooling reforms in England: from quasi-markets to co-opetition? Journal of Education Policy, 18(4), 393-406.

Altrichter, H., Rürup, M., & Schuchart, C. (2016). Schulautonomie und die Folgen. In H. Altrichter & K. Maag Merki (Eds.), Handbuch Neue Steuerung im Schulsystem (pp. 107-149). Wiesbaden: VS.

Bass, B. M., & Avolio, B. J. (1995). MLQ Multifactor Leadership Questionnaire. Technical Report. Redwood City: Mind Garden.

Belfield, C. R., & Levin, H. M. (2002). The effects of competition between schools on educational outcomes: A review for the United States. Review of Educational Research, 72(2), 279-341.

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

Bliese, P. D. (2016). Multilevel Modeling in R (2.6)—A Brief Introduction to R, the multilevel package and the nlme package. Available at https://cran.rproject.org/doc/contrib/Bliese_Multilevel.pdf

Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods In leadership: The articles, the model, and the data set. The Leadership Quarterly, 13(1), 3–14.

Bureau, J. S., Gagné, M., Morin, A. J., & Mageau, G. A. (2017). Transformational Leadership and Incivility: A Multilevel and Longitudinal Test. Journal of Interpersonal Violence, https://doi.org/10.1177/0886260517734219.

Boyce, J., & Bowers, A. J. (2017). Toward an evolving conceptualization of instructional leadership as leadership for learning: Meta-narrative review of 109 quantitative studies across 25 years. Journal of Educational Administration, https://doi.org/10.1108/JEA-06-2016-0064.

Boyce, J., & Bowers, A. J. (2018) Different Levels of Leadership for Learning: Investigating Differences Between Teachers Individually and Collectively Using Multilevel Factor Analysis of the 2011-12 Schools and Staffing Survey. International Journal of Leadership in Education, 21(2), 197-225.

Bradley, S., Crouchley, R., Millington, J., Taylor, J. (2001). Testing for Quasi-Market Forces in Secondary Education. Oxford Bulletin of Economics and Statistics, 62, 357-390.

Bradley, S., & Taylor, J. (2002). The Effect of the Quasi–market on the Efficiency–equity Trade–off in the Secondary School Sector. Bulletin of Economic Research, 54(3), 295-314.

Bush, T., & Glover, D. (2014). School leadership models: what do we know? School Leadership & Management, 1–19.

Cravens, X. C., Goldring, E., & Penaloza, R. (2012). Leadership practice in the context of US school choice reform. Leadership and Policy in Schools, 11(4), 452-476.

Erikson, R., & Goldthorpe, J. H. (1992). The Constant Flux: A Study of Class Mobility in Industrial Societies. Oxford: Oxford University Press.

Fitz, J., Gorard, S., & Taylor, C. (2003). *Schools, markets and choice policies*. London: Routledge.

Friedman M. (1962). Capitalism and Freedom. Chicago: Chicago University Press.

Gewirtz, S., Ball, S. J., & Bowe, R. (1995). Markets, choice, and equity in education. Buckingham: Open University Press.

Gibbons, S., Machin, S., & Silva, O. (2008). Choice, competition, and pupil achievement. Journal of the European Economic Association, 6(4), 912-947.

Hallinger, P. (2011). Leadership for learning: lessons from 40 years of empirical research. Journal of Educational Administration, 49(2), 125–142.

Hallinger, P. (2016). Bringing context out of the shadows of leadership. Educational Management Administration & Leadership. https://doi.org/10.1177/1741143216670652

Hallinger, P., & Heck, R. H. (2010). Leadership for Learning: Does Collaborative Leadership Make a Difference in School Improvement? Educational Management Administration & Leadership, 38(6), 654–678.

Hallinger, P., & Wang, W.-C. (2015). Assessing instructional leadership with the principal instructional management rating scale. Cham: Springer.

Handcock, M. S., Raftery, A. E., & Tantrum, J.E. (2007). Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(2), 301-354.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. Journal of the American Statistical Association, 97, 1090-1098.

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers? American Economic Review, 90(5), 1209-1238.

Hoxby, C. (2003). School choice and school productivity (or could school choice be a rising tide that lifts all boats?). In C. Hoxby (Ed.), The economics of school choice (pp. 287-342). Chicago: University of Chicago Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55.

Jennings, J. L. (2010). School choice or schools' choice? Managing in an era of accountability. Sociology of Education, 83(3), 227-247.

