Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Variabilität schulischer Lernumwelten

- Unterschiede in der Unterrichts- und Schulqualität zwischen Schulen und das Zusammenspiel mit Kontextmerkmalen und Schülerleistungen -

Dissertation
zur Erlangung des akademischen Grades
Doktorin der Philosophie (Dr. phil.)

vorgelegt von
Marina Wenger, M.A.

Berlin, 2020

Erstgutachter: Prof. Dr. Martin Brunner

Zweitgutachterin: Prof. Dr. Bettina Hannover

Weitere Kommissionsmitglieder:
Prof. Dr. Felicitas Thiel, PD Dr. Holger Gärtner, Dr. Karoline Koeppen

Tag der Disputation: 12.06.2020

Danksagung

Die Danksagung ist aus Gründen des Datenschutzes nicht enthalten.

Inhalt

Zusammenfassung

Die Schule stellt als institutionelle Bildungseinrichtung eine besonders relevante Lernumwelt von Kindern und Jugendlichen dar. Die schulische Lernumwelt beinhaltet das Lehren und Lernen im Unterricht und in der Schule unter spezifischen Rahmenbedingungen. Die Relevanz der Einzelschule zeigt sich in verschiedenen Forschungsperspektiven, beispielsweise in den Analysen Fends zu Einzelschulen als „pädagogische Handlungseinheit“ (Fend, 1986), in der Schuleffektivitätsforschung (vgl. bspw. Scheerens, 1990; Scheerens & Bosker, 1997), in Analysen der PISA-Daten (vgl. bspw. Baumert, Stanat & Watermann, 2006a; OECD, 2005) oder in der schulpolitischen Praxis im Rahmen von Schulinspektionen und Schulentwicklungsanliegen (vgl. bspw. KMK, 2015; Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013).

Überlegungen zur Qualität von Schulen basieren auf der Grundlage, dass sich die Bedingungen an Einzelschulen unterscheiden und auch verändern können. Diese Dissertation nutzte Erkenntnisse der Schuleffektivitätsforschung, der Lehr-Lern- und Schulqualitätsforschung um die *Variabilität* schulischer Lernumwelten zu ergründen. Dabei ist das Zusammenspiel schulischer Faktoren bedeutsam für die Analyse der Variabilität zwischen schulischen Lernumwelten. Als schulische Faktoren wurden Kontextmerkmale, Prozessmerkmale (Merkmale der Unterrichts- und Schulqualität) sowie als Output die Schulleistungen betrachtet (in Anlehnung an Ditton, 2000; Kunter & Voss, 2011; Scheerens, 1990).

Diese Arbeit untersuchte zur Variabilität schulischer Lernumwelten folgende zwei übergreifende Fragestellungen in drei empirischen Studien:

(1) Wie unterschiedlich sind Schulen in ihrer Unterrichtsqualität? (Studie 1)

(2) Unterscheiden sich schulische Prozessmerkmale und Outputs je nach Kontext?

   a) Kontext im Hinblick auf Schülerzusammensetzung (Komposition; Studie 2)

   b) Kontext im Hinblick auf bildungspolitische Intervention durch die Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ (Studie 3)

Mittels Daten der internationalen PISA-Erhebungen (Studie 1), sowie einer Kombination aus Daten der amtlichen Statistik, Vergleichsarbeiten und Schulinspektionsdaten (Studie 2 und 3)

wurden die Forschungsfragen analysiert. In Studie 1 wurde ein systematischer Überblick der Übereinstimmung, Variabilität und Reliabilität von Merkmalen der Unterrichtsqualität aus Sicht von Schülerinnen und Schülern auf Schulebene erstellt: es ergab sich eine moderate bis starke Übereinstimmung in den Schulen, „kleine“ bis „mittlere“ Unterschiede (LeBreton & Senter, 2008) zwischen den Schulen sowie eine unzureichende Reliabilität aggregierter Schülerurteile. In Studie 2 zeigte sich, dass Kompositionsmerkmale (mittlerer SES, mittlere Schulleistung) in positivem Zusammenhang mit Unterrichtsqualität stehen: Je höher der mittlere SES oder die mittlere Schulleistung, desto höher fiel die Klassenführung und kognitive Aktivierung (als Merkmale der Unterrichtsqualität) aus. Ein Zusammenhang zwischen Kompositionsmerkmalen und Schulqualitätsmerkmalen war kaum vorhanden. Eine Mediation des Zusammenhangs der Kompositionsmerkmale mit den mittleren Schulleistungen (als Output) durch Unterrichts- oder Schulqualitätsmerkmale konnte kaum nachgewiesen werden. In Studie 3 zeigten sich in Folge der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ an diesen Schulen lediglich geringe Verbesserungen der Unterrichts- und Schulqualität und keine signifikanten Verbesserungen der mittleren Schulleistungen im Vergleich zu den Schulen ohne diagnostizierten Entwicklungsbedarf. Es ergaben sich jedoch Hinweise auf einen nicht-intendierten Nebeneffekt für die Entwicklung der Zusammensetzung der Schülerschaft: Der Anteil an Schülerinnen und Schülern mit Migrationshintergrund scheint sich in Folge der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ zu erhöhen.

Die Ergebnisse der drei Studien tragen zur Erklärung der Variabilität schulischer Lernumwelten bei. Die Erkenntnisse daraus werden hinsichtlich ihrer Bedeutung für die Forschung, die Bildungspolitik sowie für die Praxis diskutiert.

Summary

As an educational institution, the school represents a particularly relevant learning environment for children and young people. The school learning environment is comprised of all teaching and learning that, under specific conditions, takes place in the classroom and at the school. The relevance of the school can be seen from various research perspectives, for example, in Fend's analyses of schools as "pedagogical units of action" (Fend, 1986), in school effectiveness research (see, for example, Scheerens, 1990; Scheerens & Bosker, 1997), in analyses of PISA data (see, for example, Baumert et al, 2006; OECD, 2005) or in school policy practice in the context of school inspections and school development issues (see e.g. KMK, 2015; Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013).

Considerations about the quality of schools are based on the fact that the conditions at schools differ and are changeable. This dissertation used findings from school effectiveness research, teaching research and school quality research to investigate the variability of school learning environments. The interaction of school factors is important for the analysis of the variability between school learning environments. School's contextual characteristics, process characteristics (instructional and school quality) as well as school achievement as output, were considered as school factors (following Ditton, 2000; Kunter & Voss, 2011; Scheerens, 1990).

This study examined the variability of school learning environments in the following two research questions over three empirical studies:

(1) How much do schools vary in their instructional quality? (Study 1)

(2) Do school processes and outputs differ according to context?

a) Context with regard to composition of the student body (Study 2)

b) Context with regard to educational policy intervention through the school inspection diagnosis "significant development needs" (Study 3)

Using data from the international PISA surveys (Study 1), as well as a combination of data from official statistics, large-scale achievement data and school inspection data (Studies 2 and 3), the research questions were analysed. Study 1 provided a systematic overview of the agreement, variability, and reliability of characteristics of instructional quality from the perspective of students at the school level: there was moderate to strong agreement within schools, "small" to

"medium" differences (LeBreton & Senter, 2008) between schools, and insufficient reliability of aggregated student ratings. Study 2 showed that compositional characteristics (mean SES, average school achievement) are positively related to instructional quality: the higher the mean SES or average school achievement, the higher the classroom management and cognitive activation (as components of instructional quality). There was hardly any correlation between compositional characteristics and school quality components. Instructional and school quality components did not significantly mediate the effects of compositional characteristics on average school achievement (as output). In Study 3, as a result of the school inspection diagnosis " significant development needs" these schools showed only slight improvements in instructional and school quality and no significant improvements in average school achievement compared with schools without a diagnosed need for development. There was, however, an unintended side effect on the development of the composition of the student body: The proportion of students with a migration background appears to have increased as a result of the school inspection diagnosis of "significant development needs".

The results of the three studies contribute to explaining the variability of school learning environments. The findings are discussed in their relevance for research, educational policy and practice.

Einleitung

„Das System Schule zeichnet sich kulturübergreifend durch formale, soziale und inhaltliche Merkmale aus, die Schule zu einer einzigartigen Lernumwelt machen.“ (Kunter & Baumert, 2008, S. 528)

Schülerinnen und Schüler haben unterschiedliche Voraussetzungen, wenn sie ihre Schullaufbahn beginnen und erfahren unterschiedliche Lebenswelten in ihren Familien und Peergroups. Von besonderer Bedeutung für die Kompetenz- und Persönlichkeitsentwicklung jedes Schülers und jeder Schülerin ist jedoch auch die Schule. Die Schule stellt eine wichtige Bildungsinstitution im Laufe des Lebens nahezu jedes Menschen dar. Schulische Lernumwelten bezeichnen dabei die institutionellen Bedingungen des Unterrichts und der Schule unter ihren spezifischen Rahmenbedingungen. Innerhalb der Forschung wandelte sich die Wahrnehmung der Bedeutung der einzelnen Schule und ihrer Lernumwelt jedoch. Fragen nach der Relevanz der Einzelschule wurden u.a. in der Schuleffektivitätsforschung bereits in den 1970er-Jahren gestellt. Unterschiedliche Forschungsfelder haben sich seitdem mit dem Thema der Einzelschule und ihrer Lernumwelt auseinandergesetzt. In einigen Fallstudien wurden Schulen analysiert und die Unterschiede zwischen ihnen herausgearbeitet (bspw. Fend, 1986). In großen Schulleistungsstudien wurden Unterschiede zwischen Schulformen und Einzelschulen für die Leistungsentwicklung in Deutschland betrachtet (Baumert et al., 2006a). Der Forschungsstrang der Schulentwicklungsforschung setzt sich mit Möglichkeiten der Weiterentwicklung einzelner Schulen auseinander (Rolff, 2010). Politisch wurde die Einzelschule im Zuge der „Neuen Steuerung“ im Zusammenhang mit höherer Autonomie sowie der Outputsteuerung durch Leistungsdaten interessant. Auch methodisch führten Neuerungen wie die Mehrebenenanalyse zu einer Auseinandersetzung mit der Schulebene, da diese dadurch getrennt von der Klassen- und Individualebene betrachtet werden konnte. Nicht zuletzt werden jährlich Schulen mit dem „Deutschen Schulpreis“ für ihre qualitätsvolle Schulpraxis und für innovative Konzepte ausgezeichnet (Robert Bosch Stiftung GmbH). All die angesprochenen Bereiche ordnen sich nicht unter einer gemeinsamen Herangehensweise oder übergeordneten

Theorie zur Einzelschule und ihrer Lernumwelt ein, beschäftigen sich aber dennoch alle mit dem Thema schulische Lernumwelt. Alle Fragen und Überlegungen zum Lernen in Schulen und zu „guter“ Qualität von Unterricht und Schule gehen davon aus, dass das Lernen und die Bedingungen des Lernens *nicht* überall gleich sind. Zudem werden Unterschiede zwischen Schulen angenommen, da sich die Leistungen von Schülerinnen und Schülern unterscheiden. Man nimmt also an, dass sich Lehren und Lernen durch Einflüsse oder bestimmte Gegebenheiten von innerhalb oder außerhalb der Schulen verändern können. Damit sind zwei Grundannahmen hinsichtlich der Variabilität von Schulen formuliert: Es bestehen (1) Unterschiede zwischen schulischen Lernumwelten und (2) schulische Lernumwelten sind veränderbar. Die Variabilität des Unterrichts und der Schule stand bislang wenig im Mittelpunkt der Forschung. Variabilität umfasst zum Einen Fragen der Variabilität zwischen Schulen in ihrem Lehren und Lernen sowie zum Anderen die Variabilität von Schulen *durch* ihre jeweiligen Kontextbedingungen. In diesem Rahmen ergeben sich aus unterschiedlichen Perspektiven bzw. unterschiedlichen Interessensträgern mehrere zentrale Fragen: Aus bildungspolitischer Perspektive ist interessant, welche Unterschiede zwischen Schulen bestehen und wie diese bei Steuerungsentscheidungen adressiert werden können, beispielsweise wenn politische Reformen zur Verbesserung der Unterrichtsqualität etabliert werden sollen. Aus Forschungsperspektive ist ebenfalls interessant, ob die Schulebene eine relevante Analyseebene darstellt. So ist es beispielsweise für die Planung groß-angelegter, clusterrandomisierter Interventionsstudien bedeutsam zu wissen, wie unterschiedlich Schulen in ihren Bedingungen überhaupt sind, um im Rahmen von statistischen Poweranalysen die notwendige Anzahl von Schulen zu bestimmen, damit Interventionseffekte zufallskritisch abgesichert werden können. Aus Perspektive der Schulpraxis ist es ein interessantes Feld, beispielsweise, wenn interne und externe Evaluation von Schulen geplant wird.

Diese Dissertation nähert sich dem Thema der Variabilität schulischer Lernumwelten in drei empirischen Studien. Diese drei Studien ordnen sich jeweils im Kontext verschiedener Forschungsfelder ein, die eingangs umrissen wurden. In Abbildung 1 wird, angelehnt an Schuleffektivitätsmodelle die Qualitätsbedingungen von Schulen beschreiben, dargestellt,

anhand welcher Merkmale die Variabilität schulischer Lernumwelten in der vorliegenden Arbeit untersucht wird: Kontextmerkmale, Prozessmerkmale sowie Output von Schulen.

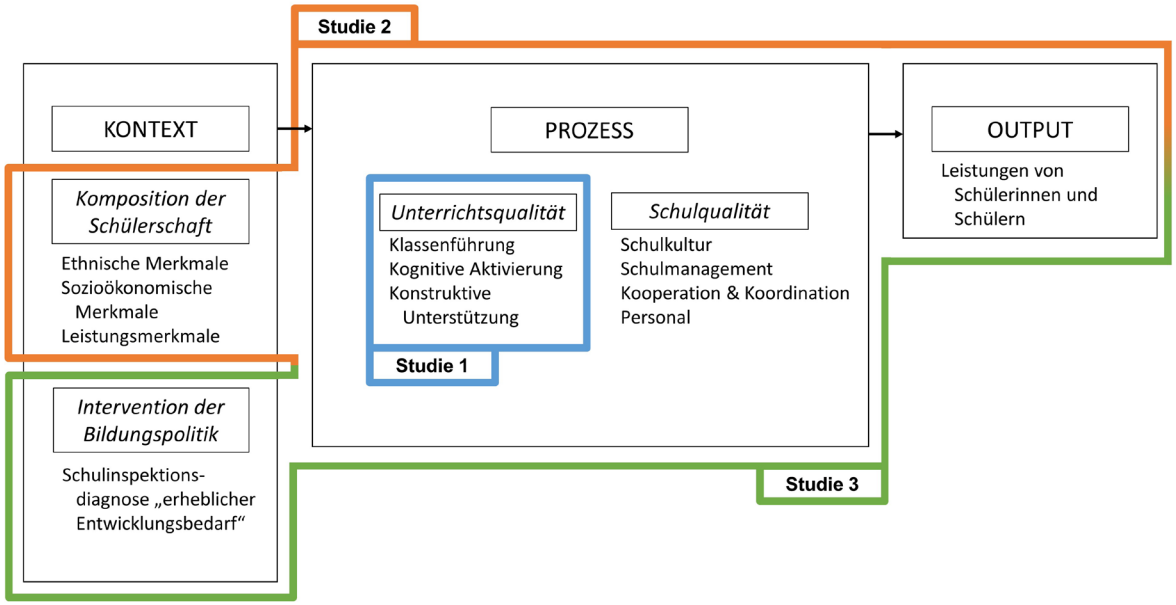


Abbildung 1: Einbettung der Studien der Dissertation in Schuleffektivitätsmodelle (angelehnt an Ditton, 2000; Kunter & Voss, 2011; Scheerens, 1990)

Studie 1 fokussiert dabei die Variabilität von Unterrichtsqualität als Teil schulischer Prozessmerkmale. In Studie 2 werden Merkmale des Schulkontexts (Komposition der Schülerschaft) mit Prozessmerkmalen (Unterrichts- und Schulqualität) sowie den Leistungen von Schülerinnen und Schülern[1] als Output verknüpft. In Studie 3 wird wiederum die Intervention der Bildungspolitik als Kontextmerkmal mit schulischen Prozessmerkmalen (Unterrichts- und Schulqualität) einerseits und Leistungen von Schülerinnen und Schülern als Output der Schulen andererseits verknüpft.

Als übergreifende Leitthemen dieser Arbeit sind daher die folgenden zu benennen:

(1) Schule als relevante Analyseeinheit

(2) Die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität sowie für Schulleistung.

[1] Leistungen von Schülerinnen und Schülern werden in den Dissertationsstudien aggregiert auf Schulebene betrachtet und daher auch als „Schulleistung“ bezeichnet.

Dabei wird in dieser Dissertation wie folgt vorgegangen: In Kapitel 1 wird dargestellt, wie die Relevanz der Einzelschule (oder die Schul*ebene*) für Lehren und Lernen in bisheriger Forschung thematisiert wurde (vgl. bspw. Baumert et al., 2006a; Fend, 1977, 1986; KMK, 2004, 2015; OECD, 2005; Sammons, Hillmann & Mortimore, 1995; Scheerens, 1990; Wurster & Gärtner, 2013; Wurster & Feldhoff, 2019). In Kapitel 2 wird das Zusammenspiel unterschiedlicher Qualitätsfaktoren von Schule thematisiert. Dies ist bedeutsam um Variabilität zwischen Schulen zu analysieren, da die Variabilität an verschiedenen Stellen des Systems Schule entstehen kann. Ein Rahmenmodell, welches Qualitätsfaktoren von Schule beschreibt, lässt sich in der Schuleffektivitätsforschung finden. Diese orientiert sich an der Effektivität der Schule im Sinne der Leistungen von Schülerinnen und Schülern. Zudem werden Aspekte der Lehr-Lernforschung beschrieben, die qualitätsvollen Unterricht bedingen. In Kapitel 3 wird darauf aufbauend ein übergreifendes Modell dargestellt, welches sowohl die Faktoren der Schuleffektivitätsforschung als auch die der Lehr-Lernforschung berücksichtigt (siehe Abb. 1). Anhand dieses Modells werden dann die übergreifenden Forschungsfragen dieser Dissertation herausgearbeitet und die Operationalisierung der Forschungsfragen durch die drei empirischen Studien dieser Dissertation beschrieben. Kapitel 4 präsentiert drei Manuskripte zu den empirischen Studien der Dissertation. In Kapitel 5 schließt sich die Gesamtdiskussion an: Zunächst werden die Befunde der einzelnen Studien in Bezug auf das Rahmenmodell (Abb. 1) zusammengefasst, bevor studienübergreifende Ergebnisse und Limitationen diskutiert werden. Darauf basierend werden Implikationen für die Forschung, für die Bildungspolitik und die Praxis abgeleitet, bevor die Gesamtdiskussion mit einem studienübergreifenden Fazit schließt.

1 Schulische Lernumwelten – Relevanz der Einzelschule

Was kennzeichnet schulische Lernumwelten? In diesem Kapitel soll die Relevanz sowie die Besonderheit der einzelnen Schule als Lernumwelt dargestellt werden, um herauszuarbeiten, welches Verständnis schulischer Lernumwelten dieser Dissertation zugrunde liegt. In welche Lernumwelt begeben sich Schülerinnen und Schüler in „ihrer" Schule? Was sind die prägenden Eigenschaften, die diese Lernumwelt bedingen? Zur Beantwortung dieser Fragen ist es relevant, sich anzusehen, wie sich bisher in der Literatur mit Schulen als Gesamtinstitutionen auseinandergesetzt wurde. Variabilität schulischer Lernumwelten beinhaltet, dass die Schule als Analyseeinheit relevant ist, zum Beispiel im Sinne von Unterschieden zwischen Lernumwelten. Die Relevanz der Einzelschule oder die Schule als einzelne Gestaltungsebene wurden in der Literatur bereits vielfach thematisiert. Daher sollen die Gedanken, mit denen sich die Autorinnen und Autoren bzw. Forschungsstränge in diesem Zusammenhang auseinandersetzten, hier zusammenfassend vorgestellt und eingeordnet werden. Ein übergreifendes Fazit zu den Befunden aus Kapitel 1 und 2 erfolgt in Kapitel 3.

1.1 Einzelschule als pädagogische Handlungseinheit (nach Fend)

> „Wer über Schulqualität spricht, macht zumindest eine wichtige Unterstellung: er nimmt an, daß es auf der Ebene der einzelnen Schule Vorgänge gibt, die positive oder negative Auswirkungen haben, welche nicht so sehr vom einzelnen Lehrer oder vom Schulsystem insgesamt ausgehen und welche nur auf dieser Ebene erreichbar sind." (Fend, 1988, S. 537)

Wenn von der Relevanz der Einzelschule die Rede ist, müssen die Arbeiten Fends aufgeführt werden, da diese den Begriff der Einzelschule als „pädagogische Handlungseinheit" (Fend, 1986) prägten. Dies beschreibt die Schule als Gesamteinheit, in der die Schulleitung und die Lehrkräfte einen „korporativen Akteur" bilden und die Vorgaben und Erwartungen des übergeordneten Systems und der Bildungspolitik für ihre jeweiligen Bedingungen der Einzelschule „rekontextualisieren" müssen (Fend, 2008). Das Konzept der Schule als pädagogische Handlungseinheit (Fend, 1986) entstand vor allem aus seiner Untersuchung von Schulen im Rahmen der „Drei-Länder-Studie" (Hessen, Nordrhein-Westfalen, Niedersachsen)

und aus der bildungspolitischen Situation der 1970er Jahre. Bildungspolitisch orientierte man sich zu diesem Zeitpunkt vorrangig auf das System als Ansatzpunkt für politische Reformen und glaubte, vor allem darüber Einfluss auf die spätere Lernentwicklung einzelner Schülerinnen und Schüler nehmen zu können (Wurster & Feldhoff, 2019 in Bezug auf Messner, 2016). Aus den zahlreichen Erfahrungen bei der Untersuchung einzelner Schulen entstand bei Fend das Bild der vielen Besonderheiten und Variationen jeder Schule, jedes Schulklimas und dessen Wirkung für die Schülerinnen und Schüler dieser Schule (Fend, 1977). Er stellt heraus, dass auch unter den gleichen organisatorischen Rahmenbedingungen unterschiedliches Unterrichten und unterschiedliche Prozesse in Schulen entstehen können (Fend, 2008). Zudem stellt er die systematische Beschreibung schulischer Lernumwelten aufgrund ihrer hohen Komplexität als problematisch dar (Fend, 1977). Er verweist daher auf die Notwendigkeit einer „inhaltlich-theoretischen Konzeption dieser Umwelten“ (Fend, 1977, S. 231) und beschreibt folgende relevante Merkmale: materielle und personelle Ausstattung, strukturelle und organisatorische Rahmenbedingungen, kulturelle Vorstrukturierungen (Curricula), Merkmale der in einer Institution lebenden Personen, Merkmale des sozialen Interaktionssystems Schule (Fend, 1977). In seiner Arbeit „Gute Schulen – schlechte Schulen“ versucht er sich einer Theorie der Schule als pädagogische Handlungseinheit weiter anzunähern, indem er einen Extremgruppenvergleich nutzt, um die Variabilität schulischer Lernumwelten aufzuzeigen (Fend, 1986). Wurster und Feldhoff (2019) beschreiben, dass sowohl aus der heutigen Sicht (methodische Kritik an Extremgruppenvergleichen, bspw. Creemers, Kyriakides & Sammons 2010) als auch aus der damaligen Sicht heraus (Messner, 2016) Kritik an Fends Vorgehen geübt wurde. Jedoch beschreibt Fend bereits 1988 in seiner „umfassenden Überarbeitung des Konzepts“ (Wurster & Feldhoff, 2019, S. 26) dass Qualitätskriterien von Lernumwelten mehrebenenanalytisch auf vier Ebenen angeordnet werden können: Schulsystem, Schulebene, Klassenebene, Personenebene (Lehrpersonen und Schülerschaft; Fend, 1988). Trotz der Kritik an Fends Vorgehen sind seine Arbeiten zur Relevanz der Einzelschule und der Bedeutung schulischer Qualität auf verschiedenen Ebenen bedeutsam für andere Forschungsentwicklungen im deutschsprachigen Raum. Nach Messner 2016 liegt das

besondere darin, dass Fend die Funktion der Schule im Zusammenspiel mit Unterricht und Kontext sieht (Messner, 2016; Wurster & Feldhoff, 2019).

1.2 Schulebene als relevante Gestaltungseinheit aus Mehrebenensicht

In Anknüpfung an das Konzept der 1970er Jahre von Fend, welches das Thema „Schule als pädagogische Handlungseinheit" hervorhob, nehmen Wurster und Feldhoff (2019) das Thema in einer Neuaufbereitung der Fendschen Daten aus der Drei-Länder-Studie und mit Hilfe moderner statistischer Verfahren wieder auf. Sie beschreiben, dass solchen Fragen nach der Relevanz der Ebene („Ist die Schule oder die Klasse die relevante pädagogische Gestaltungseinheit?", Wurster & Feldhoff, 2019) heute mit Mehrebenenanalysen und Intraklassenkorrelationen nachgegangen werden kann (Wurster & Feldhoff, 2019). In ihren Analysen nehmen sie zusätzlich zur Ebene der Schule, der Klasse und des Individuums (Schülerinnen und Schüler) auch die Ebene Schulform hinzu, die sie nach Fend dem System zuordnen. Die größte Variabilität der schulischen Prozessmerkmale (bspw. Lehrerkooperation) zeigte sich mittels Varianzzerlegung bei Wurster und Feldhoff (2019) auf der Individualebene (zwischen 67% und 92%, gegenüber 6%-25% auf Schulebene und 0%-21% auf Schulformebene). Dies wird durch Messfehler, die in den subjektiven Einschätzungen der Lehrkräfte inkludiert sind, begründet (Wurster & Feldhoff, 2019). Weiterhin bestätigen sie die These zur Schule als pädagogische Handlungseinheit von Fend auch in den neuen Analysen, da sich Schulen stärker in ihren Merkmalen unterscheiden als Schulformen, also das System. Die Variabilität zwischen den Schulen unterscheidet sich jedoch je nach betrachtetem Merkmal. Bei der Wahrnehmung von Merkmalen des Unterrichts finden Wurster und Feldhoff (2019) die meiste Variation ebenfalls auf Individualebene (zwischen 68% und 75%), jedoch auch relevante Teile auf Klassen- (16%-24%) und kleinere Anteile auf Schulebene (5%-10%). Die Schulform ist laut ihren Ergebnissen keine relevante Einheit für Unterrichtsmerkmale (0%-3%), wohingegen sich in anderen Studien durchaus Unterschiede zwischen den Schulformen zeigen (vgl. bspw. Kunter et al., 2005). Die Schlussfolgerung der Autoren lautet, dass es nicht *die* eine Ebene als pädagogische Handlungseinheit gibt, sondern je nach betrachtetem Merkmal unterschiedliche Ebenen relevant sind und zudem auch über die Interaktion über die Ebenen

hinweg Schul- und Unterrichtsqualität entsteht (Wurster & Feldhoff, 2019). Diese Befunde zeigen jedoch auch, dass die Betrachtung der Einzelschule, sowohl auf der Ebene der gesamten Schule, als auch auf Ebene des Unterrichts, bedeutsam für die Betrachtung schulischer Lernprozesse ist.

1.3 Die Relevanz von Schulfaktoren für Schülerleistungen in PISA 2000

Ähnlich zum Vorgehen von Wurster und Feldhoff (2019) auf Basis der Fendschen Daten, wurden auch die PISA-Daten (Programme For International Student Assessment) herangezogen, um die Bedeutung von Einzelschulen zu analysieren. Zuallererst steht bei PISA, welches von der OECD initiiert wurde, die Untersuchung der Leistungen der Schülerinnen und Schüler im Mittelpunkt. Dabei wird in allen teilnehmenden Ländern alle drei Jahre die Mathematik, Lese- oder Naturwissenschaftskompetenz der 15-Jährigen anhand einer repräsentativen Stichprobe von mindestens 150 Schulen untersucht (OECD, 2002). Dennoch wurden die PISA-Daten auch in weiteren Forschungsfeldern analysiert (s. bspw. auch Bischof, 2014). Dazu gehört beispielsweise die Schuleffektivität (die in Kapitel 1.5 in den Mittelpunkt rückt). In diesem Zusammenhang wurde das Buch „PISA 2000 – School factors related to quality and equity“ (OECD, 2005) veröffentlicht. Da bei PISA nicht nur umfangreiche Kompetenzdaten erhoben wurden, sondern auch Daten zur Ausstattung der Schulen, zum Schulcurriculum, zu schulinternen Evaluationspraktiken sowie zu Aspekten des Unterrichts und des Schulmanagements, der Schulkultur etc., konnten die Daten auch hinsichtlich der Relevanz und Effektivität von Schulen Erkenntnisse liefern (OECD, 2005).

Bezüglich der Leistungsunterschiede stellen die Autoren gegenüber, wieviel Varianz in den Leistungen der Schülerinnen und Schülern zwischen den Schulen und innerhalb der Schulen besteht – gemessen an OECD-Durchschnittsunterschieden in Leistungen zwischen Schülerinnen und Schülern. Dabei werden die Länder eingeteilt in Gruppen, in denen vergleichsweise große, mittlere oder geringe Unterschiede zwischen den mittleren Schulleistungen bestehen. In der Gruppe der Länder mit großen Leistungsunterschieden liegen 50-75% der Leistungsunterschiede zwischen den Schulen. Anzumerken ist hierbei, dass keine Analysen auf Klassenebene berechnet werden können, da das Studiendesign von PISA die

Erfassung aller 15-Jährigen einer Schule beinhaltet und keine Klassen- oder Jahrgangszugehörigkeit erfasst wird (OECD, 2002). Ein Großteil der Varianz der Leistung zwischen Schülerinnen und Schülern liegt also auf Schulebene (OECD, 2005). Zu dieser Gruppe gehört auch Deutschland. Es wird konstatiert, dass die Schule, die Schülerinnen und Schüler besuchen, einen signifikanten Unterschied für ihre Lesekompetenz ausmacht (OECD, 2005, 2010, 2013). Werden die großen Unterschiede in den Leistungen zwischen den Schulen in Bezug zur mittleren Höhe der Leseleistungen insgesamt gesetzt, befindet sich Deutschland erneut in der Gruppe der ungünstigen Bedingungen: Die großen Leistungsunterschiede zwischen den Schulen gehen mit einer unterdurchschnittlichen Leistung in Bezug zum OECD Durchschnitt einher (sowie in den meisten anderen Ländern mit großen Unterschieden zwischen den Schulen). Es zeigt sich demnach das Muster, dass Länder, in denen große Leistungsunterschiede zwischen Schulen bestehen, gleichzeitig insgesamt unterdurchschnittliche Leistungen erzielen (OECD, 2005).

Ein weiteres wichtiges Ergebnis der schulspezifischen Auswertung der PISA-Daten ist, dass Charakteristika der Schülerinnen und Schüler, der Schulkontext (Schultyp, Standort, durchschnittlicher sozioökonomischer Hintergrund) sowie Schulklima, Schulpolitik und -ressourcen insgesamt 75 % der Varianz in der Schulleistung auf Schulebene erklären (OECD, 2005). Im OECD-Durchschnitt wird dabei 50 % der Varianz zwischen Schulen durch Charakteristika der Schülerinnen und Schüler, 18 % durch den Schulkontext und 6 % durch das Schulklima, Schulpolitik und schulische Ressourcen erklärt, 25 % bleiben unerklärt. Im Allgemeinen überwiegt der Effekt der Zusammensetzung der Schülerschaft im Vergleich zu den anderen Schulmerkmalen (Schulklima, Schulpolitik und schulische Ressourcen), in Deutschland ist er jedoch besonders groß (OECD, 2005). Von diesen anderen Schulmerkmalen erklärt das Schulklima (hier zusammengefasst aus verschiedenen Skalen zu relevanten Prozessmerkmalen auf Unterrichts- und Schulebene) den größten Teil der Varianz, im OECD-Durchschnitt sind das 8 % (OECD, 2005). Von den Autoren wird betont, dass der geringere Einfluss der Merkmale Schulklima, Schulpolitik und schulische Ressourcen diese nicht weniger bedeutsam macht, sondern hier potentielle Möglichkeiten der politischen Intervention gesehen werden, um Leistungen von Schülerinnen und Schülern zu verbessern (OECD, 2005).

Die PISA 2000-Erhebung ist eine Querschnittsuntersuchung, sodass keine kausalen Interpretationen von Effekten der Schulvariablen auf die Leistungen der Schülerinnen und Schüler gezogen werden können (OECD, 2005). Zudem lag das Ziel der PISA-Erhebungen in der repräsentativen und validen Erhebung von individuellen Schülerdaten der verschiedenen Länder und nicht in der genaueren Beschreibung der Schulebene, sonst wäre es günstiger gewesen, mehr als 150 Schulen pro Land zu erfassen (OECD, 2005). Dennoch konnte mit der schulspezifischen Auswertung ein wertvoller Beitrag zum Wissensstand über Zusammenhänge von Faktoren und Prozessen auf Einzelschulebene mit Leistungen, also auch zur Schuleffektivität, geleistet werden.

1.4 Schulen als differenzielle Lern- und Entwicklungsmilieus

Im Zuge der PISA-Studie 2000 wurde eine Erweiterungserhebung für Deutschland durchgeführt (PISA-E). Diese wurde kombiniert mit den BIJU-Daten (Bildungsverläufe und psychosoziale Entwicklung im Jugend- und jungen Erwachsenenalter) unter anderem hinsichtlich des Themas „Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus“ ausgewertet (Baumert, Stanat & Watermann, 2006b). Dies ist für die Beschreibung der Relevanz der Einzelschulen in dieser Dissertation bedeutsam, da in dieser Studie einerseits Schulformen als unterschiedliche schulische Lernumwelten herausgearbeitet werden sowie andererseits die Bedeutung von Kompositionseffekten auf Leistungen verdeutlicht wird.

In der Studie von Baumert et al. (2006b) wird auf die Schulstruktur Deutschlands mit ihrer Aufteilung in bestimmte Schulformen Bezug genommen. Auf der Grundlage von pädagogischen Vorstellungen zum Wohle der Schülerinnen und Schüler werden diese nach der Grundschule leistungsbezogen in verschiedene Schulformen aufgeteilt. Somit entsteht eine gewollte Homogenisierung von Schülerinnen und Schülern hinsichtlich ihrer Leistungen innerhalb von Schulformen (Baumert et al., 2006a). Mit dieser Leistungshomogenisierung geht auch eine Homogenisierung hinsichtlich sozioökonomischer Gesichtspunkte einher. Dies liegt daran, dass die Leistungen sehr stark mit sozioökonomischen Aspekten zusammenhängen (zumindest in Deutschland). Laut dieser Analyse von Baumert et al. (2006b) werden 70 % der

Varianz im Fähigkeitsniveau der Schulen über das mittlere soziale Niveau der Schülerschaft erklärt. Eine weitere Erkenntnis dieser Studie ist, dass sich nicht nur die Schulen unterschiedlicher Schulformen in ihren Leistungen und in ihrer sozio-ökonomischen Zusammensetzung deutlich unterscheiden, sondern auch Schulen derselben Schulform. Die Zusammensetzung der Schülerschaft hinsichtlich verschiedener Merkmale kann sich als „kumulative Privilegierung oder Benachteiligung von Schulen" (Baumert et al., 2006b, S. 97) äußern. Diese Unterschiede der schulischen Lernumwelten werden als „differenzielle Lern- und Entwicklungsmilieus" (Baumert et al., 2006b) wie folgt definiert:

> „Wenn wir von differenziellen Lern- und Entwicklungsmilieus sprechen, ist damit gemeint, dass junge Menschen unabhängig von und zusätzlich zu ihren unterschiedlichen persönlichen, intellektuellen, kulturellen, sozialen und ökonomischen Ressourcen je nach besuchter Schulform differenzielle Entwicklungschancen erhalten, die schulmilieubedingt sind und sowohl durch den Verteilungsprozess als auch durch die institutionellen Arbeits- und Lernbedingungen und die schulformspezifischen pädagogisch-didaktischen Traditionen erzeugt werden." (Baumert et al., 2006b, 98f.)

Weiterhin wurden drei Erklärungsmöglichkeiten für fortschreitende unterschiedliche Leistungsentwicklungen zwischen den Schulformen geliefert: (1) individueller Matthäus-Effekt: Durch das unterschiedliche Niveau der Vorkenntnisse der Schülerinnen und Schüler machen diese nun auch unterschiedliche Lernfortschritte. (2) Institutionelle Unterschiede beispielsweise durch unterschiedliche Curricula und Lernkulturen. (3) Institutioneller Matthäus-Effekt: durch Kompositionseffekte der Schülerschaft entstehen für alle Schülerinnen und Schüler „förderliche" oder „hinderliche" Effekte (Baumert et al., 2006b). Die Studie betont dabei nicht nur die daraus entstehenden Folgen für die Leistungsentwicklung, sondern auch für die Persönlichkeitsentwicklung, beispielsweise im Rahmen des „Big-Fish-Little-Pond-Effects" (Baumert et al., 2006b).

Mit der beschriebenen Studie wird zunächst die Erkenntnis, dass Schul*formen* unterschiedliche Lernumwelten bilden können, in den Mittelpunkt gerückt. Dennoch birgt sie auch Hinweise auf die Relevanz von Einzelschulen, da aufgezeigt wird, dass erhebliche

Leistungsunterschiede zwischen Schulen bestehen, die unter anderem durch ihre Lernumwelt erklärt werden können. Die drei möglichen Erklärungen für die bestehenden Leistungsunterschiede wären somit auch im Hinblick auf Einzelschulen anwendbar. Zudem wird durch die Betrachtung der Effekte, die bei Baumert et al. (2006b) durch die Schülerkomposition gezeigt werden, die besondere Bedeutung des Kontexts von Schulen deutlich.

1.5 Schuleffektivitätsforschung

> "School effectiveness refers to the performance of the organizational unit called 'school'. The performance of the school can be expressed as the output of the school, which in turn is measured in terms of the average achievement of the pupils at the end of a period of formal schooling. The question of school effectiveness is interesting because it is well known that schools differ in performance. The next question is how much they differ, or, more precisely, how much schools differ when they are more or less equal in terms of pupils' innate abilities and socio-economic background."
> (Scheerens, 2000)

Eine Forschungsrichtung, die die Einzelschule dezidiert in ihren Mittelpunkt rückt, ist die Schuleffektivitätsforschung, auch international „school effectiveness resesearch" oder „educational effectiveness research" benannt. Sie befasst sich grundsätzlich mit der Wirkung der einzelnen Schule auf die Leistungen von Schülerinnen und Schülern. Darin drückt sich demnach die „Qualität" einer Schule aus, was häufig auch als „Schuleffekt" bezeichnet wird (Kunter & Baumert, 2008). Bei der Betrachtung der Lernergebnisse von Schülerinnen und Schülern werden verschiedene Faktoren der Schule berücksichtigt.

Die Schuleffektivitätsforschung hat seit ihrem Beginn in den 1960er/1970er-Jahren verschiedene Phasen durchlaufen, die laut Bischof (2014) in der internationalen Literatur mit unterschiedlichen Schwerpunkten beschrieben werden. Gemeinsam scheint jedoch, dass als Ursprung der Coleman-Report (Coleman et al., 1966) sowie die Studie von Jencks (1972) benannt werden. Mit diesen Studien wurde der „schools make no difference" – Gedanke

prominent. Demnach habe die Schule keine Bedeutung für Unterschiede in den Leistungen von Schülerinnen und Schülern, wenn andere Merkmale wie der soziale Hintergrund der Lernenden berücksichtigt werden. Diese Studien wurden daraufhin vielfach bezüglich ihrer methodischen Vorgehensweise kritisiert und führten wiederum zu einigen Studien, die zu einem gegensätzlichen Schluss kommen: „school matters" (Brookover, Beady, Flood, Schweitzer & Wisenbaker, 1979; Mortimore, Sammons, Stoll, Lewis & Russell, 1988; Teddlie & Stringfield, 1993). Als Folge wurde nicht nur der Frage nachgegangen, ob die Schule grundsätzlich einen Einfluss auf das Lernen der Schülerinnen und Schüler hat, sondern auch, *welche* spezifischen Merkmale der Schule das Lernen beeinflussen. Dabei wurden mehr Faktoren mit einbezogen als ursprünglich bei Coleman et al. (1966) und Jencks (1972). Eine weitere Entwicklung betrifft den Versuch, den „Effekt der Schule" vom „rohen" Wert einer Schulleistung im Gesamten zu trennen (vgl. bspw. Raudenbush & Willms, 1995), beispielsweise als „value added" (vgl. bspw. Lenkeit, 2012; OECD, 2008; Saunders, 1999). Als ein „Herzstück" (Scheerens, 2004; Thillmann, 2012) der Schuleffektivitätsforschung kann die „effective schools research" (Scheerens, 2004) bezeichnet werden. Dies ist die Suche nach den essentiellen Faktoren auf Schulebene, die maßgeblich den Effekt einer Schule, also die Qualität einer Schule bestimmen. Dies wird häufig im Hinblick auf spätere Leistungen von Schülerinnen und Schülern betrachtet (Scheerens, 1990, 2004). Dafür wurden insbesondere Schulen, die sich als außergewöhnlich „effektiv" zeigten, zur Analyse verwendet. Das besondere dieser Antwort der Schuleffektivitätsforschung auf den vorherigen Pessimismus bezüglich der Rolle der Schule für die spätere Leistungsentwicklung war, dass sie damit versuchte, die *black box* der Prozesse innerhalb der Schule aufzubrechen und einen Einblick in relevante Prozessmerkmale zu bekommen (Scheerens, 2004). Sammons et al. (1995) beschrieben auf Basis zahlreicher einzelner Studien die folgenden elf Hauptmerkmale: Professionelle Führung, gemeinsame Visionen und Ziele, ein lernfreundliches Umfeld, Konzentration auf das Lehren und Lernen, zielgerichtetes Lehren, hohe Erwartungen, positive Verstärkung, Überprüfung des Fortschritts, Schülerrechte und -verantwortung, Kooperation zwischen Schule und Elternhaus, lernende Organisation. In Anlehnung an die von Sammons et al. (1995) beschriebenen relevanten Merkmale effektiver Schulen formulierte Scheerens (2004) fünf Faktoren, die seiner Ansicht

nach die damaligen Ergebnisse der Schuleffektivitätsforschung zusammenfassen: starke pädagogische Führung, Betonung des Erwerbs von Grundfertigkeiten, ein geordnetes und sicheres Umfeld, hohe Erwartungen an die Leistungen der Schülerinnen und Schüler, häufige Überprüfung der Fortschritte der Schülerinnen und Schüler.

Die Forschungstradition der Schuleffektivitätsforschung liefert somit wesentliche Erkenntnisse für die Variabilität schulischer Lernumwelten: Die Schule ist als Analyseeinheit zentral, um Faktoren auszumachen, die Leistungen von Schülerinnen und Schülern beeinflussen können und es gibt bestimmte Merkmale von Schulen, die die Effektivität von Schulen bedingen.

1.6 Schulentwicklung durch Outputsteuerung

Die Bedeutung der Einzelschule in der Praxis, zeigt sich unter anderem im Hinblick auf bildungspolitische Vorgänge, die Schulentwicklungsprozesse initiieren und eine Verbesserung der Leistungen der Schülerinnen und Schüler erzielen sollen. Dabei spielen Unterrichtsentwicklung, Organisationsentwicklung und Personalentwicklung eine Rolle (Rolff, 2010, eine Zusammenfassung der Tradition der Schulentwicklung findet sich bspw. bei Bischof, 2014; Thillmann, 2012). Im Rahmen dieser Dissertation spielt die Schulentwicklung vor allem durch das neue Steuerungsparadigma eine Rolle. Während das Bildungssystem bis zu den Reformvorhaben im Nachgang von PISA 2000 primär durch Input-Vorgaben gesteuert wurde („Inputsteuerung"), u.a. durch Curricula, Anforderungen an die Lehrerbildung oder finanzielle Ressourcenzuweisungen (KMK, 2004, 2015) kam es unter anderem durch die Veröffentlichung der „schockierenden" PISA- und TIMSS-Ergebnisse 2000 zu einer verstärkten „Outputorientierung" (Altrichter & Maag Merki, 2010; KMK, 2004, 2006, 2015). Diese gab den Einzelschulen neue Relevanz, dadurch dass die Autonomie der Einzelschulen und deren Entscheidungsfähigkeit hinsichtlich bestimmter Vorgänge und Umsetzungsmöglichkeiten der Schule betont wurde (Altrichter & Maag Merki, 2010). Eine der Folgen dieser „Neuen Steuerung" (vgl. bspw. Altrichter & Maag Merki, 2010) war die Etablierung der Schulinspektion sowie der Vergleichsarbeiten in Deutschland (KMK, 2006, 2015).

Durch die Schulinspektion werden die Schulen hinsichtlich normativer Qualitätskriterien bewertet, die sich aus den landesspezifischen Qualitätsrahmen ergeben (vgl. bspw. Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013). Die Schulinspektion tut dies, da sie die Schule „als Ganzes" als essentiell für die Qualitätsentwicklung und als „lernende Organisation" ansieht (Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013). Diese Form der externen Evaluation soll Wissen generieren, das unter anderem schulinterne Prozesse zur Schulentwicklung auslösen kann (Landwehr, 2011). Die Schulinspektion hat somit die Funktion inne, Unterrichts- und Schulqualität sowie Leistungen von Schülerinnen und Schülern zu verbessern (Ehren & Visscher, 2008). Bei den Vergleichsarbeiten handelt es sich um bundesweite Kompetenztests der Leistungen von Schülerinnen und Schülern im Vergleich zu den Bildungsstandards, die bundesweit durch die Kultusministerkonferenz der Länder (KMK) formuliert wurden (KMK, 2004, 2015). Diese Bildungsstandards beschreiben Fähigkeiten, die die Schülerinnen und Schüler im Durchschnitt in der vierten bzw. zehnten Jahrgangsstufe erreichen sollen (KMK, 2004). Aus den Ergebnissen der Schulinspektion und der Vergleichsarbeiten sollen dann wiederum Folgen für die Unterrichts- und Schulentwicklung der einzelnen Schulen abgeleitet werden (KMK, 2015; Kotthoff & Böttcher, 2009). Diese Instrumente dienen also der Qualitätsentwicklung in Schulen. Es kann daher festgehalten werden, dass diese Instrumente der Schulpraxis, die an Schulentwicklung orientiert sind, die besondere Relevanz der Einzelschule betonen.

Im Zusammenhang mit der Leitidee der Schulinspektion, dass die Schulebene eine relevante Rolle einnimmt, haben sich Wurster und Gärtner (2013) ebenfalls mit der Frage der Varianzverteilung von Unterrichtsmerkmalen auf verschiedenen Ebenen auseinandergesetzt. Im Vorgehen ähnlich wie die in Kapitel 1.2 und 1.3 vorgestellten Studien, haben sie dies explizit für Daten der Schulinspektion aus Brandenburg für die Ebenen Schülerin und Schüler, Klasse, Schule und Schulform überprüft. Die Ergebnisse zur Beurteilung des Unterrichts aus Sicht der Visitatorinnen und Visitatoren zeigen einen geringen Unterschied der Inspektionsbewertungen zwischen den Schulformen (durchschnittlich 2,5 % Varianzanteil), einen „relativ gering[en]" (Wurster & Gärtner, 2013, S. 225) Unterschied zwischen den Schulen (durchschnittlicher Varianzanteil 9,1 %) und eine größere Varianzverteilung innerhalb der Schule: 88,4 %. Die

Bewertung des Unterrichts durch die Schülerinnen und Schüler variiert stärker zwischen den Schulformen (8,8 % Varianzanteil), weniger zwischen den Schulen (durchschnittlich 5,7 % Varianzanteil) und auch nur wenig mehr zwischen den Klassen (durchschnittlich 7,5 %). Der größte Anteil der Varianz liegt hierbei auf Ebene der einzelnen Schülerinnen und Schüler (78 % Varianzanteil). Es kann somit in beiden Fällen die größte Varianz auf der Ebene des jeweiligen Beurteilenden gefunden werden (Wurster & Gärtner, 2013). Die Autoren schlussfolgern daraus, dass der Varianz der Unterrichtsbewertung innerhalb der Schulen größere Bedeutung durch die Schulinspektion beigemessen werden sollte, beispielsweise durch genauere Angaben zur Heterogenität der Unterrichtsqualität im Schulinspektionsbericht (Wurster & Gärtner, 2013). Die Untersuchung der Varianzverteilung auf die verschiedenen Ebenen ist laut der Autoren insbesondere dadurch interessant, dass die Schulinspektion ausschließlich die Zielebene Schule beleuchtet. Dennoch wird durch die Befunde auch erkennbar, dass in den Einschätzungen der Unterrichtsqualität aus Sicht von Visitatorinnen und Visitatoren, der Einzelschule eine größere Bedeutung zukommt, als der Schulform. Zudem zeigt sich durch die Varianzaufklärung der Urteile von Schülerinnen und Schülern zur Unterrichtsqualität, dass die größten Unterschiede zwischen den einzelnen beurteilenden Personen liegen und die Klassenebene einen kaum höheren Anteil der Varianz in der Unterrichtsqualität erklärt, als die Schulebene.

2 Zusammenspiel unterschiedlicher Faktoren schulischer Qualität

Kapitel 1 hat die Relevanz der Einzelschule für Forschung und schulische Praxis dargelegt. Die Schuleffektivitätsforschung hat dabei aufgezeigt, dass die Einzelschule immer in ein Gefüge von Voraussetzungen und Faktoren eingebettet ist und die Qualität einer Einzelschule (Schuleffektivität) durch dieses Gesamtgefüge bestimmt wird. Ziel dieses Kapitels ist es, relevante Faktoren zu beleuchten, die Schuleffektivität und damit auch schulische Lernumwelten ausmachen, sowie das Zusammenspiel dieser Faktoren zu betrachten. Die Schuleffektivitätsforschung ist relevant für die Analyse der Variabilität schulischer Lernumwelten, da sie Modelle über Gelingensfaktoren von Schulen bietet und dadurch (1) Bereiche identifiziert, in welchen sich Schulen unterscheiden können sowie (2) mögliche Ursachen für Unterschiede zwischen Schulen liefern kann.

In diesem Kapitel wird das kombinierte Kontext-Input-Prozess-Output-Modell von Scheerens (1990) vorgestellt, dann werden die einzelnen Bereiche des Modells beleuchtet, um diese um andere theoretische Vorstellungen aus den Bereichen der Schulqualitäts- sowie Lehr-Lernforschung zu erweitern (Kapitel 2.1-2.5). Zudem werden empirische Befunde zum Zusammenspiel von Faktoren schulischer Qualität zusammengefasst (Kapitel 2.6).

Auf Grundlage der „school effectiveness“ und unter Hinzunahme der Forschung zu „instructional effectiveness” bildete Scheerens (1990) ein integriertes Modell der Schuleffektivität (Scheerens, 1990, 2004; Wang, Haertel & Walberg, 1993), welches eines der grundlegendsten im internationalen Bereich wurde (dargestellt in Abb. 2).