Klein, E. D. (2018). Erfolgreiches Schulleitungshandeln an Schulen in sozial deprivierter Lage. Eine Zusammenschau zentraler Grundlagen und Befunde aus der nationalen und internationalen Bildungsforschung. Expertise im Auftrag der Wübben Stiftung. SHIP Working Paper Reihe, No. 02. Essen: Universität Duisburg-Essen. https://doi.org/10.17185/duepublico/45206

Kanape-Willingshofer, A., Altrichter, H., & Kemethofer, D. (2016). Accountability Policies and School Leadership in Austria: Increasing Competition and Little Accountability. In J. Easley II & P. Tulowitzki (Eds.), Educational Accountability: International Perspectives on Challenges and Possibilities for School Leadership (S.142-154). London: Routledge.

Krivitsky, P. N. & Handcock, M.S. (2008). Fitting position latent cluster models for social networks with latentnet. Journal of Statistical Software, 24(5), 1-23.

Krivitsky, P. N. & Handcock, M.S. (2014*). latentnet: Latent Position and Cluster Models for Statistical Networks.* The Statnet Project (http://www.statnet.org). R package version 2.4.2, https://CRAN.R-project.org/package=latentnet.

Lee, M., Walker, A., & Chui, Y. L. (2012). Contrasting effects of instructional leadership practices on student learning in a high accountability context. Journal of Educational Administration, 50(5), 586–611.

Leithwood, K. (2001). School leadership in the context of accountability policies. International Journal of Leadership in Education, 4(3), 217-235.

Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. School leadership and management, 28(1), 27-42.

Lubienski, C. (2006). School diversification in second-best education markets: International evidence and conflicting theories of change. Educational Policy, 20(2), 323-344.

Lubienski, C. (2009). Do Quasi-markets foster Innovation in Education? A Comparative Perspective. OECD Education Working Papers, 25. Paris: OECD.

MacDonald, J. (2006). The international school industry: Examining international schools through an economic lens. Journal of Research in International Education, 5(2), 191-213.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. Structural Equation Modeling: A Multidisciplinary Journal, 11(3), 320–341.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects. Educational Psychologist, 47(2), 106–124.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error. Multivariate Behavioral Research, 44(6), 764–802.

Muthén, L. K., & Muthén, B. O. (2012). Mplus software (Version 7). Los Angeles: Muthén & Muthén.

OECD. (2009). Creating effective teaching and learning environments : first results from TALIS. Paris: OECD.

Pietsch, M., Scholand, B., Graw, S., Hengstmann, E., & Kulin, S. (2013). Skalenhandbuch der Schulinspektion Hamburg. Fragebögen für Pädagoginnen und Pädagogen, Eltern und Schülerinnen und Schüler. Hamburg: Institut für Bildungsmonitoring und Qualitätsentwicklung.

Reay, D., & Lucey, H. (2003). The Limits of Choice. Children and Inner City Schooling. Sociology, 37(1), 121-142.

Schulte, K., Hartig, J., & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann & N. Maritzen (Eds.), Grundlagen für eine daten-und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (pp. 67–80). Münster: Waxmann.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

Sahlgren, G. (2013). Incentivising excellent: school choice and education quality. London: CMRE.

Templin, J. (2008). Methods for Detecting Subgroups in Social Networks. In T. Little (Ed.), *Modeling Dyadic and Interdependent Data in the Developmental and Behavioral Sciences* (309-334). New York: Routledge.

Thiel, F., Cortina, K. S., & Pant, H. A. (2014). Steuerung im Bildungssystem im internationalen Vergleich. Zeitschrift für Pädagogik, Beiheft 60, 123-138.

Unger, C. (2015). Wettbewerbssteuerung im Primarschulbereich. Wiesbaden: VS.

Wahlstrom, K. L., & Louis, K. S. (2008). How Teachers Experience Principal Leadership: The Roles of Professional Community, Trust, Efficacy, and Shared Responsibility. Educational Administration Quarterly, 44(4), 458–495.

Widhiarso, W., & Ravand, H. (2014). Estimating reliability coefficient for multidimensional measures: A pedagogical illustration. Review of Psychology, 21(2), 111-121.

Appendix A

Tab. A.1: Proportions of realized connections (Network Density) in the regional clusters

Cluster	Network Density	No. of Schools
Östliche Alster	5.2%	63
Bergedorf	8.8%	40
Wandsbek/Nord	2.9%	115
Altona/Eimsbüttel	2.3%	136
Süderelbe	6.3%	52