Grundsätzlich konzeptualisiert Scheerens (1990) das Handeln von Schulen im Gesamtzusammenhang von Inputmerkmalen mit den tatsächlichen Prozessmerkmalen innerhalb der Schule und den daraus resultierenden Outputs. Zusätzlich zeigt das Modell auf, dass die Realisierung der Prozesse (zum Beispiel die Unterrichtsqualität) auch in Abhängigkeit zum Kontext der Schulen steht. Die Prozesse sind die essentiellen Merkmale der Schuleffektivität: Sie sind strukturell unterteilt in eine Schulebene und eine Unterrichtsebene (Klassenebene).

Dieses Modell von Scheerens (1990) wurde durch andere Autoren in verschiedenen Varianten abgewandelt, erweitert oder es wurde auf verschiedene Aspekte fokussiert (vgl. bspw. Ditton, 2000; Helmke, 2009).

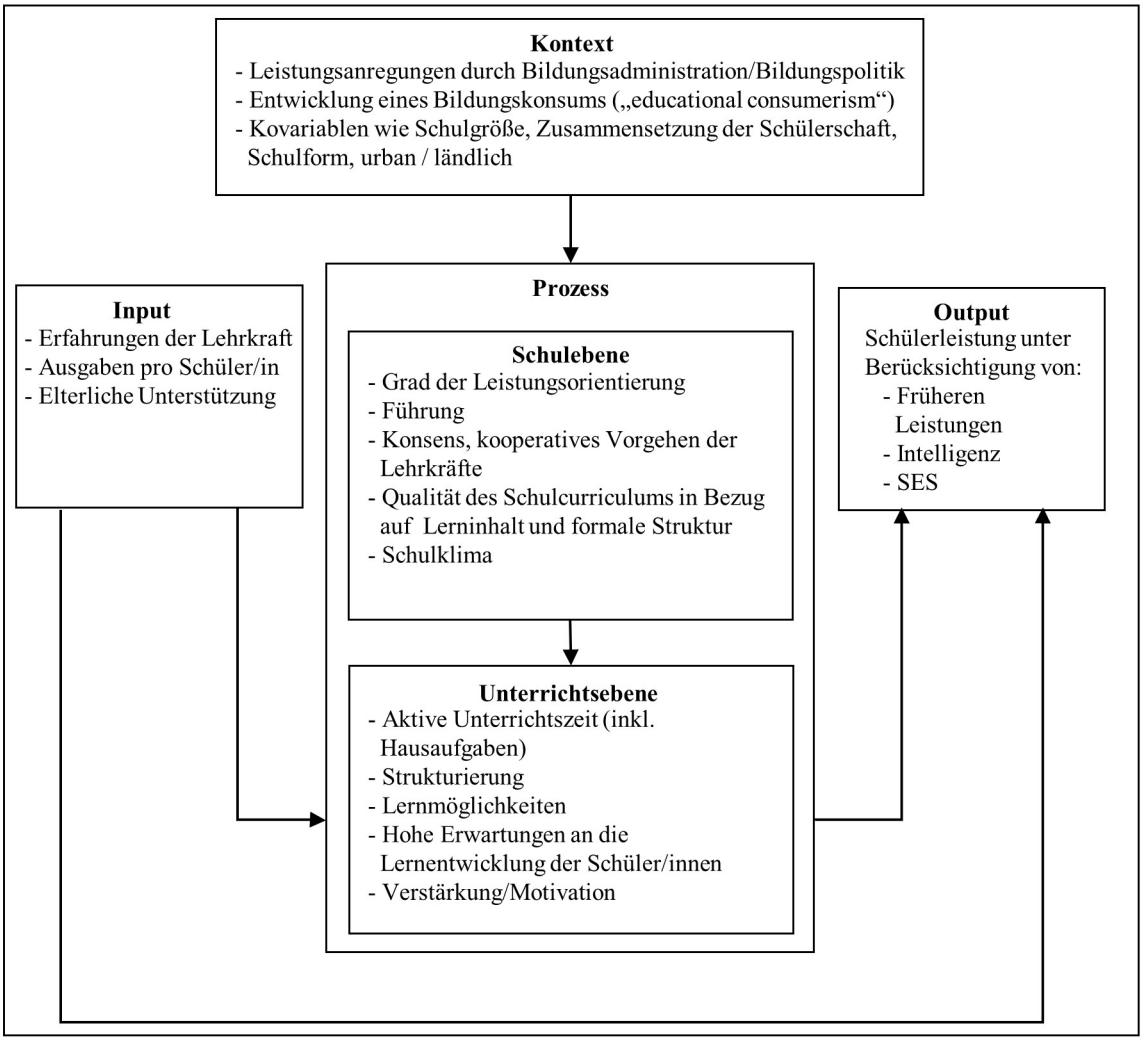


Abb. 2 eigene Übersetzung des Modells der Schuleffektivität nach Scheerens, 1990

2.1 Kontext

Kontextmerkmale sind nach Scheerens (1990) beispielsweise Leistungsanreize durch die Bildungsadministration/Bildungspolitik sowie Merkmale der Schule, wie die Schulgröße, die Zusammensetzung der Schülerschaft, die Schulart sowie die Umgebung, in die die Schule eingebettet ist (bspw. ländlich vs. urban).

2.2 Input

Inputmerkmale sind nach Scheerens (1990) beispielsweise die Erfahrungen und Kompetenzen der Lehrkräfte, die finanziellen Mittel pro Schülerin oder Schüler, die investiert werden sowie die Unterstützung der Schule durch Eltern (Abb. 2). In anderen Modellen wird im Sinne eines Inputs der Schule und des Unterrichts stärker auf professionelle Kompetenzen von Lehrkräften oder auf inhaltliche Vorgaben zu Lerninhalten und Lernzielen fokussiert (Helmke, 2009; KMK, 2004).

2.3 Prozess – Unterrichtsebene

Die Unterrichtsebene beinhaltet durch die unmittelbare Nähe zum Lernen der Schülerinnen und Schüler die „proximalen" Faktoren, welche die Leistungen der Schülerinnen und Schüler beeinflussen können (Ditton, 2000; Kunter & Baumert, 2008; Scheerens, 1990). Auf dieser Ebene lokalisiert Scheerens (1990) die Merkmale aktive Unterrichtszeit/Lernzeit (Slavin, 1994), Strukturierung, Möglichkeiten des Erlernens von Inhalten, Erwartungen an die Lernentwicklung der Schülerinnen und Schüler, Bestärkung und Motivation der Schülerinnen und Schüler (Scheerens, 1990). Die auf der Unterrichtsebene handelnden Akteure sind vorrangig die Lehrkräfte. Scheerens (1990) bezieht sich bei der Ausgestaltung der Unterrichtsebene überwiegend auf das QAIT-Modell von Slavin (1994), welches vier relevante Merkmale beschreibt: „quality of instruction", „appropriate levels of instruction", „incentive" und „time". Hier berührt die Schuleffektivitätsforschung die Lehr-Lernforschung, „die die Qualität des Unterrichts an dessen Wirkungen misst" (Ophardt & Thiel, 2008). Zahlreiche Autoren haben sich neben Slavin (1994) damit auseinandergesetzt, die Merkmale des Unterrichts zu identifizieren und zu beschreiben, die essentiell für das Lernen von Schülerinnen und Schülern sind, aber auch für andere Outcomes, wie beispielsweise sozio-emotionale Merkmale. Auch wenn die Begrifflichkeiten sich teils unterscheiden, besteht breiter Konsens über drei Merkmale, die Unterrichtsqualität ausmachen (Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Helmke, 2009; Klieme, Schümer & Knoll, 2001; Kunter & Voss, 2011; Ophardt & Thiel, 2008; Pianta & Hamre, 2009; Seidel & Shavelson, 2007):

(1) Klassenführung (Klassenmanagement),

(2) kognitive Aktivierung,

(3) konstruktive Unterstützung.

(1) Klassenführung beinhaltet die Prävention von Störungen während des Unterrichts sowie die effektive Nutzung von Unterrichtszeit. Unterricht ist eine komplexe soziale Situation, in der die Lehrkräfte gleichzeitig viele Geschehnisse koordinieren, einordnen und regulieren müssen (Kunter & Voss, 2011). (2) Kognitive Aktivierung bedeutet, dass die Lehrkraft aktivierende Aufgabenstellungen für die Klasse vorbereitet, die kognitiv herausfordern. Diese sollen an bestehendes Wissen anknüpfen, aber die Schülerinnen und Schüler auch herausfordern Probleme zu lösen und über das Bekannte hinaus zu denken. Kognitive Aktivierung kann beispielsweise auch durch Diskussionen und Austausch verschiedener Lösungswege erzielt werden (Kunter & Voss, 2011). (3) Durch konstruktive Unterstützung kann die Lehrkraft den Schülerinnen und Schülern in ihrem Lernprozess helfen. Dies kann beispielsweise durch Struktur im Unterricht und Aufbereiten der Inhalte (scaffolding) erfolgen, aber auch durch individuelles Feedback und durch das Erzeugen einer angenehmen Lernatmosphäre und Wohlfühlen im Unterricht (Kunter & Voss, 2011). Durch die COACTIV-Studie (Kunter, Baumert, Blum & Neubrand, 2011) sind diese drei Merkmale auch als „Basisdimensionen von Unterrichtsqualität“ bekannt.

2.4 Prozess – Schulebene

Die Schulebene bildet die übergeordnete Ebene der Schule, die auch den Unterricht beeinflusst. Die Schulleitung ist der Akteur, welcher überwiegend auf dieser Ebene handelt. Auf der Schulebene befinden sich die folgenden Prozessmerkmale: Grad des leistungsorientierten Vorgehens, Führung, Konsens/kooperatives Vorgehen der Lehrkräfte, Qualität des Schulcurriculums in Bezug auf den Inhalt und die formale Struktur, Schulklima, das Nutzen von Evaluationen (Scheerens, 1990).

Betrachtet man die relevanten Merkmale aus der Schuleffektivitätsforschung (Sammons et al., 1995; Scheerens, 2004) erscheinen diese Merkmale nicht systematisiert zusammengefasst oder „nach einem unbekannten Strukturierungsprinzip erstellt zu sein“ (Ditton, 2000). Stringfield (1994) legt in Anlehnung an das QAIT-Modell (Slavin, 1994) eine Systematisierung

wichtiger Merkmale auf Schulebene im MACRO-Modell vor. Die Schule wird darin als „Unterstützungssystem“ (Ditton, 2000) des Unterrichts betrachtet, die hemmend oder förderlich für den Unterricht sein kann. MACRO steht für die folgenden fünf Merkmale: “meaningful, universally understood goals”; “attention to daily academic functioning”; “coordination among programs and between school and parents over time”; “recruitment of prospective teachers, development of staff, and, when necessary, the removal of longitudinally unsuccessful teachers from the school”; “organization of the school to support universal student learning” (Stringfield, 1994).

Auf Stringfields MACRO-Merkmalen basierend, entwickelte Ditton (2000) wiederum eine Einteilung (Tab. 1) für den deutschsprachigen Raum, die in Modellen der Schulqualität und Qualitätsentwicklung eine große Rolle spielt:

(1) Schulkultur,

(2) Schulmanagement,

(3) Kooperation und Koordination und

(4) Personalpolitik und Personalentwicklung.

Ditton beschreibt diese als „übergreifende und miteinander zusammenhängende Bereiche“ (Ditton, 2000, S. 84, Tab. 1 zur Beschreibung der bedeutsamen Faktoren auf Schulebene).

Tabelle 1: Bedeutsame Faktoren auf der Schulebene nach Ditton (2000)

Bedeutsame Faktoren auf der Schulebene	
Schulkultur	**Schulmanagement**
Gemeinsam akzeptierte, handlungsrelevante und eindeutige Ziele; Einigkeit hinsichtlich der primär zu erfüllenden Aufgabe; Organisatorische und pädagogische Leitung; Geregelte Zuständigkeiten und Verantwortlichkeiten;	Gemeinsam geteiltes Aufgabenverständnis; Eine gemeinsame *Vision* Geklärte Entscheidungsbefugnisse und -verfahren; Geregelte Aufgabenverteilungen
Kooperation und Koordination	**Personalpolitik und Personalentwicklung**
Koordinierter Schul- und Unterrichtsbetrieb; Kooperation mit Partnern außerhalb der Schule (Eltern, Administration, Berater…) Einführung neuer Lehrer; Erfahrungsaustausch / Wissens-Sharing	Kooperation innerhalb der Schule (Schulleitung-Lehrer-Schüler); Rekrutierung, Sozialisation und Weiterbildung der Lehrer; Regelungen der Fort- und Weiterbildung

2.5 Output

Als Output beschreibt Scheerens (1990) die Leistungen der Schülerinnen und Schüler und verdeutlicht auch hier noch einmal die Bedeutung des Zusammenspiels der Kontext-Input-Prozess-Output-Faktoren durch die Berücksichtigung oder Adjustierung der Leistungswerte durch Lernausgangslagen der Schülerinnen und Schüler wie die frühere Leistung, die Intelligenz und den sozioökonomischen Status (SES).

2.6 Empirische Befunde zum Zusammenspiel von Kontext, Input, Prozess, Output

Anschließend an die Erläuterung der einzelnen Bereiche und Faktoren des Modells soll daher an dieser Stelle beschrieben werden, welche Verbindungen, also Zusammenhänge, zwischen verschiedenen Merkmalen bereits häufig in der Literatur empirisch überprüft wurden. Dies stellt keine lückenlose Darstellung aller Studien zu Zusammenhängen von Faktoren des Modells dar, sondern lediglich einen Überblick über die wichtigsten Ergebnisse.

Verbindung Input-Prozess-Output

Bezüglich der Verbindung von Input und Prozess ist ein häufig beforschtes Thema das, der Fähigkeiten/des Wissens oder auch der Einstellungen der Lehrkraft und der Zusammenhang dieser Aspekte mit Unterrichtsqualität. Exemplarisch für die Input-Prozess-Verbindung soll an dieser Stelle daher hierauf eingegangen werden. Weiterhin wird betrachtet, wie diese Inputbedingung vermittelt über den Unterricht auf die Leistungen der Schülerinnen und Schüler wirken (Input-Prozess-Output). Die Ausprägung der Kompetenz der Lehrkraft ist dabei positiv mit Indikatoren der Unterrichtsqualität assoziiert. In diesem Zusammenhang ist insbesondere auf die COACTIV-Studie zu verweisen, in der auf Basis einer repräsentativen Stichprobe von Mathematiklehrkräften die professionelle Kompetenz umfangreich untersucht wurde (Kunter, Klusmann & Baumert, 2009; Kunter & Voss, 2011; Kunter et al., 2011). Beispielsweise konnte darin gezeigt werden, dass das fachdidaktische Wissen und die Freude am Unterrichten (Motivation der Lehrkraft) zentral für eine hohe Unterrichtsqualität sind (Kunter et al., 2009). Zudem zeigte sich bei Lehrkräften, die transmissive lerntheoretische Überzeugungen haben (bspw. der Überzeugung sind, dass nur der eine richtige Lösungsweg geübt werden muss), dass dies mit geringerer Unterrichtsqualität einherging. Hierin wurde unter anderem auch die, über die Unterrichtsqualität vermittelte, Wirkung auf die Leistungen der Schülerinnen und Schüler untersucht. Dabei zeigte sich, dass das fachdidaktische Wissen, die motivationale Orientierung und die Überzeugungen der Lehrkraft auch für die Leistungen relevant sind (Kunter et al., 2009).

Verbindung Kontext-Prozess

Der Zusammenhang zwischen Kontextmerkmalen und den innerschulischen Prozessen ist keine typische Verbindung, die viel beforscht wird. Van Ewijk und Sleegers (2010) beschreiben den Umstand, dass der Einfluss von Kontext häufig direkt am Output gemessen wird, jedoch oft keine Details zu den Mechanismen *innerhalb* der Schule betrachtet werden. Sie bezeichnen diese Wirkungswege des Kontexts, durch die Prozesse der Schule hindurch auf den Output daher als *„black box"* (van Ewijk & Sleegers, 2010). Die prominenteste Forschungsfrage zur Untersuchung von Kontextmerkmalen, ist die zu Kompositionseffekten. Diese beschäftigen

sich jedoch explizit *nicht* mit den Prozessen in der Schule, sondern sehr ausführlich mit dem Einfluss der Zusammensetzung der Schülerschaft auf den Output. Die Vermittlung durch die Prozesse der Schule wird somit übersprungen. Dumont, Neumann, Maaz und Trautwein (2013) vermuten, dass die Prozesse selten Berücksichtigung finden, da diese sehr komplex und schwierig zu erfassen sind. Sie beschreiben jedoch, dass sich häufig theoretische Annahmen zu diesen Prozessen innerhalb der Schulen finden lassen, die sie als „Wirkmechanismen" (Dumont et al., 2013) bezeichnen und in drei Gruppen unterteilen: (1) Lehrkräfte passen ihr Verhalten bzw. ihren Unterricht in einer bestimmten Form an die vorhandene Schülerschaft an. Hierbei ist anzumerken, dass es für deutsche weiterführende Schulen aufgrund des gegliederten Schulsystems (tracking) sehr schwierig ist, diese Anpassungen an die Schülerschaft von intendierten Vorgängen im Unterricht zu unterscheiden, die durch unterschiedliche Bildungsgänge entstehen (Dumont et al., 2013). (2) Der zweite beschriebene Wirkmechanismus wird in der internationalen Literatur auch als „peer effect" (Wilkinson, 2002) bezeichnet und sieht die Gründe für unterschiedliche Outputs je nach Komposition in der Interaktion der Schülerinnen und Schüler miteinander. Hier wird angenommen, dass sich bestimmtes Verhalten und bestimmte sozio-emotionale Bedingungen „ansteckend" auswirken (Dumont et al., 2013). (3) Der dritte Mechanismus sind die Ressourcen, die die Schule durch eine bestimmte Schülerschaft, durch ihren bestimmten Kontext zur Verfügung hat (Dumont et al., 2013). Weiterhin beschreiben Dumont et al. (2013), dass die drei Wirkmechanismen wiederum auch in Wechselwirkung zueinander stehen können. Diese theoretischen Vorstellungen stehen dem wenig ausgeprägten empirischen Forschungsstand zur Verbindung von Komposition und Prozessebene der Schulen gegenüber. Ein detaillierter Überblick zu empirischen Studien, die Zusammenhänge der Zusammensetzung der Schülerschaft mit Merkmalen der Prozessebene der Schulen überprüft haben, erfolgt in Studie 2 dieser Dissertation (siehe Appendix A, Studie 2).

Bei Scheerens (1990) wird auch die Einflussnahme bzw. Intervention durch die Bildungspolitik/-verwaltung, die zu einer Leistungsverbesserung der Schülerinnen und Schüler anregen soll, als Kontext eingeordnet. Im deutschen Bildungssystem sollen daher an dieser Stelle auch Ergebnisse zur Verbindung von Steuerungsmechanismen oder Maßnahmen der

Schulentwicklung und Prozessmerkmalen der Schulen thematisiert werden. Beispielsweise gibt es einige Studien, die sich mit den Folgen der Schulinspektion für die Unterrichts- und Schulqualität von Schulen auseinandersetzen (vgl. bspw. Brimblecombe, Shaw & Ormston, 1996; Ehren & Visscher, 2008; Gärtner, Hüsemann & Pant, 2009; Reezigt & Creemers, 2005). Ein detaillierter Überblick hierzu erfolgt in Studie 3 dieser Dissertation.

Verbindung Kontext-Output

Die relevanteste Forschungsrichtung zum Zusammenhang zwischen Kontextmerkmalen und Outputs, ist, wie bereits angesprochen, die der Kompositionsforschung. Die Effekte der Schülerzusammensetzung für die Leistungen der Schülerinnen und Schüler werden in zahlreichen Publikationen untersucht (für einen Überblick vgl. bspw. Dumont et al., 2013). Obwohl der Effekt der Zusammensetzung der Lerngruppe, zusätzlich zum individuellen Einfluss von familiären Hintergrundvariablen (De Fraine, Van Damme, Van Landeghem, Opdenakker & Onghena, 2003; Harker & Tymms, 2004; Opdenakker & Van Damme, 2001) zum Teil auch als statistisches Artefakt diskutiert wird, besteht weitestgehend Einigkeit über seine Existenz und besondere Bedeutung für schulisches Lernen. Es können laut Dumont et al. (2013) drei besonders relevante Merkmale der Schülerzusammensetzung unterschieden werden: Merkmale der ethnischen Zusammensetzung, der sozio-ökonomischen Zusammensetzung und der leistungsbezogenen Zusammensetzung. Eine Zusammenfassung (vgl. bspw. Harker & Tymms, 2004) der Ergebnisse für die drei Merkmalsgruppen und ihre Effekte auf Leistungen von Schülerinnen und Schülern findet sich in Studie 2 dieser Dissertation.

Des Weiteren gibt es Studien, die sich mit anderen Kontextmerkmalen, wie „Interventionen“ durch die Bildungspolitik und -verwaltung zur Leistungsverbesserung, beschäftigen. Dies bezieht sich in Deutschland häufig auf Instrumente der Outputsteuerung (siehe Kapitel 1.6, KMK, 2015), beispielsweise die Einführung der Schulinspektion, zentraler Abschlussprüfungen und der Vergleichsarbeiten (vgl. bspw. Thiel, Hannover & Pant, 2014). Zur Entwicklung von Leistungen durch zentrale Abschlussprüfungen und die Nutzung von

Vergleichsarbeiten zeigen sich sowohl positive (Richter, Böhme, Becker, Pant & Stanat, 2014; Wößmann, 2009) als auch negative (bzw. keine) Effekte (Maag Merki, 2010).

Ein Überblick zu Studien, die sich mit den Folgen der Schulinspektion für Leistungen von Schülerinnen und Schülern beschäftigen, findet sich in Studie 3 dieser Dissertation.

Verbindung Prozess-Output

Hier geht es um die Untersuchung von relevanten schulischen Prozessmerkmalen, die sich auf Outputs, also auf Leistungen auswirken. Dies stellt eine vielfach untersuchte Verbindung dar, sowohl in der Lehr-Lernforschung als auch in der Schuleffektivitätsforschung: es trifft den Kern der Forschung zu effektiven Schulen. Hierbei erwiesen sich in zahlreichen Studien die benannten Unterrichts- und Schulmerkmale aus den Kapiteln 2.3 und 2.4 als relevante Kriterien für die Leistungen der Schülerinnen und Schüler. Auf Unterrichtsebene kann auch an dieser Stelle auf das COACTIV-Projekt verwiesen werden, dass die Relevanz der drei Basisdimensionen von Unterricht (effiziente Klassenführung, kognitive Aktivierung, konstruktive Unterstützung) für die Leistungen von Schülerinnen und Schülern zeigen konnte (Kunter & Voss, 2011; Kunter et al., 2011). Doch auch zahlreiche andere Studien zeigten die Bedeutung der in Kapitel 2.3 beschriebenen Unterrichtsmerkmale für schülerische Leistungen: Je höher die Ausprägung des Unterrichtsmerkmals, desto höher die Leistung (Haertel, Walberg & Weinstein, 1983; Matsumura, Garnier, Pascal & Valdes, 2002; Pianta & Hamre, 2009; Scheerens & Bosker, 1997; Seidel & Shavelson, 2007).

Auf Schulebene gibt es wiederum zahlreiche Studien, die die zentrale Bedeutung der beschriebenen Merkmale für Schulqualität (Schulkultur, Schulmanagement, Kooperation und Koordination, Personalentwicklung) aus Kapitel 2.4 im Hinblick auf Leistungen zeigten (Ditton & Müller, 2011; Fend, 1988; Scheerens & Bosker, 1997; Slavin, 1994; Stringfield, 1994; Teddlie & Stringfield, 1993).

Kapitel 2 hat somit das Gesamtgefüge der unterschiedlichen Qualitätsfaktoren von Einzelschulen aufgezeigt sowie das Zusammenspiel der unterschiedlichen Faktoren beleuchtet. Dies zeigt, dass Variabilität schulischer Lernumwelten an verschiedenen Stellen entstehen kann

und dass einige Zusammenhänge bereits häufiger untersucht wurden (bspw. der Zusammenhang Kontext-Output) als andere (bspw. Kontext-Prozess).

3 Forschungsfazit und Ziel der Studien

In der Einleitung wurden zwei übergreifende Leitthemen dieser Dissertation zur Variabilität schulischer Lernumwelten formuliert: (1) Schule als relevante Analyseeinheit und (2) die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität sowie für Schulleistung. In diesem Kapitel sollen die eingangs formulierten Leitthemen aus der Theorie bzw. bisheriger Forschung abgeleitet werden. Dazu sollen die Forschungsdesiderate bisheriger Arbeiten zu den zwei Leitthemen dieser Arbeit zusammengefasst (Kapitel 3.1) und daraus die übergreifenden Forschungsfragen dieser Arbeit beschrieben (Kapitel 3.2) werden. Anschließend wird ihre Umsetzung und Operationalisierung in den Studien dieser Dissertation mit Hilfe eines Überblicksmodells erläutert (Kapitel 3.3).

3.1 Fazit und Forschungsdesiderate

(1) Schule als relevante Analyseeinheit

Die Wahrnehmung der Relevanz der Einzelschule bzw. der Schulebene hat sich, wie in Kapitel 1 beschrieben, immer wieder verändert. Aus unterschiedlichen Blickrichtungen, die teilweise wenig miteinander zu tun haben, wurde die Relevanz der Einzelschule in bisheriger Forschung beleuchtet. So wurde in den 1970er-Jahren von Fend die Einzelschule als pädagogische Handlungseinheit (Fend, 1986, Kapitel 1.1) beschrieben, im Gegensatz zur vorherigen Fokussierung auf die Systemebene. Fend verwendete jedoch damals noch nicht alle statistischen Möglichkeiten zur Datenanalyse, wie sie heute zur Verfügung stehen. Dies wurde von Wurster und Feldhoff (2019) nachgeholt (Kapitel 1.2). Weiterhin brachte die PISA-Studie der Bedeutung der schulischen Lernumwelt neuen Aufschwung (Kapitel 1.3, 1.4). Obwohl die PISA-Ergebnisse in erster Linie für die reliable Erhebung der Leistungen von Schülerinnen und Schülern gedacht war, ergaben sich aus den Analysen auch Erkenntnisse für die Relevanz der Einzelschule bzw. den Einfluss der Schulebene: Durch aktuelle Methoden wie die Mehrebenenanalyse und die Beleuchtung der Varianzaufklärung unterschiedlicher Ebenen, konnten beispielsweise Unterschiede in der Leistung zwischen Schulen in den verschiedenen Ländern angegeben werden (OECD, 2005, 2010, 2013, Kapitel 1.3). Jedoch wurde hier noch keine systematische Übersicht zu Unterschieden von Unterrichtsmerkmalen zwischen und

innerhalb von Schulen gegeben. Eine solche Übersicht könnte einen weiteren Beitrag zum Verständnis der Unterschiede zwischen Schulen leisten. Ergebnisse der OECD (2005) deuten darauf hin, dass das Zusammenspiel von Kontextmerkmalen und Einzelschulen relevant ist. Dazu, inwieweit die Einzelschulen jedoch als Gesamtsysteme auf Kontextmerkmale reagieren, besteht noch Forschungsbedarf. Weiterhin wurden durch die Ausführungen und Analysen zu „Schulen als differenzielle Lern- und Entwicklungsmilieus“ (Baumert et al., 2006b, Kapitel 1.4) wichtige Erkenntnisse zur Variabilität von Schulen bekannt. Diese rückten insbesondere die grundlegenden Unterschiede der Lernumwelten an unterschiedlichen Schulformen in den Blick. Die Autoren sprechen zwar auch die große Variabilität innerhalb der Schulformen an, beispielsweise wenn sie angeben, dass Schulen mit mittleren Bildungsabschlüssen sehr breit in ihren Leistungen und in ihrem sozialen Hintergrund variieren, jedoch wird hier nicht vertiefend auf die Einzelschule eingegangen. Auch durch den Forschungsstrang der Schuleffektivitätsforschung (Kapitel 1.5) erfolgte die Diskussion zur Bedeutung der Einzelschule und ihres Beitrags zu Leistungen der Schülerinnen und Schüler, insbesondere indem Merkmale „effektiver Schulen“ festgehalten wurden. Zudem hat sich der Bereich der Schulentwicklung mit Einzelschulen und ihrer Steuerung durch die Bildungspolitik bzw. den Möglichkeiten ihrer Veränderungen und Weiterentwicklungen beschäftigt. Die Einführung der Outputsteuerung (Kapitel 1.6) ging zwar teils mit der Evaluation bestimmter Wirkungen (wie bspw. Folgen der Schulinspektion für Unterrichts-/Schulqualität und Leistungen) einher, andere bildungspolitische Vorgänge, wie die Einordnung von Schulen als erheblich entwicklungsbedürftig durch die Schulinspektion, wurden wiederum noch nicht evaluiert.

Es kann insgesamt zum Leitthema „Schule als relevante Analyseeinheit“ festgehalten werden, dass sich aus verschiedenen Blickrichtungen bereits mit der Schule als Analyseeinheit befasst wurde, jedoch noch etliche Desiderate bestehen, bei deren Aufklärung diese Dissertation einen Beitrag leisten soll: So ist die Bewertung einer gewissen Varianzaufklärung weiterhin unklar – ist beispielsweise 10 % Varianzaufklärung eines Unterrichtsmerkmals auf Schulebene als „hoch“ zu werten? Dafür bedarf es einer systematischen Übersicht, welche die Unterschiede der Unterrichtsmerkmale zwischen Schulen quantifizieren kann. Zudem bedarf

es weiterer Forschung zum Zusammenspiel schulischer Kontextmerkmale mit Prozess- und Outputmerkmalen der Einzelschulen als Gesamtsystem.

(2) Die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität sowie für Schulleistung

Ein zweites Leitthema dieser Dissertation stellt die Bedeutung des Schulkontexts (im Sinne von Komposition und Intervention), insbesondere für Prozessmerkmale wie die Unterrichts- und Schulqualität, aber auch für Leistungen von Schülerinnen und Schülern dar. In Kapitel 2 wurde das Modell der Schuleffektivität von Scheerens (1990) dargestellt und aufgezeigt, dass es verschiedene Faktoren gibt, die die Qualität von Schulen bedingen. Diese Struktur wurde dann um Aspekte der Lehr-Lernforschung bzw. der Schulqualitätsforschung erweitert (Kapitel 2.3 und 2.4) um die innerschulischen Prozesse zu strukturieren. Eine systematische empirische Überprüfung des Kontext-Input-Prozess-Output-Modells hat bislang nicht stattgefunden. Es wurden hingegen einzelne Verbindungen zwischen Qualitätsfaktoren von Schule häufiger untersucht (Kapitel 2.6), wie beispielsweise der Zusammenhang zwischen Komposition (Kontext) und Leistungen (Output), während andere Verbindungen (bspw. Kontext und Prozess) weniger häufig untersucht wurden. Die Beschreibung der innerschulischen Prozesse als *black box* (bspw. van Ewijk & Sleegers, 2010) bringt die Vernachlässigung dieses Themas in bisherigen Studien auf den Punkt (Kapitel 2). Zudem wurden in bisherigen Studien eher nur Zusammenhänge einzelner Merkmale schulischer Qualität analysiert, anstatt umfassender und systematischer vorzugehen und auf mehrere Merkmale zu fokussieren (Kapitel 2.6). Diese Dissertation soll einen Beitrag dazu leisten, *mehrere* Verbindungen zwischen Kontextmerkmalen, Prozessmerkmalen und Outputs zu analysieren, die bislang nur unzureichend betrachtet wurden.

Ein weiterer Aspekt des Leitthemas betrifft die Besonderheit des Schulkontexts (im Sinne von Komposition und Intervention) bei der Analyse von Fragen, die die Einzelschule und ihre Unterrichts- und Schulqualität sowie ihre Leistungen betreffen. Die Relevanz von Kompositionseffekten als eine Variante von Kontexteffekten wurde insbesondere durch PISA und in Deutschland durch die Analyse der PISA-E-Daten prominent (Baumert et al., 2006b; Deutsches PISA-Konsortium, 2003; OECD, 2005, siehe Kapitel 1.3 und 1.4). Die

Berücksichtigung der Komposition bei der Betrachtung von individuellen Leistungen von Schülerinnen und Schülern stellt ein vielfach untersuchtes Feld dar (zum Überblick siehe Dumont et al., 2013; van Ewijk & Sleegers, 2010). Auch für die Schuleffektivitätsforschung (siehe Kapitel 1.5 und 2.1) beinhaltet der Fokus auf die Effektivität von Schulen (durch Leistungen von Schülerinnen und Schülern) stets Merkmale des Kontexts der Einzelschule. Hinsichtlich der Unterschiede zwischen Einzelschulen wird ebenso bei Baumert et al. (2006b) beschrieben, wie unterschiedlich schulische Lernumwelten allein schon durch bestimmte Kompositionsmerkmale der Schülerschaft sind. Bislang wurde der Beitrag bestimmter Kompositionseffekte für die Einzelschule, also Faktoren, die für unterschiedliche schulische Lernumwelten sorgen, jedoch unzureichend berücksichtigt, da hauptsächlich Zusammenhänge von Kontextmerkmalen mit Outputs, also Leistungen betrachtet wurden. Die Verbindung von Kontext (im Sinne von Komposition) mit Prozessmerkmalen von Schulen fand bisher kaum statt. Dazu kann diese Dissertation einen Beitrag leisten.

Bezüglich des Kontexteinflusses durch Maßnahmen der Bildungspolitik zur Schulentwicklung (Intervention), bzw. zur Leistungssteigerung gibt es bisher zwar Forschung zu Effekten der Schulinspektion bzw. der Vergleichsarbeiten (siehe Kapitel 2.6), andere Maßnahmen wie beispielsweise die Diagnose einer Schule als entwicklungsbedürftig im Rahmen der Schulinspektion (international unter „special measure“ bekannt“), wurden nur wenig untersucht. Die Schulinspektion stellt dabei ein Instrument der Steuerung dar, indem die Prozessebene von Schulen beleuchtet wird. Durch die Diagnose „erheblicher Entwicklungsbedarf“ wird für Schulen ein Kontext geschaffen, der zahlreiche Konsequenzen für die Schule haben kann. Dazu gehören einerseits ein erhöhter Druck auf die Lehrkräfte und die Schulleitung (Reezigt & Creemers, 2005) und die Stigmatisierung der Schulen, andererseits erhalten diese Schulen zusätzliche Unterstützung bei ihrer Arbeit (bspw. durch Beratung). Auch zu dieser Verbindung von Kontext, im Sinne von bildungspolitischer Intervention, mit Prozess- und Outputdaten kann mit dieser Dissertation ein Beitrag geleistet werden.

3.2 Forschungsfragen

Vor dem Hintergrund der Leitthemen, lassen sich zwei übergreifende Fragestellungen dieser Dissertation ableiten, die bisherige Forschungslücken füllen können:

(1) Wie unterschiedlich sind Schulen in ihrer Unterrichtsqualität?

Diese Frage berührt insbesondere das Thema, ob die Schule eine relevante Analyseeinheit darstellt (Leitthema 1), speziell für Unterrichtsmerkmale (Klassenführung, kognitive Aktivierung, konstruktive Unterstützung). Unterscheidet sich Unterrichtsqualität überhaupt zwischen Schulen? Beinhaltet Unterrichtsqualität Merkmale, die auf Schulebene liegen? Oder gibt es kein „Unterrichtsprofil" einer Schule?

(2) Unterscheiden sich schulische Prozessmerkmale und Outputs je nach Kontext?
   a) Kontext im Hinblick auf Schülerzusammensetzung (Komposition)
   b) Kontext im Hinblick auf bildungspolitische Intervention

Die zweite Forschungsfrage stellt den Zusammenhang von verschiedenen Faktoren schulischer Qualität in den Mittelpunkt: Dabei wird die Bedeutung des Schulkontexts für Prozessmerkmale und Outputs von Schulen besonders hervorgehoben (Leitthema 2). Macht es für die Unterrichts- und Schulqualität einen Unterschied, welchen Kontexteinflüssen eine Schule unterliegt? Spielen Kontexteinflüsse eine Rolle für die Leistungen von Schülerinnen und Schülern? Bei der Untersuchung der zweiten Forschungsfrage steht ebenfalls die gesamte Schule als Analyseeinheit im Fokus (Leitthema 1). Der Kontext soll hierbei a) die Schülerzusammensetzung (Komposition) und b) bildungspolitische Intervention sein. Dabei wird der Einfluss von Komposition und bildungspolitischer Intervention auf die Variabilität schulischer Lernumwelten untersucht.

3.3 Konkretisierung und Überblick über die Studien

In Kapitel 1 dieser Arbeit wurden empirische Studien, die sich mit der Relevanz der Einzelschule beschäftigt haben, kurz erläutert. In Kapitel 2 wurde das Kontext-Input-Prozess-Output-Modell und seine einzelnen Bereiche und Faktoren theoretisch vorgestellt sowie übliche Forschungsthemen, die sich mit Zusammenhängen zwischen einzelnen Merkmalen des Modells beschäftigt haben, beschrieben. In Kapitel 3.1 wurden die Forschungsdesiderate aus Kapitel 1 und 2 in Bezug auf die zwei Leitthemen dieser Arbeit zusammengefasst und daraus die übergreifenden zwei Forschungsfragen dieser Arbeit abgeleitet (Kapitel 3.2). Nun soll an dieser Stelle erläutert werden, wie die Forschungsfragen für die vorliegende Dissertation operationalisiert wurden. Weiterhin wird das Modell vorgestellt, in welches die drei Studien dieser Dissertation eingeordnet werden. Es stellt eine Synthese aus theoretischen Ideen der Bereiche Schuleffektivitätsforschung, Lehr-Lern-Forschung und Schulqualitätsforschung dar (Synthese aus Ditton, 2000; Kunter & Voss, 2011; Scheerens, 1990).

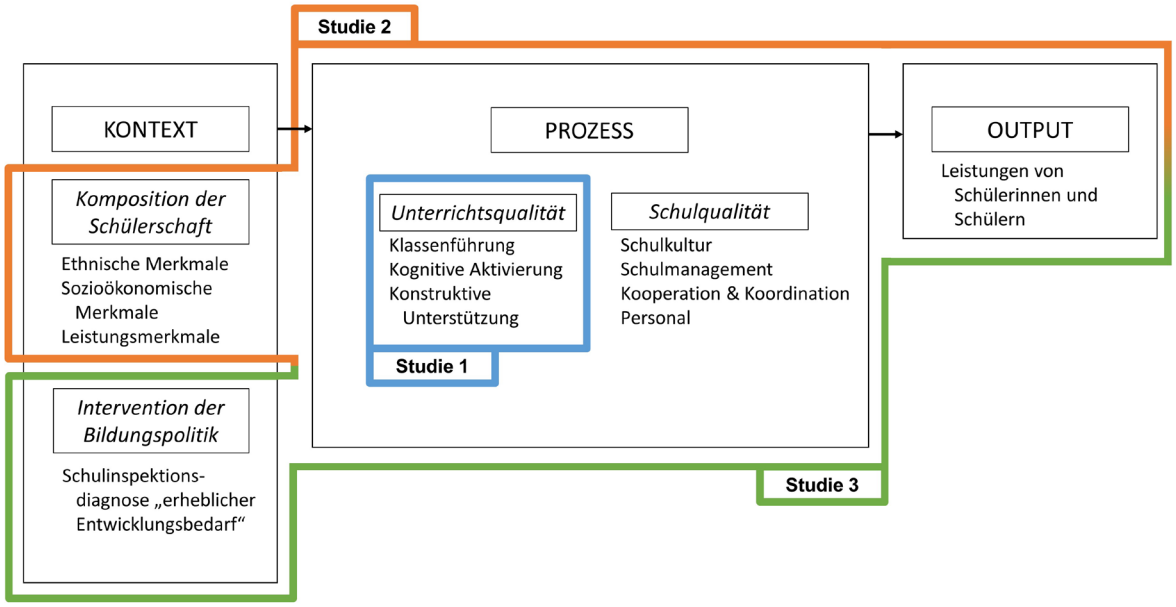


Abb. 1 Übersicht der Dissertationsstudien angelehnt an Ditton, 2000; Kunter & Voss, 2011; Scheerens, 1990

Das Kontext-Input-Prozess-Output-Modell (Scheerens, 1990) weist verschiedene Faktoren schulischer Qualität auf, welche zwischen Schulen variieren können. Daher wird dieses Modell in dieser Dissertation zur Analyse der Variabilität schulischer Lernumwelten herangezogen und auf die Besonderheiten der Dissertation angepasst. Der Kontext wird in dieser Dissertation

mittels zweier verschiedener Bereiche berücksichtigt: a) der Komposition der Schülerschaft und b) durch bildungspolitische Intervention, konkret durch die Diagnose „Schule mit erheblichem Entwicklungsbedarf“ im Rahmen der Schulinspektion. Damit rückt die Berücksichtigung des Schulkontexts bei der Betrachtung von schulischen Prozessmerkmalen und Outputs verstärkt in den Fokus und stellt eine Besonderheit dieser Dissertation dar (Leitthema 2). Der Input wird in der vorliegenden Arbeit nicht weiter berücksichtigt. Bezüglich der Prozessmerkmale wird in dieser Dissertation in Anlehnung an die bisherige Literatur zur Schuleffektivität (vgl. bspw. Ditton, 2000; Fend, 1988; Scheerens, 1990; Scheerens & Bosker, 1997) zwischen der (1) Unterrichtsqualität und der (2) Schulqualität unterschieden (siehe Abb. 1). Diese umfassen somit Prozessmerkmale, die innerhalb der Schule liegen und insgesamt relevant für die Effektivität der Schule sind. (1) Für die Unterrichtsebene werden hierbei die Merkmale verwendet, die als „Basisdimensionen für Unterrichtsqualität“ breit akzeptiert sind (Klieme et al., 2001; Kunter & Voss, 2011): Effiziente Klassenführung, kognitive Aktivierung, konstruktive Unterstützung (zu alternativen, aber äquivalenten Begrifflichkeiten siehe Kapitel 2.3). (2) Für die Schulebene wurden die Merkmale verwendet, die Ditton (2000) als relevante Faktoren der Schulqualität identifiziert hat: Schulkultur, Schulmanagement, Kooperation und Koordination, Personal. Anders als im ursprünglichen Kontext-Input-Prozess-Output-Modell werden die Unterrichtsqualität und Schulqualität in den Studien dieser Dissertation nicht hierarchisch strukturiert (die übergeordnete Schulebene beeinflusst die Unterrichtsebene), sondern Unterrichtsqualität und Schulqualität stehen gleichwertig nebeneinander. Damit sind beide für die Schule insgesamt relevant und Unterricht wird ebenso auf Schulebene aggregiert. Bezüglich des Output-Bausteins werden ausschließlich Leistungen der Schülerinnen und Schüler verwendet, ebenfalls aggregiert auf Schulebene, also als mittlere Schulleistung.

Mit dieser spezifizierten Version des Kontext-Input-Prozess-Output-Modells werden die zwei übergreifenden Fragestellungen in dieser Dissertation analysiert. In Abbildung 1 ist dargestellt, wie sich die drei Studien dieser Dissertation in das übergeordnete Modell einfügen. Im Folgenden sollen die drei Studien zusammenfassend vorgestellt werden.

Studie 1 fokussiert dabei auf den Teil der Unterrichtsqualität und geht der Frage nach, wie unterschiedlich schulische Lernumwelten überhaupt sind und wie sich diese mit statistischen Kennwerten quantifizieren lassen (Forschungsfrage 1). Versteht man Einzelschulen als Handlungs*einheit*, sollten auch die Urteile der Schülerinnen und Schüler einer Schule in gewissem Maße übereinstimmen. Die drei statistischen Kennwerte, die im Fokus der Studie standen, dienen (1) der Messung der Übereinstimmung von Schülerurteilen zu Unterrichtsqualität innerhalb einer Schule, (2) der Messung der Variabilität, also der Unterschiede zwischen den Schulen in Merkmalen der Unterrichtsqualität und (3) der Messung der Reliabilität von aggregierten Schülerurteilen. Bei der Aggregation von Schülerurteilen zu Merkmalen von Unterrichtsqualität einer Schule (bspw. relevant im Rahmen von Schulinspektionen), wird schlichtweg davon ausgegangen, dass Schülerurteile innerhalb einer Schule übereinstimmen, Unterschiede zwischen Schulen bestehen und die Messung dieser Werte reliabel ist. Dies sind Grundvoraussetzungen, die bis dahin nicht systematisch überprüft wurden. In Studie 1 wird in Anlehnung an Metaanalysen ein systematischer Überblick über die Höhe der drei Kennwerte gegeben. Um eine systematische Übersicht der drei Kennwerte zu erhalten, wurde auf die umfassende Datenbasis der PISA-Erhebungen der Jahre 2000, 2003, 2006, 2012 zurückgegriffen und alle Unterrichtsqualitätsmerkmale der drei Basisdimensionen von Unterrichtsqualität verwendet. Damit wurden für alle vorhandenen PISA-Länder (81) die Kennwerte zur Übereinstimmung innerhalb von Schulen, zur Variabilität zwischen Schulen und zur Reliabilität der Variabilitätswerte berechnet. Dafür wurden prominente Indizes verwendet (vgl. bspw. Bliese, 2000; James, Demaree & Wolf, 1984; Lüdtke, Trautwein, Kunter & Baumert, 2006): (1) das Übereinstimmungsmaß $r_{WG(J),}$ (2) die Varianzaufklärung mit Hilfe von Intraklassenkorrelationen (ICC(1)) und (3) die ICC(2) zur Reliabilitätsanalyse. Bezüglich der Variabilität schulischer Lernumwelten kann Studie 1 somit aufzeigen, wie groß Unterschiede sind, die zwischen Schulen (und innerhalb von Schulen) hinsichtlich ihrer Unterrichtsqualität bestehen.

Studie 2 beschäftigt sich damit, wie sich Prozess und Output von Schulen in Abhängigkeit von der Komposition unterscheiden (Forschungsfrage 2a). Sie widmet sich also der Frage nach der Variabilität zwischen Schulen im Zusammenspiel mit ihren

Kontextmerkmalen, hier: durch Merkmale der Zusammensetzung der Schülerschaft. Dabei werden (1) die Zusammenhänge der Merkmale von Schülerinnen und Schülern mit Merkmalen der Unterrichts- und Schulqualität als Prozessmerkmale untersucht sowie (2) inwieweit sich das Zusammenspiel von Kontextmerkmalen und Merkmalen der Unterrichts- und Schulqualität wiederum in Leistungen als Output der Schule niederschlägt. Insbesondere die Betrachtung des Zusammenhangs zwischen der Schülerzusammensetzung als Kontext und der Unterrichts- und Schulqualität als Prozessmerkmale wird in der bisherigen Forschung kaum untersucht (eine umfassende Zusammenstellung der Literatur dazu findet sich in Appendix A der Studie 2). Als Datengrundlage der zweiten Studie diente eine Kombination verschiedener umfassender Datensätze der Berliner Grundschulen: Schulinspektionsdaten der Jahre 2011-2017, Daten der amtlichen Statistik (zur Erfassung des sozioökonomischen Status, des Migrationshintergrunds von Schülerinnen und Schülern auf Schulebene) sowie Daten zu den Ergebnissen der Vergleichsarbeiten (VERA) der dritten Jahrgangsstufe in Mathematik und Deutsch. Diese besondere Datenbasis wurde herangezogen und mittels Korrelationsanalysen und Pfadmodellen ausgewertet, um die Forschungsfragen zu beantworten. Studie 2 kann daher in dieser Dissertation zur Beantwortung der Frage nach Unterschieden zwischen der Unterrichts- und Schulqualität (sowie der Leistungen) in Abhängigkeit der Komposition als Kontext der Schule beitragen.

Studie 3 widmet sich der Frage, wie sich Prozessmerkmale und Outputs von Schulen in Abhängigkeit einer bildungspolitischen Intervention unterscheiden (Forschungsfrage 2b). Sie nähert sich dem Thema Variabilität schulischer Lernumwelten, indem sie eine bestimmte Intervention durch die Bildungspolitik in den Fokus rückt. In Anlehnung an Scheerens (1990) kann diese Einflussnahme der Bildungspolitik und -verwaltung zur Schulentwicklung und Leistungserhöhung („achievement stimulants from higher administrative levels“, Scheerens, 1990) als Kontext von schulischen Prozessen sowie deren Outputs verstanden werden. Die Intervention besteht darin, dass im Rahmen der Schulinspektion die Diagnose „Schule mit erheblichem Entwicklungsbedarf“ vergeben werden kann. Dieses Vorgehen, welches in Deutschland in verschiedenen Bundesländern, aber auch international üblich ist, wurde bisher kaum evaluiert. Es gibt zwar einige Studien zu Folgen der Schulinspektion für die Unterrichts-

und Schulentwicklung sowie für die Leistungsentwicklung von Schulen, jedoch gibt es zur Besonderheit der Diagnose eines „erheblichen Entwicklungsbedarfs“ durch die Schulinspektion bislang kaum Literatur. Die dritte Studie dieser Dissertation untersucht daher mögliche Folgen der Diagnose „erheblicher Entwicklungsbedarf“ hinsichtlich der Entwicklung der Unterrichts- und Schulqualität, der Entwicklung der Schulleistung, aber auch hinsichtlich der Entwicklung der Schülerschaft. Die Datengrundlage dieser Studie ist erneut eine Kombination mehrerer umfassender Datensätze von Berliner Grundschulen der Jahre 2011 bis 2017: Schulinspektionsdaten, Daten der Vergleichsarbeiten (VERA) in den Fächern Mathematik und Deutsch, Daten der amtlichen Statistik zur Lernmittelzuzahlungsbefreiung und zum Migrationshintergrund der Schülerinnen und Schüler der Jahre 2011-2017. Die Ergebnisse aus Studie 3 können daher in dieser Dissertation zur Beantwortung der Frage nach dem Einfluss von Kontext durch bildungspolitische Intervention auf die Variabilität schulischer Prozessmerkmale und Outputs beitragen. Diese drei vorgestellten Studien (siehe Abb. 1) sollen im folgenden Kapitel (4) präsentiert werden.

4 Studien

Die drei empirischen Dissertationsstudien (siehe Abb. 1) sollen in diesem Kapitel präsentiert werden:

Studie 1:

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene. *Zeitschrift für Erziehungswissenschaft, 21*(5), 929–950. https://doi.org/10.1007/s11618-018-0813-3

Studie 2:

Wenger, M., Gärtner, H. & Brunner, M. (2020). To what extent are characteristics of a school's student body, instructional quality, school quality and school achievement interrelated? *School Effectiveness and School Improvement*. Online Vorveröffentlichung. https://doi.org/10.1080/09243453.2020.1754243

Studie 3:

Wenger, M., Gärtner, H. & Brunner, M. (eingereicht). Wie entwickeln sich Schulen nach der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf"? *Zeitschrift für Erziehungswissenschaft.*

4.1 Studie 1: Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene. *Zeitschrift für Erziehungswissenschaft, 21*(5), 929–950.

Dieser Teil der Dissertation entspricht dem Manuskript, welches zur Publikation der *Zeitschrift für Erziehungswissenschaft* angenommen wurde.

Die finale Publikation ist als Zeitschriftenbeitrag veröffentlicht und online verfügbar unter:

https://doi.org/10.1007/s11618-018-0813-3

Das zugehörige Online-Supplement im Excel-Format ist nicht in der Dissertation enthalten, kann aber ebenfalls unter dem angegebenen Link abgerufen werden.