Appendix B

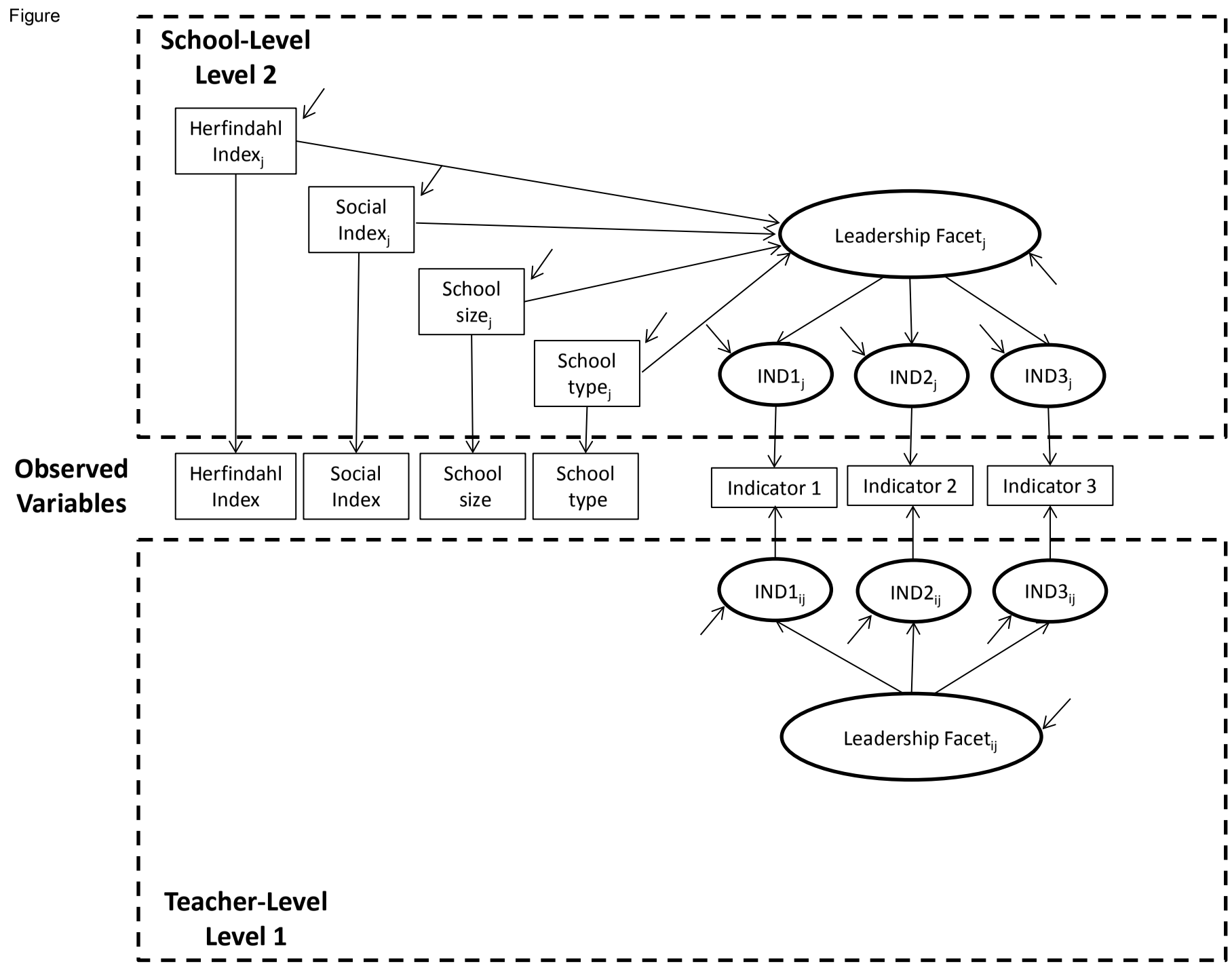
Tab. A.1: BICs for the estimation of local schooling markets within regional markets

Regional Schooling Market	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
Östliche Alster	927.7	892.4	931.4	916.2	949.6	921.0
Bergedorf	467.9	486.6	-*	-	-	-
Wandsbek/Nord+	2124.6	2114.4	2092.5	2086.8	2075.6	2111.5
Altona/Eimsbüttel+	1236.2	1232.0	1255.4	1238.6	1222.4	-*
Süderelbe	778.7	787.2	715.0	722.0	735.5	-*

* solution contained a cluster without schools assigned

+ One cluster was discarded due to a lack of connectivity to the rest of the Schooling Market. This cluster contains special schools and small non-government schools.

Figure


School-Level
Level 2
Herfindahl Index$_j$
Social Index$_j$
School size$_j$
School type$_j$
Leadership Facet$_j$
IND1$_j$
IND2$_j$
IND3$_j$
Observed Variables
Herfindahl Index
Social Index
School size
School type
Indicator 1
Indicator 2
Indicator 3
IND1$_{ij}$
IND2$_{ij}$
IND3$_{ij}$
Leadership Facet$_{ij}$
Teacher-Level
Level 1