Zusammenfassung

Für die Analyse der Unterrichtsqualität von Schulen durch Schülerurteile sollten drei Voraussetzungen erfüllt sein: (1) eine angemessene Übereinstimmung der Schülerurteile innerhalb der Schulen, (2) systematische Variabilität der Schülerurteile zwischen Schulen, (3) ein ausreichendes Maß an Reliabilität der aggregierten Urteile. Diese Studie untersucht mit internationalen PISA-Daten (Zyklen 2000-2012; 81 Länder, über 55 300 Schulen, über 1.3 Millionen 15-Jährige), inwiefern dies für Indikatoren der Qualitätsdimensionen des Unterrichts (Klassenführung, Kognitive Aktivierung, Konstruktive Unterstützung) zutrifft. Dafür bestimmten wir das Übereinstimmungsmaß $r_{WG(J)}$ sowie die Intraklassenkorrelationen ICC(1) und ICC(2). Es zeigte sich, dass (1) die Mehrzahl der Unterrichtsmerkmale eine moderate oder starke Übereinstimmung in Schulen aufwies, (2) sich Unterrichtsmerkmale aus Sicht der Schülerschaft systematisch zwischen Schulen unterschieden, jedoch (3) die Reliabilität der aggregierten Schülerurteile in vielen Ländern nicht ausreichte. Die Ergebnisse diskutieren wir vor dem Hintergrund von Konventionen zur Beurteilung der Übereinstimmung, Variabilität und Reliabilität auf Schulebene.

Schlagwörter: Unterrichtsqualität, Schülerurteile, PISA

Abstract

Three requirements should be fulfilled using student ratings to assess instructional quality of schools: (1) an appropriate level of inter-rater agreement within schools, (2) systematic variance of student ratings between schools, (3) an adequate reliability level of aggregated student ratings. Using international PISA-data (2000 to 2012; 81 countries, over 55 300 schools, over 1.3 million 15-year olds) this study investigated how these requirements were met for indicators of instructional quality (classroom management, cognitive activation, individual learning support). We computed the inter-rater agreement index $r_{WG(J)}$, the intraclass correlations ICC(1) and ICC(2). Our results showed that (1) student ratings demonstrated a moderate or strong level of agreement for most indicators of instructional quality and (2) instructional quality assessed by students varied systematically between schools. Yet, (3) reliability of aggregated student ratings was not sufficient in many countries. We discuss these results regarding conventions to evaluate agreement, variability, and reliability of student ratings at school level.

Keywords: instructional quality, PISA, student ratings

1. Einleitung

Die Analyse der Unterrichtsqualität ist zentral für die empirische Schul- und Unterrichtsforschung, die Schuleffektivitätsforschung und die pädagogische Praxis, z. B. im Rahmen der externen Evaluation von Einzelschulen durch die Schulinspektion. Es besteht breiter Konsens, dass die effiziente Klassenführung, kognitive Aktivierung und konstruktive Unterstützung zentrale Basisdimensionen der Unterrichtsqualität darstellen (Helmke 2009; Lipowsky 2009; Kunter et al. 2011). Zur Erfassung dieser Unterrichtsmerkmale werden häufig Schülerurteile[2] durch Fragebogenverfahren erhoben (Gruehn 2000; Church et al. 2001; Klieme und Rakoczy 2003). Die Schülerurteile werden dann auf Schulebene aggregiert, um die Unterrichtsqualität einer Schule als Ganzes abzubilden. Die aggregierten Schülerwahrnehmungen geben so Auskunft über die geteilte Wahrnehmung der Unterrichtsqualität einer Schule. Um Schulen hinsichtlich ihrer Unterrichtsqualität differenzieren zu können, wird grundsätzlich davon ausgegangen, dass die aggregierten Schülerurteile innerhalb der Schulen (1) übereinstimmen, (2) systematisch zwischen Schulen variieren und (3) hinreichend reliabel sind. Bislang gibt es aber kaum Studien, die im internationalen Kontext diese Kriterien zur Verwendung von Schülerurteilen der Unterrichtsqualität systematisch untersuchten. Die vorliegende Arbeit analysiert deshalb auf Grundlage der Daten aus 81 Ländern, die an PISA (Programme for International Student Assessment) in den Jahren 2000 bis 2012 teilnahmen, die Übereinstimmung, die Variabilität und die Reliabilität von zentralen Merkmalen der Unterrichtsqualität von Schulen aus der Sicht von Schülerinnen und Schülern.

2. Theoretischer Hintergrund

2.1 Statistische Kennwerte zur Analyse der Unterrichtsqualität von Schulen aus Sicht von Schülerinnen und Schülern

Unterrichtsmerkmale werden in der Unterrichtsforschung und in Verfahren der externen Evaluation, wie z. B. der Schulinspektion, häufig durch Urteile von Schülerinnen und Schülern

[2] Bei den Begriffen wie „Schülerurteil", „Schülerwahrnehmung" etc. werden sowohl Schülerinnen als auch Schüler mitgedacht.

erhoben. Eine Vielzahl von Arbeiten beschreiben die Vorteile von Schülerbefragungen (bspw. Gruehn 2000; Clausen 2002; Gärtner 2010): Schülereinschätzungen sind ökonomisch zu erheben, zeigen eine hohe prädiktive Validität für Entwicklungsverläufe und basieren auf einem langen Erfahrungszeitraum, der eine breite Basis relevanter Verhaltensstichproben repräsentiert. Um die Unterrichtsqualität von Klassen oder Schulen zu messen, werden die individuellen Schülereinschätzungen auf Klassen- bzw. Schulebene aggregiert. Da der vorliegende Beitrag auf Unterrichtsmerkmale auf Schulebene fokussiert, wird nachfolgend nur diese Analyseebene dargestellt.

Bei der Aggregation von Schülerwahrnehmungen zur Unterrichtsqualität auf Schulebene wird von sogenannten „Kompositions-Modellen" (Chan 1998) ausgegangen. Diese Modelle beschreiben die funktionalen Beziehungen von Konstrukten auf verschiedenen Ebenen (z. B. Schülerebene und Schulebene), die sich auf dieselben Inhalte beziehen, sich jedoch zwischen den Ebenen unterscheiden können. Zudem geben diese Modelle die notwendigen (Rechen-)Operationen an, wie aus Konstruktindikatoren der niedrigeren Ebene (z. B. Schüler) Konstruktindikatoren der höheren Ebene (z. B. Schulen) gebildet werden (Chan 1998). In nahezu allen empirischen Studien zur Analyse der Unterrichtsqualität und auch in der angewandten schulischen Evaluationsforschung werden implizit oder explizit Modelle des Konsens („Direct Consensus-Modell" oder „Referent-shift Consensus-Modell"; siehe Chan 1998) verwendet. Hierbei wird der jeweilige Schulmittelwert auf Grundlage der individuellen Urteile der Schülerinnen und Schüler gebildet. In der organisationspsychologischen Forschung sowie der empirischen Unterrichtsforschung besteht Einigkeit darüber, dass der Schulmittelwert bei angemessener Übereinstimmung zwischen den Schülerinnen und Schülern ein von allen bzw. der Mehrzahl geteiltes Urteil der Unterrichtsqualität repräsentiert (Lüdtke et al. 2009; Stapleton et al. 2016). Dies ist auch eine wichtige Bedingung für die Konstruktvalidität der aggregierten Maße (Chan 1998; Bliese 2000; Cohen et al. 2001; LeBreton und Senter 2008; Fischer 2009).

Um zu bestimmen, inwiefern die Wahrnehmung der Schülerinnen und Schüler übereinstimmt, kann auf Indizes, die in der Organisationspsychologie entwickelt wurden, zurückgegriffen werden. Einer der wichtigsten Indizes ist der r_{WG}. Bei diesem Index wird davon

ausgegangen, dass sich bei einer rein zufälligen Abgabe von Urteilen der Schülerinnen und Schüler eine Gleichverteilung (eine sogenannte „Nullverteilung“) auf die verschiedenen Antwortmöglichkeiten ergeben sollte. Eine Gleichverteilung erscheint bei Ratingskalen eher unwahrscheinlich, sie dient jedoch als Referenz dafür, wie groß die Varianz wäre, wenn sich die Antworten mit gleicher Wahrscheinlichkeit auf alle Antwortstufen verteilen würden. Die tatsächlich gefundene Varianz der Schülerurteile wird dann an dieser Gleichverteilung relativiert. Wenn die beobachtete Varianz kleiner als die Varianz der Nullverteilung ausfällt, zeigt dies eine überzufällige, systematische Übereinstimmung an (Finn 1970). Dabei gilt: Umso kleiner die an der Nullverteilung relativierte Varianz der Schülerurteile ausfällt, desto größer ist der r_{WG}. Wenn die Urteile der Schülerinnen und Schüler perfekt übereinstimmen, ist die Varianz der Schülerurteile null. In diesem Fall ist der $r_{WG} = 1$. Je mehr sich die Verteilung der Schülerurteile der Nullverteilung annähert, desto mehr nähert sich der r_{WG} null. Bei großer Diskrepanz der Schülerurteile kann der r_{WG} theoretisch sogar negative Werte annehmen, wird jedoch nach Empfehlung von James et al. (1984) in diesem Fall bei null abgeschnitten. Der r_{WG} bezieht sich auf die Übereinstimmung bei einem Einzelitem. Zur Bestimmung der Übereinstimmung von Schülerurteilen auf der Basis mehrerer Items J (z. B. von Items zur Messung der Unterrichtsqualität, die zu einem Skalenwert aufsummiert werden) haben James et al. (1984) den Index $r_{WG(J)}$ vorgeschlagen:

$$r_{WG(J)} = \frac{J[1-(\bar{s}^2/\sigma_{EU}^2)]}{J[1-(\bar{s}^2/\sigma_{EU}^2)]+(\bar{s}^2/\sigma_{EU}^2)} \qquad (1)$$

Hierbei ist $\bar{s}^2$ die durchschnittliche Varianz der J Items einer Skala und σ_{EU}^2 die Varianz der Nullverteilung. Vergleiche von r_{WG}-Werten zwischen Skalen sind nur unter bestimmten Bedingungen wie bspw. einheitlichen Antwortformaten möglich (LeBreton und Senter 2008). Ab wann von einer „akzeptablen“ Übereinstimmung gesprochen werden kann, ist umstritten. Als kritische Grenze wird in der Literatur meist ein Wert von $r_{WG} = .70$ genannt. Diese Grenze wird jedoch selbst kritisch betrachtet (LeBreton und Senter 2008), da sie den Anwendungskontext der aggregierten Maße nicht berücksichtigt (z. B. Forschungskontext vs.

Evaluation der Schulinspektion mit weitreichenden Folgen für die Einzelschule). LeBreton und Senter (2008, Tab. 3) schlagen daher folgende differenzierten Beurteilungskriterien vor: keine Übereinstimmung ($.00 \leq r_{WG} \leq .30$), schwache Übereinstimmung ($.31 \leq r_{WG} \leq .50$), moderate Übereinstimmung ($.51 \leq r_{WG} \leq .70$), starke Übereinstimmung ($.71 \leq r_{WG} \leq .90$), sehr starke Übereinstimmung ($.91 \leq r_{WG} \leq 1.00$). Der Index r_{WG} gibt an, wie gut Urteile der Schülerinnen und Schüler derselben Schule übereinstimmen. Dieser Index betrifft also die Einzelschule und nicht die Gesamtstichprobe. Weiterhin informiert der r_{WG} Index nicht darüber, ob sich Schulen aus Sicht ihrer Schülerinnen und Schüler systematisch hinsichtlich ihrer Unterrichtsqualität unterscheiden, bzw. wie reliabel die aggregierten Schulmittelwerte zur Unterrichtsqualität sind. Um diese Fragen zu beantworten, können Indizes verwendet werden, die zur Bestimmung der Reliabilität von aggregierten Urteilen (wie z. B. Schülerurteilen) entwickelt wurden: die Intraklassenkorrelationen ICC(1) und ICC(2) (Bliese 2000). Diese Indizes werden jeweils für die Gesamtstichprobe und nicht (wie der r_{WG}) für Einzelschulen berechnet.

Die ICC(1) gibt an, welcher Anteil der Varianz (in der Gesamtstichprobe) eines Unterrichtsmerkmals auf die Zugehörigkeit zu einer Schule zurückzuführen ist (Lüdtke et al. 2006). Die ICC(1) kann mit folgender Formel berechnet werden:

$$\mathrm{ICC}(1) = \tau^2 / (\tau^2 + \sigma^2) \qquad (2)$$

Hierbei steht τ^2 für die Varianz zwischen, σ^2 für die Varianz innerhalb der Schulen. Die Werte von ICC(1) können dabei zwischen 0 und 1 liegen: ein Wert von ICC(1) = 0 bedeutet, dass sich die Unterrichtsqualität von Schulen aus Sicht ihrer Schülerinnen und Schüler nicht systematisch unterscheidet. Ein Wert von ICC(1) = .10 bedeutet beispielsweise, dass 10 % der Varianz in den Unterrichtsurteilen der Schülerinnen und Schüler auf systematische Unterschiede zwischen den Schulen zurückzuführen ist. Ein Wert von ICC(1) = 1 bedeutet, dass die beobachtbare Varianz vollständig auf Unterschiede zwischen Schulen zurückgeht. Da die ICC(1) hierbei wie eine Effektgröße interpretiert wird, empfehlen LeBreton und Senter (2008) auch die Anwendung traditioneller Konventionen für die Interpretationen von Effektgrößen: ICC(1) = .01 wird als „kleiner“, ICC(1) = .10 als „mittlerer“ und ICC(1) = .25

und größer wird als „großer“ Effekt angesehen. Empfehlenswert ist die Anwendung dieser Konvention vor allem dann, wenn keine anderen Evaluationskriterien – wie z. B. empirische Verteilungen der ICC(1) zu einer bestimmten Forschungsfrage – vorliegen (vgl. Cohen 1988).

Die ICC(1) ist ein wichtiges Maß, um zu bestimmen, inwiefern sich die auf Schulebene gemittelten Urteile der Schülerinnen und Schüler zur Unterrichtsqualität systematisch zwischen Schulen unterscheiden. Die ICC(1) informiert jedoch nicht über die Reliabilität dieser Mittelwerte und berücksichtigt nicht die Anzahl an Schülerurteilen, die in die Berechnung der Mittelwerte einging. Hier setzt die ICC(2) an, die angibt, wie reliabel die Schulmittelwerte zur Unterrichtsqualität ausfallen, die auf Grundlage der Urteile der Schülerinnen und Schüler berechnet wurden (Lüdtke et al. 2006). Die ICC(2) wird wie folgt berechnet (Bliese 2000):

$$\mathrm{ICC}(2) = \frac{k \cdot \mathrm{ICC}(1)}{1 + (k-1) \cdot \mathrm{ICC}(1)} \tag{3}$$

Hierbei steht k für die durchschnittliche Anzahl der Schülerinnen und Schüler pro Schule (Stapleton et al. 2016). Aus Gleichung 3 wird ersichtlich, dass der Wert der ICC(2) zunimmt, wenn (a) sich Schulen stärker systematisch hinsichtlich ihrer Unterrichtsqualität unterscheiden oder wenn (b) der Schulmittelwert auf einer größeren Zahl von Urteilen der Schülerinnen und Schüler pro Schule basiert. LeBreton und Senter (2008) beschreiben, dass als untere Grenze akzeptabler Werte der Reliabilität häufig ein Wert von .70 angesehen wird. Diese Grenze wird jedoch erneut als kritisch und lediglich als eine „Heuristik“ (LeBreton und Senter 2008) betrachtet.

2.2 Theorien und Befunde zu Unterschieden der Unterrichtsqualität von Schulen

Die Unterrichtsqualität ist ein zentrales Maß zur Charakterisierung schulischer Lernumwelten. Dass sich die Lernumwelten zwischen Schulen unterscheiden, ist in mehreren Strömungen der Bildungsforschung relevant, in denen implizit oder explizit von der Variabilität zwischen Schulen ausgegangen wird. So schreibt zum Beispiel Fend, „dass selbst im Rahmen gleicher organisationeller, administrativer und curricularer Strukturen sehr unterschiedliche

Gestalten des Schullebens und des Unterrichtens entstehen können“ (Fend 2008, S. 153). Unterschiede in der Unterrichtsqualität sind dabei unter anderem in Modellen der Schulqualität bzw. Schuleffektivität zentral (Scheerens und Bosker 1997; Ditton 2000). Denn in diesen Modellen stellen Unterschiede im Unterricht einen wichtigen Faktor dar, der Unterschiede in der Effektivität der Schulen erklären soll. Die Variabilität von Schulen ist weiterhin insbesondere auch für solche Forschungsfragen relevant, die sich mit der Bedeutung der Zusammensetzung (Komposition) der Schülerschaft (z. B. hinsichtlich ihres sozioökonomischen Hintergrunds oder ihrer Leistungsfähigkeit) für die Leistungs- und Persönlichkeitsentwicklung beschäftigen. Im Rahmen dieser Forschung wurden, z. B. auf Grundlage der Daten der PISA-Studie 2000, neben Schulformen auch Einzelschulen als „differenzielle Lern- und Entwicklungsmilieus“ beschrieben (Baumert et al. 2003), was die Unterschiedlichkeit von Einzelschulen unterstreicht. In der pädagogischen Praxis sind Unterschiede der Unterrichtsqualität zwischen Schulen vor allem auch für die Evaluation von Schulen im Rahmen der Schulinspektion relevant.

Trotz ihrer Bedeutung für die empirische Bildungsforschung gibt es interessanterweise bislang wenig Studien, die sich systematisch mit der Beurteilungsübereinstimmung, systematischen Unterschieden in der Unterrichtsqualität zwischen Schulen aus Sicht der Schülerinnen und Schüler oder der Reliabilität von aggregierten Schülerurteilen befassen. Unterschiede zwischen Schulen wurden häufig mit Blick auf Leistungsunterschiede ausgewertet (z. B. Baumert et al. 2003). Unterschiede in der Unterrichtsqualität wurden häufig mit Blick auf systematische Unterschiede zwischen Schulklassen (z. B. Lüdtke et al. 2006), aber nicht im Hinblick auf Unterschiede zwischen Schulen analysiert. Eine der wenigen Studien zu dieser Fragestellung stammt von Klieme und Rakoczy (2003), die die internationalen PISA 2000 Daten analysierten. In dieser Arbeit unterschieden sich Schulen innerhalb der teilnehmenden Staaten systematisch hinsichtlich ihrer Unterrichtsqualität. Über alle Staaten hinweg lag der Median der ICC(1)-Werte für „Unterstützung“ und „Beziehungsqualität“ jeweils bei .11 (für Deutschland: ICC(1) = .17 bzw. ICC(1) = .11). Bei dem Merkmal „Leistungsdruck“ lag der Median bei ICC(1) = .10 (Deutschland ICC(1) = .12) und für „Disziplinprobleme“ lag er bei ICC(1) = .13 (für Deutschland bei ICC(1) = .18). Sogar

innerhalb des gymnasialen Bildungsgangs, dessen Unterricht für viele einen hohen Grad der Standardisierung aufweisen sollte, unterschieden sich die Schulen in Deutschland hinsichtlich ihrer Unterrichtsmerkmale: die ICC(1)-Werte lagen hier zwischen .11 und .19 (Klieme und Rakoczy 2003). Übereinstimmungsmaße der Schülerurteile in den Schulen wurden nicht berichtet. Wurster und Gärtner (2013) untersuchten Unterrichtsurteile von Schülerinnen und Schülern, die im Rahmen der Schulinspektion abgegeben wurden. ICC(1)-Werte (für Unterschiede zwischen Schulen) lagen zwischen .06 (für die Skala „respektvoller Umgang“) und .23 („klare Struktur des Unterrichts“); der durchschnittliche ICC(1)-Wert lag bei .14 (Wurster und Gärtner 2013). Auch in dieser Studie wurden keine Übereinstimmungsmaße berichtet.

3. Forschungsfragen

Die Analyse der Unterrichtsqualität ist zentral für die Grundlagenforschung (empirische Schul-, Unterrichts- und Schuleffektivitätsforschung) sowie für die pädagogische Praxis (z. B. die Evaluation von Schulen durch die Schulinspektion). Eine etablierte Methode zur Erfassung der Unterrichtsqualität stellen Schülerurteile dar, die auf Schulebene gemittelt werden, um die Qualität eines bestimmten Unterrichtsmerkmals an einer bestimmten Schule zu bestimmen. Je nach Anwendungs- oder Forschungskontext sollten hierfür eine oder mehrere Voraussetzungen erfüllt sein: (1) es sollte eine angemessene Übereinstimmung der Schülerurteile innerhalb der Schule, (2) eine hinreichende Variation der aggregierten Werte zwischen den Schulen sowie (3) eine hinreichende Reliabilität der aggregierten Unterrichtsmerkmale vorliegen. Bislang gibt es jedoch nur wenige Studien, die systematisch untersuchen, in welchem Ausmaß diese Voraussetzungen erfüllt sind. Dies ist insbesondere auch deshalb relevant, da bislang häufig nur Konventionen aus anderen Forschungs- oder Anwendungskontexten vorliegen und nicht klar ist, inwiefern diese die empirischen Verteilungen der Übereinstimmung, Variabilität und Reliabilität der Unterrichtsqualität aus Schülersicht approximieren. In der vorliegenden Studie untersuchen wir deshalb auf Grundlage der Daten der PISA-Studien der Jahre 2000 bis 2012 die folgenden Forschungsfragen: (1) Inwieweit stimmen Schülerurteile von Unterrichtsmerkmalen in den Schulen überein? (2) Wie stark unterscheidet sich die

Unterrichtsqualität aus Sicht der Schülerinnen und Schüler zwischen Schulen? (3) Wie reliabel sind die aggregierten Schülerurteile zur Unterrichtsqualität? Auf Basis der Ergebnisse dieser drei Forschungsfragen kann der Forderung bspw. von LeBreton und Senter (2008) nach der Berücksichtigung des spezifischen Anwendungskontextes nachgekommen werden: Wir diskutieren die bestehenden Konventionen zur Beurteilung der Höhe von Übereinstimmung, Variabilität sowie Reliabilität im Kontext der Evaluation von Unterrichtsqualität auf Schulebene.

4. Methode

4.1 Stichprobe/Datengrundlage

Die Analysen dieser Studie basieren auf den Daten der 15-jährigen Schülerinnen und Schüler, die an den internationalen PISA-Erhebungen in den Jahren 2000, 2003, 2009 und 2012 teilnahmen (bei PISA 2006 wurde die Unterrichtsqualität nicht aus Schülersicht erfasst). PISA wird von der OECD initiiert und untersucht primär die Leistungen der 15-Jährigen in den Bereichen Lesekompetenz, mathematische und naturwissenschaftliche Kompetenz im 3-jährlichen Zyklus. In jedem Zyklus wurden aus meist mehrfach stratifizierten Listen mindestens 150 Schulen innerhalb eines Staates zufällig gezogen. Innerhalb der Schulen wurden zufällig 35 Schülerinnen und Schüler, die zum Testzeitpunkt rund 15 Jahre alt waren, zur Teilnahme ausgewählt (OECD 2002). Die Gesamtstichprobengröße für die vorliegenden Analysen der jeweiligen PISA-Zyklen lag zwischen 223 649 15-Jährigen aus 43 Ländern mit insgesamt 8 492 Schulen (bei PISA 2000) und 506 803 15-Jährigen aus 74 Ländern mit insgesamt 18 519 Schulen (bei PISA 2009; s. Tabelle 1). Hierbei ist anzumerken, dass bei PISA 2012 ein Rotationsdesign für den Schülerfragebogen angewendet wurde, sodass die Items zu Unterrichtsmerkmalen jeweils nur von zwei Drittel der PISA-Stichprobe beantwortet wurden. Insgesamt gingen somit Daten von über 1,3 Millionen Jugendlichen aus über 55 300 Schulen aus 81 verschiedenen Ländern in die vorliegende Studie ein.