Figure

Local Schooling Markets in Hamburg 2014

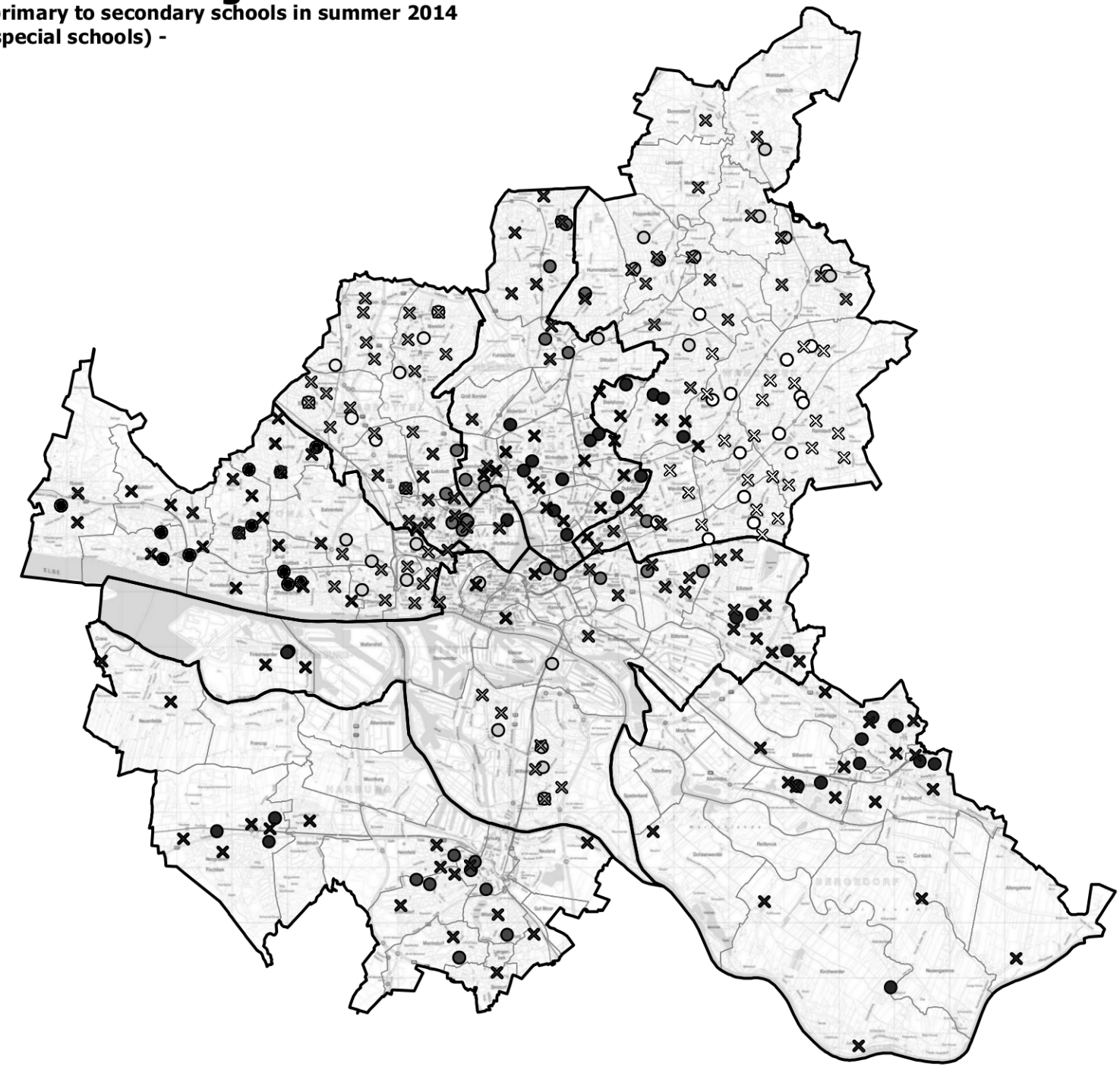
based on flows of students moving from primary to secondary schools in summer 2014
- only public schools are displayed (excl. special schools) -

Locations of secondary schools

- östl. Alster (Süd)
- östl. Alster (Nord)
- Bergedorf/Billstedt
- Dulsberg/Bramfeld/Steilshoop
- Rahlstedt/Jenfeld/Farmsen
- Walddörfer
- Langenhorn/Hummelsbüttel/Fuhlsbüttel
- Eimsbüttel (Süd)
- Eimsbüttel (Nord)
- Elbvororte/Osdorf/Lurup
- Altona (Ost)
- Elbinseln
- Harburg-Kern
- Neugraben/Finkenwerder

Locations

- Primary Schools



Clustering: Latent Cluster Position Model
(Krivitsky & Handcock 2009)
Line-cut <1 SuS

Spatial relations: without urban hinterland

Source: Statistics of School Year 2014/15
Data Status: 5th of September 2014
IfBQ, generated on 22nd of December 2015

0 1 2 4 6 8 10 12 km