Tabelle 1

Angaben zur Schüler- und Schulanzahl sowie zu fehlenden Werten bei den erfassten Unterrichtsmerkmalen in PISA 2000, 2003, 2009 und 2012

	Klassenführung					Kogn. Aktiv.	Konstruktive Unterstützung											
	Disziplin (S)		Disziplin (M)		Klassenführung (M)	kognitive Aktivierung (M)	Unterstützung (S)	Unterstützung (M)		Leistungsdruck (S)	Lehrersteuerung (M)	Schülerorientierung (M)	Strukturierung (S)	Rückmeldung (M)	Beziehungsqualität			
	2000[a]	2009[b]	2003[a]	2012[a]	2012[b]	2012[a]	2000[a]	2003[a]	2012[a]	2000[a]	2012[a]	2012[a]	2009[a]	2012[a]	2000[b]	2003[b]	2009[b]	2012[b]
pro Teilnahmeland																		
$N_{SuS\ Min}$	313	328	330	185	189	191	311	330	185	312	190	190	327	188	308	331	328	185
$N_{SuS\ Max}$	29 122	37 807	28 743	22 254	22 127	22 202	29 115	29 152	22 308	29 118	22 239	22 223	37 634	22 201	29 205	29 710	37 919	22 178
$N_{Sch\ Min}$	11	12	12	12	12	12	11	12	12	11	12	12	12	12	11	12	12	12
$N_{Sch\ Max}$	1 117	1 535	1 122	1 455	1 452	1 455	1 117	1 122	1 455	1 117	1 455	1 455	1 534	1 455	1 117	1 124	1 535	1 454
Gesamtanzahl																		
Miss %	2	2	3	1	2	1	2	3	1	2	1	1	3	1	2	2	2	1
N_{SuS}	224 122	505 421	267 893	314 775	312 708	314 553	223 789	268 903	316 369	224 010	315 518	315 114	502 859	314 678	223 649	271 777	506 803	313 857
N_{Sch}	8 493	18 515	10 216	18 101	18 099	18 106	8 492	10 216	18 104	8 493	18 110	18 108	18 512	18 103	8 494	10 222	18 519	18 097
$N_{Länder}$	43	74	41	68	68	68	43	41	68	43	68	68	74	68	43	41	74	68

Anmerkungen. (S) = fachspezifische Skala zum Sprachunterricht, (M) = fachspezifische Skala zum Mathematikunterricht. Die Angaben zu fehlenden Werten bezieht sich jeweils nur auf die Schülerinnen und Schüler, denen eine Skala zur Beurteilung eines Unterrichtsmerkmals tatsächlich vorlag und wird über alle Länder angegeben.
SuS = Schüler und Schülerinnen; Sch = Schulen; Miss = fehlende Werte

[a]: verwendete Antwortskala: „in jeder Stunde", „in den meisten Stunden", „in manchen Stunden" und „nie oder fast nie"; bei der Skala zur kognitiven Aktivierung (M) 2012 wurden die Antwortkategorien leicht variiert („immer oder fast immer", „oft", „manchmal", „nie oder fast nie")

[b]: verwendete Antwortskala: „stimme überhaupt nicht zu", „stimme eher nicht zu", „stimme eher zu", „stimme ganz zu"

4.2 Instrumente

In den PISA-Studien wurden Schülerfragebögen zur Erfassung von Unterrichtsmerkmalen eingesetzt, die von Expertenteams entwickelt und in Feldtests intensiv psychometrisch überprüft wurden. Die Analysen der vorliegenden Studie beziehen alle Unterrichtsskalen mit ein, die in den Erhebungen der PISA-Studie zwischen 2000 und 2012 zum Einsatz kamen (s. Tabelle 1). Dabei wurden die Schülerinnen und Schüler befragt, wie häufig bestimmte Situationen im Unterricht auftreten oder wie stark sie einer Aussage zum Unterricht zustimmen. Alle Antwortskalen waren vierstufig, die Antwortkategorien variierten leicht zwischen den Erhebungszeitpunkten und Merkmalen (vgl. Tabelle 1). Die nachfolgenden Angaben zum Median bzw. Mittelwert der internen Konsistenz (Cronbachs Alpha α) der Unterrichtsskalen haben wir aus den technischen Berichten der OECD entnommen (OECD 2002, 2005, 2012, 2014). Die Unterrichtsskalen ordneten wir den drei Basisdimensionen von Unterrichtsqualität wie folgt zu.

4.2.1 Skalen zur Klassenführung

Eine Skala zur Disziplin (Gesamtitemzahl: 5 Items; Beispielitem: „Wir Schülerinnen und Schüler hören nicht auf das, was der Lehrer/die Lehrerin sagt.“) wurde fachspezifisch für den Unterricht der jeweiligen Landessprache in 2000 ($\alpha_{M\,OECD} = .81$) und 2009 ($\alpha_{Med\,OECD} = .88$) verwendet, für den Mathematikunterricht wurde sie 2003 ($\alpha_{Med\,OECD} = .83$) und 2012 ($\alpha_{Med\,OECD} = .89$) eingesetzt. Im Jahr 2012 wurde zudem eine Skala zur Klassenführung im Mathematikunterricht eingesetzt (4 Items; „Mein/e Lehrer/in sorgt für Ordnung in der Klasse.[3]“, $\alpha_{Med\,OECD} = .72$).

4.2.2 Skala zur kognitiven Aktivierung

Im PISA-Durchgang 2012 wurde eine Skala eingesetzt (9 Items; „Der Lehrer/die Lehrerin stellt Aufgaben, die durch verschiedene Wege gelöst werden können.“), in der es darum geht, inwieweit die Schülerinnen und Schüler durch die Lehrkraft im Mathematikunterricht kognitiv herausgefordert werden ($\alpha_{Med\,OECD} = .83$).

[3] Eigene Übersetzung der Items von 2012, da hier der deutsche Fragebogen nicht öffentlich zugänglich ist.

4.2.3 Skalen zur konstruktiven Unterstützung

In der Skala zur Unterstützung wird erhoben, wie ausgeprägt die Unterstützung der Lehrkraft im Unterricht durch die Schülerinnen und Schüler wahrgenommen wird (5 Items; „Unser Lehrer/unsere Lehrerin interessiert sich für den Lernfortschritt jedes einzelnen Schülers/jeder Schülerin."). Im Jahr 2000 wurde diese Skala ($\alpha_{M\,OECD}$ = .87) für den Unterricht in der Landessprache eingesetzt, 2003 ($\alpha_{Med\,OECD}$ = .83) und 2012 für den Mathematikunterricht ($\alpha_{Med\,OECD}$ = .85).

In PISA 2000 wurde eine Skala zum Leistungsdruck (4 Items; „Unser Lehrer/unsere Lehrerin will, dass wir uns richtig anstrengen.") für den Unterricht der Landessprache erhoben ($\alpha_{M\,OECD}$ = .54). In PISA 2012 kam eine Skala zum Lehrerverhalten im Mathematikunterricht zum Einsatz. Darin geht es um lehrerzentrierte Instruktion (5 Items; „Am Anfang der Stunde fasst der Lehrer/die Lehrerin kurz die letzte Stunde zusammen.", $\alpha_{Med\,OECD}$ = .73). Weiterhin wurde im Jahr 2012 eine fachspezifische Skala zur Schülerorientierung im Mathematikunterricht verwendet (4 Items; „Der Lehrer/die Lehrerin lässt uns in Kleingruppen arbeiten um zu gemeinsamen Lösungen für ein Problem oder eine Aufgabe zu gelangen.", $\alpha_{Med\,OECD}$ = .68). In den Erhebungen 2009 wurde eine Skala zu Strukturierungsstrategien der Lehrkräfte im Unterricht der jeweiligen Landessprache eingesetzt (9 Items; „Unsere Lehrerin/unser Lehrer gibt den Schülerinnen/Schülern die Möglichkeit, Fragen zu den Leseaufgaben zu stellen.", $\alpha_{Med\,OECD}$ =.83). In 2012 wurde eine Skala zum Verhalten der Lehrkraft bei Rückmeldungen an die Schülerinnen und Schüler eingesetzt (4 Items; „Der Lehrer/die Lehrerin erklärt uns, was von uns erwartet wird, wenn wir einen Test, ein Quiz oder einen Arbeitsauftrag bekommen."). Diese Skala wurde fachspezifisch im Mathematikunterricht erhoben ($\alpha_{Med\,OECD}$ = .76). Die fachunabhängige Skala zur Beziehungsqualität (5 Items; „Die meisten meiner Lehrer/Lehrerinnen interessieren sich für das, was ich zu sagen habe.") zwischen Schülerinnen und Schülern und Lehrkraft wurde in allen hier analysierten PISA-Durchgängen herangezogen: PISA 2000 ($\alpha_{M\,OECD}$ = .79), 2003 ($\alpha_{Med\,OECD}$ = .76), 2009 ($\alpha_{Med\,OECD}$ = .83) und 2012 ($\alpha_{Med\,OECD}$ = .83).

4.3 Statistische Analysen

Um die Frage zu beantworten, inwieweit Schülerurteile von Unterrichtsmerkmalen in Schulen übereinstimmen, berechneten wir mit dem R-Paket „multilevel“ (Bliese 2013) für jede Einzelschule und für jede Skala zur Unterrichtsqualität den $r_{WG(J)}$ Index über die Einzelitems. Für die Ergebnisdarstellung haben wir für ein bestimmtes Unterrichtsmerkmal die $r_{WG(J)}$ Indizes aller Schulen innerhalb eines Landes gemittelt. Darüber hinaus berechneten wir auch (bezogen auf alle Schulen, die an einem bestimmten PISA-Zyklus teilnahmen) den prozentualen Anteil an Schulen, deren $r_{WG(J)}$-Wert über dem Grenzwert von $r_{WG} = .70$ lag.

Um die Fragen zu beantworten, wie stark sich Schulen hinsichtlich der Unterrichtsqualität unterschieden bzw. wie reliabel die aggregierten Schülerurteile zur Unterrichtsqualität waren, berechneten wir Werte für die ICC(1) und ICC(2) mit dem R-Paket „lme4“ (Bates et al. 2015). Als Skalenwerte aller Unterrichtsmerkmale nutzten wir die WLE-Personenschätzer, die in den public-use Datensätzen der OECD zur Verfügung gestellt wurden. Höhere WLEs drücken dabei eine höhere Ausprägung des Skalenwerts aus (OECD 2002, 2005, 2012, 2014). Anzumerken ist hier, dass im technischen Bericht von PISA 2012 (OECD 2014) für die Unterrichtsskalen von PISA 2012 und 2003 ICC(1)-Werte für OECD-Länder angegeben werden. Diese weichen jedoch geringfügig von den hier berichteten Werten ab, da nicht genau dieselben Auswertungsmethoden verwendet wurden. Für ICC(2) geben wir an, in wie vielen Ländern der Wert über dem Grenzwert von ICC(2) = .70 lag. Weiterhin ermittelten wir für die ICC-Werte 95 %-Konfidenzintervalle (s. Online Supplement (OS)) mittels Bootstrapverfahren (bootMer in lme4), da bei Intraklassenkorrelationen keine etablierten analytischen Verfahren zur Bestimmung der Standardfehler existieren (Bates 2010; Bates et al. 2015). Nach sorgfältiger Überlegung haben wir uns dafür entschieden, bei der Nutzung dieser Verfahren auf die Verwendung der Stichprobengewichte zu verzichten[4]. Zudem wurde für Unterrichtsmerkmale, die in mehreren PISA-Zyklen erhoben wurden, die Varianz zwischen den Ländern und

[4] Die 95 %-Konfidenzintervalle könnten sonst nicht berechnet werden. Zudem wurden in Analysen mit einem Teil der Daten bei der ICC(1)-Berechnung Stichprobengewichte verwendet. Diese ergaben lediglich sehr geringe Abweichungen von maximal 0.03. Das Nicht-Verwenden der Stichprobengewichte hat jedoch zur Folge, dass die Befunde nur auf die Grundgesamtheit der teilnehmenden Schülerinnen und Schüler in den Ländern verallgemeinert werden können und insgesamt weniger präzise ausfallen, als unter Einbeziehung der Stichprobengewichte.

zwischen den Zeitpunkten berechnet (s. OS): Somit ist es möglich zu bestimmen, inwiefern Unterschiede in $r_{WG(J)}$, ICC(1) bzw. ICC(2) zwischen den Ländern zeitlich stabile Merkmale der Länder sind oder auch innerhalb der Länder zwischen den PISA-Zyklen variieren. Je höher der Varianzanteil, der auf die Länder zurückgeht, desto zeitlich stabiler sind differentielle Unterschiede in $r_{WG(J)}$, ICC(1) bzw. ICC(2) zwischen den Ländern. Um die Ergebnisse der Intraklassenkorrelationen zusammenzufassen, haben wir zusätzlich zur deskriptiven Beschreibung der Verteilung der jeweiligen Parameter meta-analytische Kennwerte mit dem R-Paket „metafor" (Viechtbauer 2016) berechnet: μ als durchschnittlichen Wert von ICC(1) bzw. ICC(2), der die Schätzgenauigkeit der PISA-Erhebungen in den verschiedenen Ländern berücksichtigt; I^2 um die Variabilität zwischen den Intraklassenkorrelationen, die auf Heterogenität beruht, in Prozent anzugeben; Q als Cochrans Q-Teststatistik zur inferenzstatistischen Überprüfung der Heterogenität.

Da der prozentuale Anteil der fehlenden Werte sehr gering ausfiel (Tabelle 1), wurden fehlende Fälle für die Analysen zu einer bestimmten Unterrichtsskala fallweise ausgeschlossen.

5. Ergebnisse

5.1 Forschungsfrage 1: Übereinstimmung von Schülerurteilen innerhalb von Schulen

Abbildung 1 stellt die Verteilung der Kennwerte, die wir in der vorliegenden Studie untersuchten, im Überblick dar. Jeder Punkt im oberen Segment von Abbildung 1 steht hier für die über alle Schulen eines Landes gemittelten $r_{WG(J)}$ -Werte einer bestimmten Skala bei einem bestimmten PISA-Zyklus. Wie aus Abbildung 1 ersichtlich ist, zeigten sich zum Teil große Unterschiede zwischen den Ländern und zwischen den untersuchten Unterrichtsmerkmalen: Die Verteilungen streuten zwischen $r_{WG(J)} = .29$ (Schülerorientierung (M) 2012 in Tunesien) und $r_{WG(J)} = .94$ (Beziehungsqualität 2003 in Thailand). Die landesspezifischen Ergebnisse aller Kennwerte sind im Online Supplement zu finden. Im Folgenden weisen wir die Wertespannen sowie die Einzelwerte, die an Schulen in Deutschland (DE) ermittelt wurden, in Klammern aus.

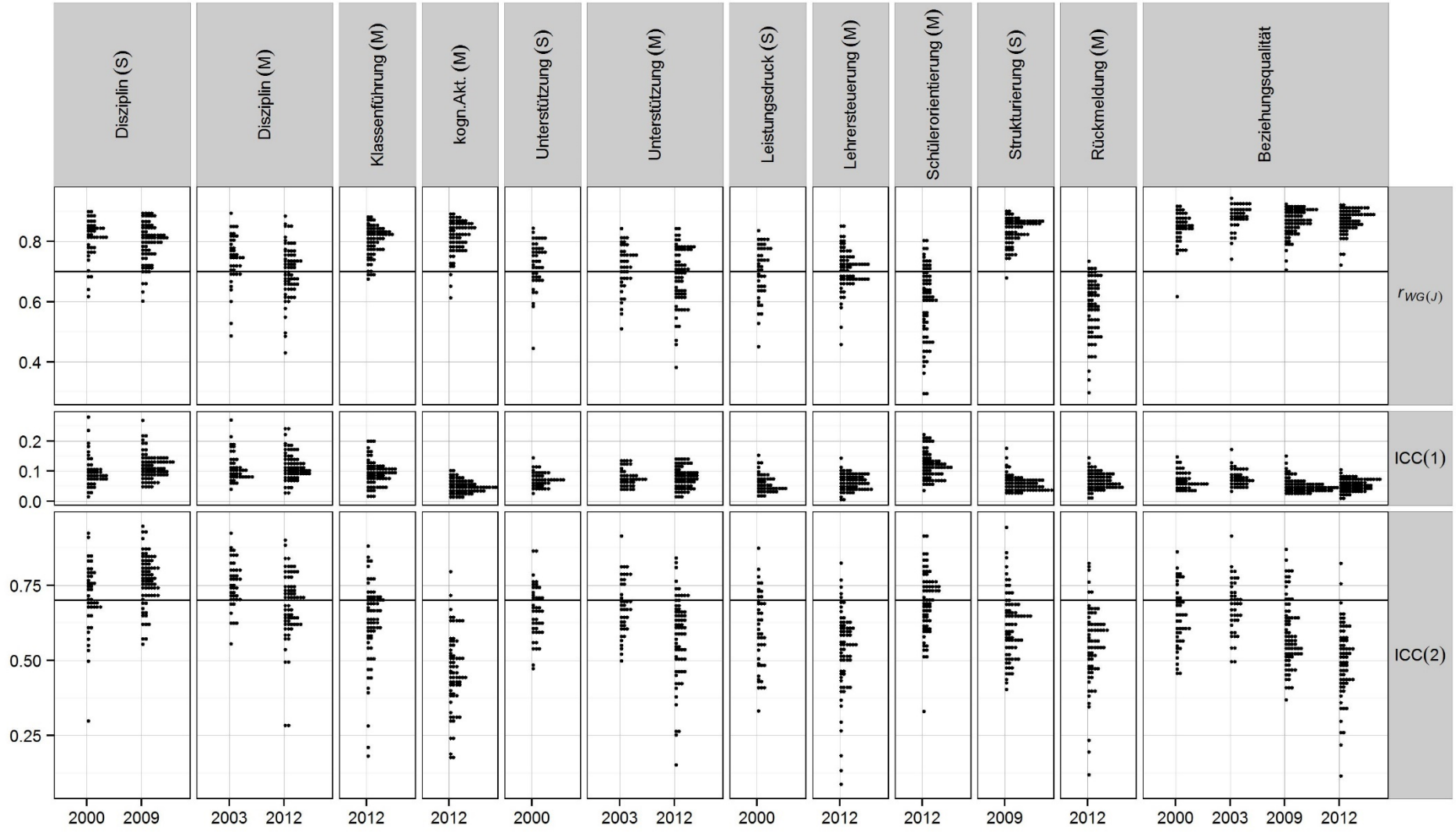


Abbildung 1: Verteilungen der Kennwerte im Überblick

Trotz dieser teils starken Variabilität der $r_{WG(J)}$-Werte, die in Abbildung 1 zwischen den Ländern zu beobachten ist, waren die Übereinstimmungsindizes fast aller untersuchten Unterrichtsmerkmale im oberen Wertebereich lokalisiert. Wie aus Tabelle 2 ersichtlich ist, lagen die Mediane der Verteilungen für Merkmale der Basisdimension Klassenführung zwischen $r_{WG(J)} = .71$ (DE: .62) und $r_{WG(J)} = .82$ (DE: .78); der Median für die Skala kognitive Aktivierung lag bei $r_{WG(J)} = .83$ (DE: .84); die Mediane der Merkmale der Basisdimension konstruktive Unterstützung variierten zwischen $r_{WG(J)} = .59$ (DE: .52) und $r_{WG(J)} = .89$ (DE: .83). Bei Unterrichtsmerkmalen, die über mehrere Jahre erhoben wurden, zeigte sich nahezu keine Varianz zwischen den Erhebungszeitpunkten (s. OS). Mit Blick auf den „Grenzwert" von $r_{WG} = .70$ für eine mindestens „akzeptable" Übereinstimmung stellen wir folgendes fest: Der überwiegende Teil der auf Landesebene gemittelten Kennwerte und der überwiegende Teil der Schulen lag für die Mehrzahl der untersuchten Unterrichtsmerkmale über dieser Marke. Anzumerken ist hierbei aber, dass an zahlreichen Schulen innerhalb der Länder, die Übereinstimmungsgüte der Schülerurteile unter dem Grenzwert von $r_{WG} = .70$ lag. Auffällig sind insbesondere die Unterrichtsmerkmale Schülerorientierung ($r_{WG(J)} = .62$; DE: .67) und Rückmeldung ($r_{WG(J)} = .59$; DE: .63), bei denen in relativ vielen Ländern an der Mehrzahl der Schulen der Wert von .70 unterschritten wurde.

Tabelle 2

Verteilungen der Kennwerte $r_{WG(J)}$, ICC(1) und ICC(2) über die Teilnehmerstaaten an PISA

	Klassenführung					Kogn. Aktiv.	Konstruktive Unterstützung											
	Disziplin (S)		Disziplin (M)		Klassen-führung (M)	kognitive Aktivierung (M)	Unter-stützung (S)	Unter-stützung (M)		Leistungs-druck (S)	Lehrersteue-rung (M)	Schülerorien-tierung (M)	Struk-turierung (S)	Rückmel-dung (M)	Beziehungs-qualität			
	2000	2009	2003	2012	2012	2012	2000	2003	2012	2000	2012	2012	2009	2012	2000	2003	2009	2012
$r_{WG(J)}$																		
Minimum	.62	.60	.49	.43	.68	.61	.45	.51	.38	.45	.46	.29	.68	.30	.62	.74	.71	.72
25%-Perzentil	.78	.76	.70	.65	.77	.78	.68	.67	.62	.64	.67	.51	.81	.51	.81	.87	.84	.85
Median	.82	.81	.75	.71	.81	.83	.74	.72	.70	.72	.72	.62	.85	.59	.85	.89	.87	.88
75%-Perzentil	.85	.85	.81	.76	.84	.86	.78	.77	.77	.78	.75	.69	.87	.66	.88	.91	.90	.90
Maximum	.90	.90	.89	.89	.89	.89	.85	.84	.85	.84	.85	.81	.90	.73	.92	.94	.93	.92
$N_{Schulen} > .7$ in %	88	83	73	63	86	87	73	68	62	66	63	46	94	34	94	98	96	96
ICC(1)																		
μ	.10	.11	.11	.12	.09	.05	.07	.08	.08	.06	.06	.12	.06	.07	.07	.08	.05	.05
Konfidenzintervall	[.08; .11]	[.10; .12]	[.10; .12]	[.10; .13]	[.08; .10]	[.04; .05]	[.06; .08]	[.07; .09]	[.07; .08]	[.05; .06]	[.06; .07]	[.11; .13]	[.05; .06]	[.06; .07]	[.06; .07]	[.07; .09]	[.04; .05]	[.05; .06]
I^2 in %	94	92	93	91	91	82	82	89	89	90	85	91	85	84	87	88	82	76
Q	424	778	333	553	580	385	204	307	628	304	458	603	407	388	253	263	336	265
Minimum	.02	.04	.04	.03	.02	.01	.03	.04	.01	.01	.00	.04	.02	.01	.03	.03	.02	.01
25%-Perzentil	.07	.09	.08	.09	.07	.03	.05	.06	.05	.04	.04	.09	.04	.05	.04	.06	.04	.04
Median	.08	.11	.10	.11	.10	.05	.07	.07	.08	.06	.06	.12	.06	.07	.06	.07	.05	.06
75%-Perzentil	.11	.14	.14	.14	.12	.06	.09	.10	.10	.07	.08	.14	.07	.09	.09	.10	.06	.07
Maximum	.28	.27	.27	.24	.20	.10	.14	.14	.14	.15	.14	.22	.18	.14	.15	.17	.15	.10
ICC(2)																		
μ	.73	.77	.76	.70	.64	.47	.67	.68	.60	.60	.55	.70	.62	.57	.65	.69	.59	.51
Konfidenzintervall	[.70; .75]	[.75; .79]	[.73; .78]	[.67; .72]	[.62; .67]	[.45; .50]	[.64; .70]	[.66; .71]	[.57; .62]	[.56; .64]	[.52; .58]	[.68; .72]	[.59; .64]	[.54; .59]	[.62; .68]	[.66; .71]	[.57; .62]	[.48; .53]
I^2 in %	93	94	92	92	92	84	84	90	90	90	87	93	90	88	86	89	88	81
Q	617	1156	556	735	599	419	270	357	516	455	392	902	993	423	284	298	622	303
Minimum	.30	.55	.56	.28	.18	.18	.47	.50	.15	.33	.09	.33	.40	.12	.46	.49	.37	.12
25%-Perzentil	.67	.72	.70	.62	.58	.39	.60	.61	.50	.49	.46	.62	.54	.47	.56	.62	.52	.43
Median	.72	.77	.75	.71	.64	.46	.67	.67	.61	.59	.55	.70	.62	.56	.64	.69	.57	.51
75%-Perzentil	.79	.83	.80	.75	.71	.54	.73	.76	.67	.70	.61	.76	.68	.62	.71	.74	.65	.57
Maximum	.92	.95	.92	.90	.88	.80	.87	.91	.84	.87	.82	.91	.94	.82	.86	.91	.87	.82
$N_{Länder} > .7$ in %	53	82	80	51	31	3	37	34	18	23	7	47	16	9	28	44	19	3

Anmerkungen. (S) = fachspezifische Skala zum Sprachunterricht, (M) = fachspezifische Skala zum Mathematikunterricht; Q(df) bei Merkmalen 2000: 42, Q(df) bei Merkmalen 2003: 40; Q(df) bei Merkmalen 2009: 73, Q(df) bei Merkmalen 2012: 67; für die $r_{WG(J)}$-Kennwerte konnten μ, I^2 und Q nicht berechnet werden, da keine Standardfehler für diesen Kennwert vorliegen

5.2 Forschungsfrage 2: Unterschiede zwischen Schulen

Das mittlere Segment in Abbildung 1 stellt die Verteilungen der Intraklassenkorrelationen ICC(1) der untersuchten Unterrichtsmerkmale in den jeweiligen Jahren dar. Dabei repräsentiert ein Punkt den ICC(1)-Wert für ein bestimmtes Unterrichtsmerkmal in einem bestimmten Land im jeweiligen PISA-Zyklus. Insgesamt zeigten sich in fast allen Ländern bzw. für fast alle Unterrichtsmerkmale statistisch bedeutsame (s. 95%-Konfidenzintervalle im OS) Unterschiede zwischen den Schulen. Die Werte der Intraklassenkorrelationen streuten zwischen ICC(1) = .00 (Lehrersteuerung (M) 2012 im US-Staat Connecticut) und ICC(1) = .28 (Disziplin (S) 2000 in Island). Wie aus Tabelle 2 ersichtlich ist, lagen die Mediane der ICC(1)-Werte für Merkmale der Basisdimension Klassenführung zwischen .08 (DE: .05) und .11 (DE: .10); der Median ICC(1)-Wert für die Skala kognitive Aktivierung lag bei .05 (DE: .04); die Mediane der ICC(1)-Werte der Merkmale der Basisdimension konstruktive Unterstützung variierten zwischen .05 (DE: .04) und .12 (DE: .16). Die durchschnittlichen ICC(1)-Werte waren dabei nahezu identisch mit den Medianwerten. Die Werte von I^2 weisen auf eine starke Heterogenität der ICC(1)-Werte zwischen den Ländern hin, die (wie die Q-Teststatistiken zeigen) auch statistisch bedeutsam von Null verschieden waren. Bemessen an den bestehenden Konventionen zur Beurteilung der ICC(1)-Werte waren die Unterschiede der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler zwischen den Schulen also meist als „klein“ oder „mittel“ zu bezeichnen. Bei Unterrichtsmerkmalen, die über mehrere Jahre erhoben wurden, lag nur ein sehr geringer Teil der Varianz der ICC(1)-Werte zwischen den Erhebungszeitpunkten (s. OS).

5.3 Forschungsfrage 3: Reliabilität der Schulmittelwerte

Mit Blick auf die Reliabilität der aggregierten Schülerurteile zeigten sich teilweise große Unterschiede zwischen den untersuchten Unterrichtsmerkmalen bzw. zwischen den Ländern für ein bestimmtes Merkmal (s. Abbildung 1, unteres Segment sowie I^2-Werte und zugehörige Q-Teststatistiken in Tabelle 2).

Die Werte der Intraklassenkorrelationen streuten zwischen ICC(2) = .09 (Lehrersteuerung (M) 2012 im US-Staat Connecticut) und ICC(2) = .95 (Disziplin (S) 2009 in

Macao-China). Wie aus Tabelle 2 ersichtlich ist, lagen die Mediane der ICC(2)-Werte für Merkmale der Basisdimension Klassenführung zwischen .64 (DE: .39) und .77 (DE: .71); der Median ICC(2)-Wert für die Skala kognitive Aktivierung lag bei = .46 (DE: .32); die Mediane der ICC(2)-Werte der Merkmale der Basisdimension konstruktive Unterstützung variierten zwischen .51 (DE: .40) und .70 (DE: .70). Die durchschnittlichen ICC(2)-Werte waren dabei nahezu identisch mit den Medianwerten. Bei Unterrichtsmerkmalen, die über mehrere Jahre erhoben wurden, zeigten sich kleine bis mittlere Unterschiede zwischen den Erhebungszeitpunkten (s. OS). Mit Blick auf den „Grenzwert" von ICC(2) = .70 für eine mindestens „akzeptable" Reliabilität stellen wir folgendes fest: In einem Großteil der Länder lag die Reliabilität der Schulmittelwerte der untersuchten Unterrichtsmerkmale unter ICC(2) = .70: Zwischen 18% (Disziplin in 2009) und 97% (Beziehungsqualität in 2012) der Länder unterschritten diese Schranke bei einem bestimmten PISA-Zyklus.

6. Diskussion

Die Analyse der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler ist zentral im Rahmen der empirischen Schul-, Unterrichts- und Schuleffektivitätsforschung sowie für die pädagogische Praxis, z. B. im Rahmen der externen Evaluation von Einzelschulen durch die Schulinspektion. Je nach Forschungs- bzw. Anwendungskontext ist die Nutzung von auf Schulebene aggregierten Schülerurteilen an eine oder mehrere Voraussetzungen gebunden. Wir untersuchten deshalb auf Grundlage der Daten der PISA-Studien (1) den Grad der Übereinstimmung von Schülerurteilen der Unterrichtsqualität, (2) die Variabilität der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler zwischen Schulen und (3) die Reliabilität der aggregierten Schülerurteile. Bei der nachfolgenden Diskussion der Ergebnisse stehen die Konventionen zur Interpretation der Höhe der Kennwerte $r_{WG(J)}$, ICC(1) und ICC(2) im Mittelpunkt.

6.1 Konventionen zur Beurteilung der Übereinstimmung, von Schulunterschieden und der Reliabilität

Zur Interpretation der Werte des Übereinstimmungsmaßes $r_{WG(J)}$, des Variabilitätsmaßes ICC(1) und des Reliabilitätsmaßes ICC(2) nutzten wir gängige Konventionen. Diese reflektieren wir nun vor dem Hintergrund der Ergebnisse aus der vorliegenden Studie für Anwendungskontexte, in denen Schülerurteile genutzt werden, um die Unterrichtsqualität von Schulen zu erfassen.

Als Grenzwert für eine akzeptable Übereinstimmung von Schülerurteilen wird häufig ein r_{WG} -Wert von .70 genannt, um die Aggregation der Schülerurteile zu einem Schulmittelwert zu rechtfertigen. Dieser Grenzwert wurde bereits in zahlreichen Arbeiten kritisiert und LeBreton und Senter (2008) haben deshalb eine differenzierte Klassifikation vorgeschlagen (s. Abschnitt 2.1). Ausgehend von dieser Klassifikation zeigte sich für die meisten der untersuchten Unterrichtsmerkmale in der großen Mehrzahl der Länder eine „moderate“ und meist sogar „starke“ Übereinstimmung der Schülerurteile in den Schulen. Ausgenommen hiervon waren lediglich zwei Unterrichtsmerkmale zur konstruktiven Unterstützung (Schülerorientierung und Rückmeldung). Bemerkenswert ist hierbei, dass bereits die mittleren $r_{WG(J)}$ -Werte, die das 25. Perzentil der jeweiligen Verteilungen markierten, alle über dem Grenzwert lagen, der eine „moderate“ Übereinstimmung indiziert. Bis auf die oben genannten zwei Ausnahmen lagen die Medianwerte der Übereinstimmungsindizes sogar im Bereich, der als „starke“ Übereinstimmung beurteilt wird. Diese Ergebnisse unterstreichen, dass für die Analyse der Unterrichtsqualität auf Grundlage von Schülerurteilen die Heuristik von LeBreton und Senter realistische Erwartungen an die Qualität der Übereinstimmung dieser Urteile formuliert, die im Kontext von large-scale Assessments bzw. im Forschungskontext erreicht werden kann. Da Schülerfragebögen zur Unterrichtsqualität, wie sie in PISA eingesetzt werden, in ähnlicher Form auch für die externe Evaluation im Rahmen der Schulinspektion genutzt werden, stimmen diese Ergebnisse jedoch bedenklich. Denn für weitreichende Entscheidungen, die Einzelschulen betreffen, sollten deutlich höhere Übereinstimmungswerte von Schülerurteilen zur Unterrichtsqualität erzielt werden (LeBreton und Senter 2008). Dabei deuten die vorliegenden Ergebnisse insbesondere daraufhin, dass eine hohe Übereinstimmung

von Schülerurteilen bei bestimmten Merkmalen nicht an allen Schulen erwartet werden kann. Beispielsweise weisen die Merkmale Schülerorientierung und Rückmeldung in vielen Schulen (in vielen Ländern) geringe Übereinstimmungen auf. Mehrere Faktoren können hierfür einzeln oder auch kombiniert den Grad der Übereinstimmung von Schülerurteilen bestimmen: (a) Lehrkräfte passen ihre Rückmeldungen und ihre Fördermaßnahmen an die Lernvoraussetzungen und den -fortschritt ihrer Schülerinnen und Schüler differenziell an. Die individuell gemachten Erfahrungen im Klassenzimmer beeinflussen dann den Grad der Übereinstimmung (Gärtner 2010). Gut beobachtbare Unterrichtsmerkmale wie Disziplin und Klassenführung sind deutlich weniger von solchen individuellen Erfahrungen abhängig. (b) Bei PISA nehmen Schülerinnen und Schüler einer Schule aus unterschiedlichen Klassen teil. Es ist also gut möglich, dass an einzelnen Schulen Schülerinnen und Schüler von unterschiedlichen Lehrkräften unterrichtet werden, die sich in bestimmten Unterrichtsmerkmalen systematisch unterscheiden. (c) Die Itemformulierung spielt bei der Erfassung von Konstrukten in Mehrebenenkontexten eine besondere Rolle. So ist davon auszugehen, dass durch die Verwendung von „ich"/„mir" anstelle von „wir"/„uns" (wie z. B. bei Rückmeldung) nicht nur „geteilte" Konstrukte erfasst werden, die auf eine höhere Aggregatebene fokussieren, sondern auch Konstrukte gemessen werden, die die Individualebene betreffen (Stapleton et al. 2016). Für solche Skalen ist daher immer ein gewisses Maß an Heterogenität auf Individualebene innerhalb von Schulen zu erwarten. Insgesamt weisen die vorliegenden Ergebnisse also darauf hin, dass die Aussagekraft von aggregierten Schülerurteilen (z. B. in Form von Schulmittelwerten) bei bestimmten Unterrichtsmerkmalen in vielen Schulen begrenzt ist und die Wahrnehmung dieser Unterrichtsmerkmale vieler Schülerinnen und Schüler an diesen Schulen nur unzutreffend abbilden. Für solche Unterrichtsmerkmale stellt sich die grundsätzliche Frage, inwiefern sie durch ein Konsensusmodell zutreffend beschrieben werden, oder ob hier nicht von Dispersionsmodellen (s. z. B. Chan 1998) ausgegangen und entsprechende Indizes (z. B. die Standardabweichung der Schülerurteile in einer Schule; s. LeBreton und Senter 2008) berechnet werden sollten. Insbesondere für die Schulinspektionspraxis legt dies nahe, dass generell auch die Heterogenität der Unterrichtsqualität innerhalb einer Schule stärker in den Blick genommen (vgl. Wurster und

Gärtner 2013) und Kennwerte berechnet werden sollten, die diese Unterschiede innerhalb von Schulen quantifizieren (z. B. Dispersionsmaße). Generell unterstreichen die vorliegenden Ergebnisse zur Übereinstimmung, dass im Kontext der Evaluation stets eine Maßzahl der Beurteilerübereinstimmung mit angegeben werden sollte, um den Adressaten der Evaluationsergebnisse eine Einschätzung zu ermöglichen, inwiefern aggregierte Urteile die Wahrnehmung aller Schülerinnen und Schüler abbilden können (Lüdtke et al. 2006; Gärtner 2010).

In ihrer einschlägigen Arbeit interpretieren LeBreton und Senter (2008) Werte der ICC(1) = .01 als „kleinen“, ICC(1) = .10 als „mittleren“ und ICC(1) = .25 als „großen“ Effekt unabhängig vom Anwendungskontext. Bezogen auf diese Konvention zeigten sich systematische (und statistisch bedeutsame) Unterschiede in der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler zwischen den Schulen. Jedoch waren diese Unterschiede zwischen Schulen in nahezu allen Ländern und für alle Unterrichtsmerkmale als „klein“ oder „mittel“ zu beurteilen, was sich auch gut mit dem bisherigen Forschungsstand deckt (Klieme und Rakoczy 2003; Wurster und Gärtner 2013). Zu bedenken ist hierbei aber, dass in einschlägigen Arbeiten zur Effektgrößeninterpretation empfohlen wird, Konventionen für den jeweiligen Anwendungskontext abzuleiten (Cohen 1988). Um hierfür empirische Anhaltspunkte zu gewinnen, wendeten wir die Rationale von Hemphill (2003) an, der Beurteilungsmaßstäbe für Korrelationen untersuchte. Dieser Rationale folgend haben wir die ICC(1)-Werte aller untersuchten Unterrichtsmerkmale zunächst in einer Gesamtliste zusammengefasst und dann in Drittel eingeteilt. Das untere Drittel der ICC(1)-Werte lag zwischen .00 und .06, das mittlere Drittel zwischen .07 und .09 und das obere Drittel zwischen .09 bis .28. Wählt man die Mittelwerte dieser Intervalle jeweils als Referenzpunkt, dann spiegelt eine ICC(1) = .03 „kleine“, ICC(1) = .08 „mittlere“ und ICC(1) = .19 „große“ Unterschiede zwischen den Schulen hinsichtlich der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler wider. Mit Blick auf die gängige Konvention aus der Arbeit von LeBreton und Senter (2008) zeigt sich also, dass sich die empirischen Referenzwerte der vorliegenden Studie zur Interpretation „kleiner“ und „mittlerer“ Effekte gut decken, die gängige Interpretation „großer“ Effekte jedoch nur eingeschränkt auf die empirischen Verteilungen der ICC(1)-Werte der

Unterrichtsmerkmale zutrifft. Die empirischen Referenzwerte der ICC(1) der vorliegenden Arbeit können nun (ergänzend oder anstelle von etablierten Konventionen) als empirisch fundierte Anhaltspunkte für die Planung von großangelegten, experimentellen Interventionsstudien zur Verbesserung der Unterrichtsqualität von Schulen verwendet werden, die zunehmend an Bedeutung im Kontext einer evidenzbasierten Bildungspolitik gewinnen. Dafür ist es im Rahmen von Poweranalysen notwendig zu wissen, wie unterschiedlich Schulen hinsichtlich des jeweiligen Zielkonstrukts sind, um die Anzahl der Schulen (bzw. der Schülerinnen und Schüler innerhalb von Schulen) zu bestimmen (Hedges und Hedberg 2007).

Legt man an die empirisch ermittelten Reliabilitäten für die Schulmittelwerte zur Unterrichtsqualität den häufig verwendeten Maßstab von ICC(2) ≥ .70 an, dann zeigte sich, dass die Reliabilität der meisten Unterrichtsmerkmale in der Mehrzahl der Länder nicht in einem akzeptablen Bereich lag. Die ICC(2) ist eine Funktion der ICC(1) und der durchschnittlichen Gruppengröße. Ceteris paribus kann auf Grundlage der vorliegenden Studie geschlussfolgert werden, dass mehr als die bei PISA üblichen 35 Schülerinnen und Schüler einer Schule befragt werden müssen, wenn eine höhere Reliabilität der Schulmittelwerte erzielt werden soll. Entsprechend den Empfehlungen von LeBreton und Senter (2008) sollte dabei der Grenzwert für akzeptable Werte mit Blick auf den Anwendungskontext festgelegt werden. Insbesondere für bedeutsame Entscheidungen auf der Grundlage von Unterrichtsbeurteilungen durch Schülerinnen und Schüler (z. B. im Kontext der Schulinspektion) werden Reliabilitäten von ICC(2) ≥ .90, ähnlich zur Einzelfalldiagnostik, benötigt. Wird diese Marke angestrebt, müssten ausgehend von Gleichung 2 und von den in dieser Arbeit ermittelten Referenzwerten für kleine ICC(1)-Werte 291 Schülerinnen und Schüler pro Schule teilnehmen; bei mittleren bzw. großen ICC(1)-Werten sind es 104 bzw. 39 Schülerinnen und Schüler.

6.2 Grenzen der Studie

Die im Rahmen von PISA verfügbaren Daten sowie die verwendete Analysesoftware setzten der vorliegenden Studie Grenzen. (1) In dieser Studie untersuchten wir Unterrichtsmerkmale aus Sicht der Schülerinnen und Schüler auf Schulebene. Die Zuordnung der Schülerinnen und Schüler zu einzelnen Klassen ist mit den PISA-Daten nicht möglich. Es

ist davon auszugehen, dass die Schülerurteile des Unterrichts zwischen den Klassen innerhalb einer Schule variieren (Wurster und Gärtner 2013). Die vorliegenden Befunde sollten deshalb nicht auf Mehrebenen-Designs übertragen werden, in denen die Klassenebene im Mittelpunkt der Analysen steht. (2) Wir untersuchten zwar drei wichtige Basisdimensionen von Unterrichtsqualität, davon jedoch jeweils nur einen kleinen Ausschnitt relevanter Unterrichtsmerkmale auf Basis von Schülerurteilen. Hier ist weitere Forschung notwendig, die prüft, inwieweit die vorliegenden Befunde auf weitere Unterrichtsmerkmale bzw. auf Urteile von mehreren externen Beobachtern pro Schule (z. B. im Rahmen der Schulinspektion) generalisiert werden können. (3) Die vorliegenden Analysen beruhten auf ungewichteten Daten. Dies ermöglichte es uns, die Analysen mit Hilfe der R-Pakete lme4 und multilevel durchzuführen, die sich durch eine große Flexibilität bezüglich der statistischen Inferenz auszeichnen (z. B. Bootstrapverfahren für Konfidenzintervalle). Einzelne Vergleiche der vorliegenden Ergebnisse mit Analysen, bei denen die Gewichte berücksichtigt wurden, weisen aber darauf hin, dass die hier berichteten Verteilungen der ICC(1)-Werte nahezu identisch sind mit den entsprechenden Verteilungen, die bei gewichteten Analysen resultieren würden. (4) Da in dieser Studie die Gesamtverteilungen der Kennwerte der Übereinstimmung, Variabilität und Reliabilität über 81 Länder hinweg fokussiert werden, kann leider nicht dezidiert auf die Analyse einzelner Länder bzw. Ländergruppen und die Höhe ihrer Kennwerte eingegangen werden.

6.3 Schlussfolgerung

Generell zeigen die vorliegenden Ergebnisse, dass in der Mehrzahl der Länder für zahlreiche Unterrichtsmerkmale Schülerurteile innerhalb von Schulen hinreichend übereinstimmen, sodass die Aggregation auf Schulebene (im Sinne eines Konsensusmodells) gerechtfertigt ist. Die auf Schulebene gemittelten Schülerurteile zur Unterrichtsqualität unterscheiden sich systematisch und statistisch bedeutsam, wenn auch in unterschiedlichem Ausmaß zwischen den Schulen: Damit ist die Wahrnehmung des Unterrichts aus Sicht der Schülerinnen und Schüler eine wertvolle Informationsquelle, um die Unterrichtsqualität von Schulen differenziert analysieren zu können. Je nach Anwendungskontext sind die Urteile von

Unterrichtsmerkmalen durch etwa 35 Schülerinnen und Schüler pro Schule (der Standardstichprobengröße innerhalb von Schulen bei PISA) aber nur zu Teilen ausreichend, um befriedigende Reliabilitäten der Schulmittelwerte zu erzielen. Insbesondere Evaluationen in der pädagogischen Praxis sollten deutlich größere Stichprobenumfänge anstreben, um die Unterrichtsqualität durch Schülerurteile verlässlich einschätzen zu können.

Literatur

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf. Zugegriffen: 31. Juli 2016.

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1).

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten - institutionelle Bedingungen des Lehrens und Lernens. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–332). Opladen: Leske + Budrich.

Bliese, P. (2000). Within-group agreement, non-independence, and reliability. Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Hrsg.), *Frontiers of industrial and organizational psychology: Multilevel Theory, Research, and Methods in Organizations. Foundations, Extensions, and New Directions* (S. 349–381). San Francisco: Jossey-Bass.

Bliese, P. (2013). *multilevel: Multilevel functions. R package version 2.5*. https://CRAN.R-project.org/package=multilevel. Zugegriffen: 2. Juni 2016.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis. A typology of composition models. *Journal of Applied Psychology, 83*(2), 234–246.

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93*(1), 43–54.

Clausen, M. (2002). *Pädagogische Psychologie und Entwicklungspsychologie. Bd. 29: Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität*. Münster: Waxmann.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological methods, 6*(3), 297–310.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Hillsdale NJ: Erlbaum.

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein, & E. Terhart (Hrsg.), *Zeitschrift für Pädagogik. Beiheft. Bd. 41: Qualität und Qualitätssicherung im Bildungsbereich; Schule, Sozialpädagogik, Hochschule* (S. 73–92). Weinheim: Beltz.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität* (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*(1), 71–76.

Fischer, R. (2009). Where is culture in cross cultural research?: An outline of a multilevel research process for measuring culture as a shared meaning system. *International Journal of Cross Cultural Management, 9*(1), 25–49.

Gärtner, H. (2010). Wie Schülerinnen und Schüler ihre Lernumwelt wahrnehmen. *Zeitschrift für Pädagogische Psychologie, 24*(2), 111–122.

Gruehn, S. (2000). *Pädagogische Psychologie und Entwicklungspsychologie. Bd. 12: Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung*. Münster: Waxmann.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (1. Aufl.). Seelze-Velber: Kallmeyer.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78–79.

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), 85–98.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 333–359). Opladen: Leske + Budrich.

Kunter, M., Baumert, J., Blum, W., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852.

Lipowsky, F. (2009). Unterricht. In E. Wild & J. Möller (Hrsg.), *Springer-Lehrbuch: Pädagogische Psychologie* (S. 73–101). Berlin, Heidelberg: Springer.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131.

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.

OECD (2002). *PISA 2000 Technical Report*. Paris: OECD Publishing.

OECD (2005). *PISA 2003 Technical Report*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1. ed.). Oxford, New York: Pergamon.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520.

Viechtbauer, W. (2016). *metafor: Meta-Analysis Package for R. R package version 1.9-9*. https://cran.r-project.org/web/packages/metafor/metafor.pdf. Zugegriffen: 31. März 2016.

Wurster, S., & Gärtner, H. (2013). Erfassung von Bildungsprozessen im Rahmen von Schulinspektion und deren potenzieller Nutzen für die empirische Bildungsforschung. *Unterrichtswissenschaft, 41*(3), 217–236

4.2 Studie 2: To what extent are characteristics of a school's student body, instructional quality, school quality and school achievement interrelated?

Wenger, M., Gärtner, H. & Brunner, M. (2020). To what extent are characteristics of a school's student body, instructional quality, school quality and school achievement interrelated? *School Effectiveness and School Improvement*. Online Vorveröffentlichung.

Dieser Teil der Dissertation entspricht dem Manuskript, welches zur Publikation der Zeitschrift *School Effectiveness and School Improvement* angenommen wurde.

Die finale Publikation ist online vorveröffentlicht und verfügbar unter:

https://doi.org/10.1080/09243453.2020.1754243

4.3 Studie 3: Wie entwickeln sich Schulen nach der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“?

Wenger, M., Gärtner, H. & Brunner, M. (eingereicht). Wie entwickeln sich Schulen nach der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf”? *Zeitschrift für Erziehungswissenschaft.*

Dieser Teil der Dissertation entspricht dem Manuskript, welches zur Publikation bei der *Zeitschrift für Erziehungswissenschaft* eingereicht wurde.

Zusammenfassung

Die Schulinspektion evaluiert Schulen mit dem Ziel der Qualitätssicherung von Unterrichts- und Schulqualität. Dies gilt insbesondere für Schulen, an denen „erheblicher Entwicklungsbedarf“ festgestellt wurde. Diese Schulen bekommen zusätzliche Unterstützung, erfahren aber auch zusätzlichen Druck durch diese Einordnung. Die weitere Entwicklung dieser Schulen ist bisher kaum erforscht. Diese Studie untersucht mit Daten der Schulinspektion, der amtlichen Statistik und Leistungsdaten von 333 Berliner Grundschulen die Entwicklung von Schulen mit erheblichem Entwicklungsbedarf hinsichtlich von Unterrichts- und Schulqualität, Schulleistung, und Zusammensetzung der Schülerschaft (SES und Migrationshintergrund). Die empirischen Analysen zeigten, dass sich bei diesen Schulen die Unterrichts- und Schulqualität nur geringfügig verbesserte, sich der Leistungsabstand zu allen anderen Grundschulen nicht statistisch signifikant verringerte, und sich die Zusammensetzung der Schülerschaft hinsichtlich des sozioökonomischen Status (SES) nicht veränderte, sich jedoch der Anteil von Kindern mit Migrationshintergrund signifikant erhöhte.

Schlagwörter: Schulinspektion, Entwicklungsbedarf, Schülerleistungen, Entwicklung

Abstract

The School Inspectorate evaluates schools with the aim of quality assurance and quality development. This applies in particular to schools where "considerable need for development" has been identified. These schools receive additional support, but also experience additional pressure as a result. The further development of these schools has hardly been researched so far. Therefore, using data from the School Inspectorate, official statistics and performance data from 333 Berlin primary schools, this study examines the development of schools with significant development needs in terms of instruction and school quality, school performance, and composition of the student body (SES and migration background). The empirical analyses showed that at these schools the quality of teaching and schooling improved only slightly, the achievement gap did not decrease significantly in statistical terms (in comparison to all other primary schools), and the composition of the student body did not change with regard to socio-economic status (SES), but the proportion of children with migration experience increased significantly.

Keywords: development, school inspection, special measure, student performance

1. Einleitung – Funktionen und Wirkannahmen der Schulinspektion

In vielen Ländern der Bundesrepublik Deutschland wurde im Zuge der Reformmaßnahmen nach der ersten Veröffentlichung der PISA-Ergebnisse eine externe Evaluation von Schulen als neue Form der Qualitätssicherung eingeführt (im Folgenden einheitlich als Schulinspektion benannt, vgl. bspw. Döbert und Dedering 2008). Das übergeordnete Ziel der Schulinspektion ist die Verbesserung der Schul- und Unterrichtsqualität, insbesondere die Verbesserung der Leistungen von Schülerinnen und Schülern (Ehren und Visscher 2006). Generell beurteilt Schulinspektion schulische Qualität auf Grundlage normativer Festlegungen, die sich meist aus den landesspezifischen Qualitätsrahmen ergeben. Dabei sollen mehrere Funktionen erfüllt werden. Die Beurteilungen liefern Wissen (Funktion der Wissensgewinnung), das mehrere Folgeprozesse auslösen soll (Landwehr 2011). So sollen schulinterne Maßnahmen ausgelöst werden, die direkt zu einer Verbesserung der Schul- und Unterrichtsqualität beitragen (Schulentwicklungsfunktion). Schulen nutzen Schulinspektionsergebnisse in diesem Sinne eigenständig, um aufgezeigten Entwicklungsbedarf zu bearbeiten und ihre Stärken zu sichern. Gleichzeitig können sowohl die Einzelschule als auch die Bildungsverwaltung mithilfe der Schulinspektionsergebnisse Rechenschaft über die Qualität der Schule ablegen (Rechenschaftsfunktion; Landwehr 2011). Im internationalen Raum ist diese daher auch zentrales Ziel der Schulinspektion und schlechte Ergebnisse sind mit Sanktionen verbunden (z.B. in Großbritannien).

Auch die Rechenschaftsfunktion soll in der Folge zur Weiterentwicklung der Schul- und Unterrichtsqualität beitragen, beispielsweise durch von der Schulaufsicht angeordnete Maßnahmen für Schulen, bei denen Entwicklungsbedarf festgestellt wurde. Weiterhin unterstützt die Schulinspektion die Implementation politisch veränderter Normen und Anforderungen (Funktion der Normendurchsetzung) (Landwehr 2011), indem sie diese in den Katalog der Bewertungskriterien aufnimmt, an die Schulen zur Vorbereitung auf eine Schulinspektion kommuniziert und entsprechend evaluiert.

Die Wirkung einer Evaluation durch die Schulinspektion ist multidimensional und multidirektional (bspw. Reezigt und Creemers 2005; Ehren und Visscher 2006). Ein wesentlicher Faktor, der die Weiterentwicklung der Schulen motiviert, ist beispielsweise der

Druck, der durch die Veröffentlichung der Ergebnisse auf die Schulen entsteht (Reezigt und Creemers 2005; Altrichter und Kemethofer 2015). Dies kann gewünschte Folgen, wie eine verbesserte Schul- und Unterrichtsqualität und dadurch vermittelt eine Leistungssteigerung der Schülerinnen und Schüler nach sich ziehen (Ehren und Visscher 2006). Denkbar sind aber auch unerwünschte Folgen, wie beispielsweise die Stigmatisierung der Schulen und der Menschen, die dort arbeiten und der Kinder, die dort lernen.

Dies betrifft insbesondere solche Schulen, die von der Schulinspektion als besonders schwach bewertet werden, also Schulen, deren Leistungen die Qualitätskriterien der Schulinspektion nicht genügend erfüllen. In Bayern werden solche Schulen zum Beispiel als Schulen mit "großer Schwäche" bezeichnet, in Hamburg als „Fallkonferenzschulen" und in Berlin als „Schulen mit erheblichem Entwicklungsbedarf"; in Großbritannien sind es Schulen, die dem „special measure"-Programm zugeordnet werden. Im Folgenden verwenden wir die Bezeichnung „Schulen mit Entwicklungsbedarf". Mit einer solchen Diagnose sind verstärkte Erwartungen verknüpft, die schulischen Prozesse nach der Schulinspektion rasch zu verbessern. Bisher gibt es international nur wenige (bspw. Matthews und Sammons 2005) und für Deutschland keine Studien, die die Entwicklung von Schulen mit Entwicklungsbedarf untersuchen. Dieser Beitrag widmet sich daher folgender Forschungsfrage: Wie entwickeln sich Schulen nach der Diagnose als Schulen mit Entwicklungsbedarf? Für die empirischen Analysen verwenden wir hierzu einen bislang einzigartigen Datensatz, der Daten der Schulinspektion, der amtlichen Statistik und Leistungsdaten von allen (333) Berliner Grundschulen verknüpft.

2. Empirische Befunde

Was ist der bisherige Kenntnisstand zu den Folgen von Schulinspektion? Die Befunde werden hier zunächst generell (also für Schulen mit und ohne Entwicklungsbedarf) beschrieben, bevor auf die Besonderheit der Schulen mit Entwicklungsbedarf eingegangen wird.

2.1 Generelle Folgen der Schulinspektion für Schulen

Die bisherige Forschung zu generellen Folgen der Schulinspektion fokussierte auf drei zentrale Aspekte: (1) die Entwicklung der Unterrichts- und Schulqualität, (2) die Leistungsentwicklung der Schülerinnen und Schüler und (3) nichtintendierte Nebeneffekte.

(1) Einige Untersuchungen zeigten, dass ein größerer Anteil von Lehrkräften in Folge der Schulinspektion etwas am Lehrstil und den Unterrichtsmethoden ändern wollte (international bspw. Brimblecombe et al. 1996: 33%). Es konnte auch festgestellt werden, dass relativ einfach umzusetzende Verbesserungen initiiert wurden, beispielsweise die Erhöhung der Zeit für den Sprachunterricht (Ehren und Visscher 2008). Gärtner et al. (2009) beschreiben, dass durchschnittlich vier konkrete qualitätssichernde oder –verbessernde Maßnahmen in Folge auf die Rückmeldung der Schulinspektion eingeleitet wurden und 45 % der Schulleitungen einen Einfluss auf den Unterricht wahrnahmen. Auch bei Preuße et al. (2019) fand ein Großteil (65 %) der Schulleitungen, dass die Schulinspektion Prozesse der Schulentwicklung anregte. Auf der anderen Seite schätzten in dieser Studie 61 % der Schulleitungen den Nutzen der Schulinspektion für die Qualitätssicherung pädagogischer Prozesse als gering ein und nur 36 % sehen eine positive Langzeitwirkung. Einige Studien zeigen auch, dass - so wie theoretisch angenommen wird (Reezigt und Creemers 2005; Ehren und Visscher 2006) - externer Druck Schulen zur Qualitätsentwicklung anregte: Schulleitungen, die höheren externen Druck verspürten berichteten über vermehrte Entwicklungsaktivitäten an ihren Schulen (Altrichter und Kemethofer 2015). Weiterhin führte die Androhung von Sanktionen dazu, dass mehr Geld für Fortbildungen und für die Unterrichtsentwicklung in leistungsschwachen Schulen eingesetzt wurden (Chiang 2009). Es wurden zudem positive Effekte durch die Veröffentlichung der Ergebnisse beschrieben und dadurch rückgeschlossen, dass externer Druck die Entwicklungstätigkeit erhöhte (Faubert 2009). Gärtner et al. (2014) fanden in einer Längsschnittstudie wiederum kaum relevante Folgen durch die Schulinspektion in verschiedenen Schulqualitätsindikatoren.

(2) Zur Leistungsentwicklung von Schülerinnen und Schülern in Folge von Schulinspektion wurden keine eindeutigen Ergebnisse gefunden. Einerseits wurde festgestellt, dass keine positive Entwicklung stattgefunden hat (Cullingford 1999; Shaw et al. 2003; Gärtner

und Pant 2011) und die Leistungen im Jahr der Schulinspektion sogar schwächer ausfielen (Rosenthal 2004). Andererseits konnten in anderen Studien auch positive Effekte gefunden werden (Hanushek und Raymond 2005; Lee 2006; Pietsch et al. 2014). Zudem fanden manche Studien auch differenzielle Effekte je nach Schulkontext, Bedingungen der Evaluation, Ressourcen, Untersuchungsmethode oder auch Fach (Shaw et al. 2003; Hanushek und Raymond 2005; Lee 2006; Luginbuhl et al. 2009; Pietsch et al. 2014; Penninckx et al. 2016).

(3) Zu den nichtintendierten Effekten, die als eine Folge von Schulinspektion berichtet wurden, gehörten beispielsweise das „window dressing“ (bspw. durch Ausschließen bestimmter Schülerinnen und Schülern von Tests, „Aufhübschen“ der Dokumente, Ehren und Swanborn 2012; Kemethofer und Helm 2017) sowie erhöhter Stress der Schulleitung bzw. des Kollegiums (Penninckx et al. 2016; Preuße et al. 2019). Weiterhin sind aufgrund der Beanspruchung durch die Schulinspektion andere Aufgaben liegen geblieben und das Arbeitsklima wurde als angespannt beschrieben. Auch die zeitliche Belastung wurde als hoch eingeschätzt (Preuße et al. 2019). Böhm-Kasper et al. (2016) benennen zudem eine „Stigmatisierung“ von Schulen, die auch das Vertrauen in die Schule schwächen kann.

Insgesamt ist die Befundlage zu generellen Folgen von Schulinspektion nicht eindeutig: Die Schulinspektion scheint insgesamt nicht zwangsläufig zu einer Verbesserung von Schul- und Unterrichtsqualität an den inspizierten Schulen sowie zu einer Leistungssteigerung der Schülerinnen und Schülern zu führen. Dabei stellt sich jedoch die Frage, inwiefern dies auch bei den Schulen mit Entwicklungsbedarf zutrifft.

2.2 Folgen der Schulinspektion für „Schulen mit Entwicklungsbedarf“

Die Schulinspektion ist durch das Erfüllen ihrer Rechenschaftsfunktion vermutlich die einzige Institution, die auf einer breiten Kriterienbasis Schulen mit besonders schwacher Schul- und Unterrichtsqualität identifizieren kann. Dabei nimmt das Feststellen eines „Entwicklungsbedarfs“ eine besondere Stellung ein, da eine solche Diagnose nur erfolgt, wenn zahlreiche Indikatoren als schwach ausgeprägt bewertet werden. Die genauen Kriterien, wann Entwicklungsbedarf diagnostiziert wird, unterscheiden sich zwischen den Systemen der Bundesländer (und international). Davon unabhängig ist diese Diagnose mit bestimmten

Konsequenzen und Zielen verknüpft. Dazu gehört meist, eine weiterführende Kontrolle durch die Schulinspektion, zum Beispiel eine zeitlich vorgezogene Nachinspektion der Schul- und Unterrichtsqualität an diesen Schulen. Dazu zählen auch verstärkte Unterstützungsangebote an die Schulen, wie beispielsweise durch Schulberaterinnen und Schulberater oder die Schulaufsicht. Betrachtet man den Forschungsstand speziell für Schulen mit Entwicklungsbedarf findet man nur wenige Studien, die bislang (1) die Entwicklung der Unterrichts- und Schulqualität, (2) die Leistungsentwicklung der Schülerinnen und Schüler oder (3) nichtintendierte Nebeneffekte für diese besondere Gruppe von Schulen untersuchten.

Ziel der Diagnose Schule mit Entwicklungsbedarf ist es, die Schulen herauszuheben, die besondere Schwächen aufweisen, um durch geeignete Interventionen die Unterrichts- und Schulqualität sowie die Schülerleistungen rasch zu verbessern. Es ist bei diesen Schulen von „verschärften Bedingungen“ (Dedering et al. 2016, S. 204) auszugehen. Durch die Außenwirkung des Urteils ist nach der Theorie des „accountability pressure“ anzunehmen, dass es zu einem verstärkten Druck und damit zu entsprechenden positiven Entwicklungen kommt (Kotthoff und Böttcher 2009; Perryman 2010; Altrichter und Kemethofer 2015). Einige wenige empirische Untersuchungen deuten darauf hin, dass sich diese Wirkung einstellt. (1) So konstatiert die Berliner Schulinspektion, dass in den Nachinspektionen deutliche Verbesserungen der Schulen mit Entwicklungsbedarf hinsichtlich ihrer Unterrichts- und Schulqualität zu beobachten sind (Senatsverwaltung für Bildung, Jugend und Wissenschaft 2014). Dies scheint auch für Großbritannien zu gelten: So beschreiben zum Beispiel Matthews und Sammons (2005) eine Abnahme von schlechten Unterrichtsstunden als Folge der Einordnung in das special measure-Programm. (2) Weiterhin zeigte sich für niederländische Schulen die aufgrund ihres „hohen Risikos“ in einem risikobasierten Schulinspektionssystem häufiger inspiziert wurden, dass sich die Leistungen von Schülerinnen und Schülern an diesen Schulen stärker verbesserten als an anderen Schulen (Ehren und Shackleton 2016).

(3) Auf der anderen Seite können sich auch nichtintendierte Nebeneffekte der Schulinspektion verstärkt zeigen (Altrichter und Kemethofer 2015). Einerseits, da die Ergebnisse der Schulinspektion nicht ignoriert werden können, andererseits, da die schulinternen Voraussetzungen für eine sinnvolle Annahme der Ergebnisse und eine darauf

basierende Weiterarbeit schwieriger sind als an anderen Schulen (Dedering et al. 2016). So zeigte sich, dass sich die pädagogischen Akteure an solchen Schulen häufig ungerecht behandelt fühlten. Der Fokus ihrer Arbeit lag dann häufig nicht auf einer Weiterentwicklung der Unterrichts- oder Schulqualität, sondern auf der Bearbeitung von Dokumenten (Sommer 2010; Dedering et al. 2016). Dies lag vermutlich auch daran, dass an diesen Schulen nach Dedering et al. (2016) häufig das strategische Wissen bzw. die Strukturen fehlten (wie auch durch die Schulinspektion bescheinigt wurde) die erwarteten Veränderungen herbeizuführen. Dabei fühlten sich die pädagogischen Akteure an Schulen mit Entwicklungsbedarf auch häufig von der Schulaufsicht nicht in ausreichendem Maß unterstützt (Preuße et al. 2019).

3. Forschungsfragen

Die Diagnose Schule mit Entwicklungsbedarf geht mit gesteigerten Erwartungen an die Verbesserung von Schul- und Unterrichtqualität einher (Perryman 2010). Der Forschungsstand zu Wirkungen der Schulinspektion im Allgemeinen zeigt, dass in Folge der Schulinspektion einige geringe Verbesserungen der Unterrichts- und Schulqualität zu beobachten sind. Die Ergebnisse zur Leistungsentwicklung von Schülerinnen und Schülern sind nicht eindeutig. Schulinspektion kann wiederum auch einige nichtintendierte Nebeneffekte herbeiführen. Die internationalen Forschungsergebnisse speziell zu Schulen mit Entwicklungsbedarf zeigen tendenziell positivere Wirkungen (Matthews und Sammons 2005; Ehren und Shackleton 2016). Über Schulen mit Entwicklungsbedarf in Deutschland ist jedoch so gut wie nichts bekannt. Auf Basis der Daten Berliner Grundschulen soll mit dieser Studie daher erstmals die Entwicklung von Schulen mit Entwicklungsbedarf in Folge der Schulinspektion untersucht werden. Dabei stehen die folgenden Forschungsfragen im Fokus: Wie entwickeln sich Schulen mit Entwicklungsbedarf hinsichtlich (a) ihrer Unterrichts- und Schulqualität, (b) ihren Schulleistungen und (c) der Zusammensetzung ihrer Schülerschaft?

Die Diagnose Schule mit Entwicklungsbedarf ist in Berlin mit einem Bündel an Maßnahmen verbunden, die zahlreiche Wirkmechanismen auslösen können. Es ist zu erwarten, dass die Schulen durch den Druck, der durch die Veröffentlichung der Ergebnisse entsteht, großes Interesse haben, bei einer Nachinspektion Verbesserungen zu zeigen. Zudem kommt es

zu einer verstärkten Kontrolle durch die Schulaufsicht, die fordert, Qualitätsdefizite abzubauen. Hierzu können Schulen mit Entwicklungsbedarf zusätzliche Unterstützung in Form von unterschiedlichen Beratungsangeboten in Anspruch nehmen. In der Summe können diese Faktoren zu einer Verbesserung der Schul- und Unterrichtsqualität führen und darüber vermittelt zu besseren Schülerleistungen. Andererseits ist es gut möglich, dass durch die Diagnose einer Grundschule als Schule mit Entwicklungsbedarf, Eltern versuchen ihre Kinder (trotz der in Berlin geltenden Einzugsgebietsregelung) auf andere Schulen zu geben (Noreisch 2007; Fincke und Lange 2012). Dabei ist ausgehend von der Forschung zum Elternverhalten bei der Schulwahl für ihre Kinder anzunehmen, dass Eltern verstärkt versuchen, Schulen mit Entwicklungsbedarf zu meiden. Dies wird angenommen, da Eltern - zumindest bei der Wahl der weiterführenden Schule - öffentliche Schulinspektionsergebnisse und Kennzahlen zur Schülerschaft bei der Schulwahl berücksichtigen (Clausen 2006; Böhm-Kasper et al. 2016; Jurczok 2019). Zudem wird für die Suche nach der passenden weiterführenden Schule auch von der Senatsverwaltung auf die öffentlichen Schulinspektionsergebnisse im Schulverzeichnis hingewiesen (Senatsverwaltung für Bildung, Jugend und Familie 2017). Es kann angenommen werden, dass Eltern solche Quellen auch für die „Wahl" der Grundschule nutzen. Berücksichtigt man Ergebnisse zu Unterschieden der Schulwahl abhängig vom sozialen Hintergrund und der Migrationserfahrung der Familien (Riedel et al. 2010; Jurczok 2019), lässt sich vermuten, dass sich insbesondere Eltern aus bestimmten Gruppen gegen Schulen mit Entwicklungsbedarf entscheiden: Eltern, die sich *nicht* in sozialer, finanzieller oder bildungsbezogener Risikolage befinden. So wäre zu erwarten, dass sich durch die Diagnose Schule mit Entwicklungsbedarf die Zusammensetzung der Schülerschaft an diesen Schulen verändert und sich der Anteil von Kindern aus sozioökonomisch benachteiligten Familien oder Familien mit Migrationshintergrund erhöht.

4. Methode

4.1 Stichprobe Berliner Grundschulen

Tabelle 1 Schulinspektionen an Berliner Grundschulen 2011-2017

Diagnose der Schulinspektion	Schuljahr						Summe
	11/12	12/13	13/14	14/15	15/16	16/17	
Schulen ohne erheblichen Entwicklungsbedarf	55	38	51	63	43	61	311
Schulen mit erheblichem Entwicklungsbedarf	1	5	6	4	2	4	22
Summe Schulinspektionen	56	43	57	67	45	65	333

Tabelle 2 Deskriptive Merkmale der Berliner Grundschulen (N = 333)

	N	*M*	*SD*	Min	Max
Schulleistung[a] im Jahr der SI	333	0.06	0.97	-2.36	2.23
Schulleistung[a] ein Jahr nach SI	268	0.13	0.97	-2.56	2.57
Schulleistung[a] zwei Jahre nach SI	223	0.07	0.99	-2.37	2.08
Schulleistung[a] drei Jahre nach SI	155	0.08	1.01	-2.45	1.98
Migrationshintergrund (%) im Jahr der SI	333	40.53	28.09	0.31	98.59
Migrationshintergrund (%) ein Jahr nach SI	268	40.87	27.46	0.00	95.70
Migrationshintergrund (%) zwei Jahre nach SI	223	41.67	27.78	0.27	96.26
Migrationshintergrund (%) drei Jahre nach SI	156	39.98	27.25	0.61	96.11
SES (%) im Jahr der SI	333	37.76	25.29	0.26	98.79
SES (%) ein Jahr nach SI	268	36.61	24.52	2.19	91.98
SES (%) zwei Jahre nach SI	223	37.93	24.85	1.32	91.95
SES (%) drei Jahre nach SI	156	35.88	24.67	0.27	96.26

Anmerkungen. Min = Minimum. Max = Maximum. SES (%): 100% – schul-spezifischer Anteil der Schülerinnen und Schüler, die lernmittelzuzahlungsbefreit sind (d.h. ein SES-Wert von 0% steht für ein sehr niedriges Niveau des SES der Schülerschaft dieser Schule). SI = Schulinspektion.

[a] z-standardisierte mittlere Schulleistung

Die Schulinspektion in Berlin begutachtet seit 2005 ca. alle fünf Jahre alle öffentlichen Berliner Schulen. In der zweiten sogenannten Runde der Berliner Schulinspektion wurden

zwischen 43 und 67 Grundschulen pro Jahr inspiziert (Tab. 1), die Auswahl pro Jahr erfolgt zufällig.

Die Analysen in dieser Studie basieren auf den Daten aller öffentlichen Berliner Grundschulen ($N = 333$), die im Zeitraum zwischen 2011 und 2017 im Rahmen der zweiten Runde der Schulinspektion evaluiert wurden. Die deskriptiven Merkmale dieser Grundschulen sind in Tab. 2 dargestellt[7]. Von diesen 333 Grundschulen wurden innerhalb der zweiten Runde an Schulinspektionen 22 öffentliche Grundschulen als Schulen mit Entwicklungsbedarf eingeordnet (6,6 % aller Grundschulen). An 10 der 22 Grundschulen fanden im Zeitraum bis 2017 die Nachinspektionen statt (2 bis 3 Jahre nach der Regelinspektion im Rahmen der zweiten Runde). Eine von diesen nachinspizierten Schulen wies dabei wiederum Entwicklungsbedarf auf. Die anderen 12 Grundschulen werden im Rahmen der dritten Runde der Schulinspektion erneut inspiziert werden.

4.2 Verfahrensweise der Berliner Schulinspektion

Im Vorfeld der Schulinspektion werden von jeder Schule Dokumente angefordert (u.a. Gremienprotokolle und das schulinterne Curriculum), anschließend besucht ein Schulinspektionsteam aus mehreren Personen mehrere Tage die Schule, um sie zu evaluieren. Im Anschluss daran wird jeder Schule Rückmeldung zu den Evaluationsergebnissen in Form eines Berichts gegeben. Eine Kurzversion des Berichts wird im Internet im Rahmen des Schulprofils veröffentlicht. Der Bericht enthält eine Zusammenfassung der Stärken und Schwächen sowie das Qualitätsprofil mit den Bewertungen zu den einzelnen Indikatoren der Schulinspektion (Senatsverwaltung für Bildung, Wissenschaft und Forschung 2009). Wenn eine Schule Entwicklungsbedarf aufweist, der „weit über dem Durchschnitt liegt“ (Senatsverwaltung für Bildung, Wissenschaft und Forschung 2009), dann wird sie als „Schule mit erheblichem Entwicklungsbedarf“ eingestuft. Die Diagnose erfolgt in Berlin wenn Schulen eines oder mehrere der folgenden Kriterien aufweisen (Stand 2019, in Bezug auf Senatsverwaltung für Bildung, Wissenschaft und Forschung 2009): (1) Schlechte Ergebnisse in

[7] Im Online Supplement sind die deskriptiven Merkmale zudem separat für Schulen mit und ohne Entwicklungsbedarf dargestellt.

Schulleistungsuntersuchungen bleiben ohne schulische Konsequenzen, (2) Mängel im Schul- bzw. Konflikt- und Beschwerdemanagement, (3) für besondere schulspezifische Problemlagen werden keine geeigneten Gegenmaßnahmen ergriffen, (4) gravierende Mängel im Personal- und Ressourcenmanagement bestehen, (5) eine durchwegs als schwach bewertete Unterrichtqualität liegt vor.

Hauptkonsequenz dieser Diagnose ist, dass ca. zwei Jahre später eine vorzeitige „Nachinspektion" der Schule stattfindet, um zu prüfen, ob sich die Schul- und Unterrichtsqualität bedeutsam verbessert hat. Andererseits wurde die prozessbegleitende Schulberatung „proSchul" eingerichtet, um insbesondere Schulen mit Entwicklungsbedarf zu unterstützen. Zudem sollen Absprachen mit der Schulaufsicht getroffen werden, um zu dokumentieren, wie der festgestellte Entwicklungsbedarf überwunden werden kann (Senatsverwaltung für Bildung, Jugend und Wissenschaft 2012).

4.3 Instrumente

Die Analysen Berliner Grundschulen in dieser Studie basieren auf der Kombination der Daten der (a) Schulinspektion, (b) amtlichen Statistik sowie (c) Vergleichsarbeiten der dritten Jahrgangsstufe (VERA 3).

4.3.1 Unterrichts- und Schulqualität

Die Schulinspektion erfasst 19 Merkmale zu Unterrichts- und Schulqualität für alle Berliner Schulen. Diese Merkmale wurden von der Berliner Schulinspektion als Evaluationskriterien entwickelt und sind aus dem „Handlungsrahmen Schulqualität in Berlin" (Senatsverwaltung für Bildung, Jugend und Wissenschaft 2013) abgeleitet. Ein Schulinspektionsteam besteht aus vier geschulten Personen. Bei der Schulinspektion werden zur Bewertung der Unterrichtsqualität mindestens 70% aller Lehrkräfte jeweils 20 Minuten in ihrem Unterricht besucht. Hieraus wird das „Unterrichtsprofil der Schule" erstellt, Einzelbewertungen auf Klassenebene werden nicht rückgemeldet. Die Bewertung der Schulqualitätsindikatoren basiert auf verschiedenen Datengrundlagen: Es erfolgen Interviews mit der Schulleitung, mit Schülerinnen und Schülern, mit Lehrkräften und mit Eltern, es werden Schuldokumente analysiert und schriftliche

Fragebögen mit verschiedenen Personengruppen eingesetzt. Die Qualitätsindikatoren der Unterrichts- und Schulqualität werden jeweils auf einer 4-stufigen Bewertungsskala von 1= „schwach ausgeprägt“, 2= „eher schwach ausgeprägt“, 3 = „eher stark ausgeprägt“ und 4 = „stark ausgeprägt“ eingeschätzt. Die Ergebnisse von Reliabilitätsanalysen der Daten aus der Brandenburger Schulinspektion (Visitation) mit einem sehr ähnlichen Verfahren, legen nahe, dass die Merkmale zur Unterrichts- und Schulqualität reliabel auf Schulebene erfasst werden können (Wurster und Gärtner 2013).

Die von der Schulinspektion evaluierten Merkmale wurden im Rahmen dieser Studie einschlägigen Dimensionen der Unterrichts- und Schulqualität zugeordnet (Tab. 3). In der Forschung zur Unterrichtsqualität besteht breiter Konsens drei Basisdimensionen zu differenzieren: Effiziente Klassenführung, kognitive Aktivierung und konstruktive Unterstützung (Klieme et al. 2001; Kunter et al. 2011). In der Forschung zu Schulqualität werden häufig die folgenden Faktoren unterschieden (Scheerens und Bosker 1997; Ditton 2000): Schulkultur, Schulmanagement, Kooperation/Koordination, Personal.

Tabelle 3 Unterrichts- und Schulqualität: Indikatoren der Schulinspektion

	M	*SD*	α	Qualitätsindikatoren der Schulinspektion	Beispielitem
				Unterrichtsqualität	
Effiziente Klassenführung	3.78	0.13	0.64	Lehr-Lernzeit	Der Anteil an Warte- und Leerlaufzeiten für die Schüler/innen ist gering.
				Verhalten der SuS im Unterricht	Schülerinnen und Schüler stören nicht den Unterricht.
Kognitive Aktivierung	2.51	0.24	0.87	Methoden- und Medienwahl	Die Lehrkraft gestaltet den Unterricht anregend und motivierend.
				Selbstständiges Lernen	Die Schüler/innen organisieren Lernprozesse/Unterrichts- bzw. Arbeitsabläufe selbstständig.
				Problemorientiertes Lernen	Im Unterricht werden ergebnisoffene Frage- und Problemstellungen behandelt.
				Leistungs- und Anstrengungsbereitschaft	Die Leistungsanforderungen sind herausfordernd.
				Kooperatives Lernen	Teamorientierte Aufgabenstellungen werden im Unterricht gestellt und behandelt.
Konstruktive Unterstützung	3.10	0.18	0.77	Pädagogisches Klima im Unterricht	Die Lehrkraft sorgt für eine angstfreie Lernatmosphäre.
				Förderung von Selbstvertrauen und Selbsteinschätzung	Das Selbstvertrauen der Schüler/innen wird gefördert (z.B. Anerkennung, Ermunterung, Lob).
				Innere Differenzierung	Für die Schüler/innen bestehen Wahlmöglichkeiten entsprechend ihren Interessen und Neigungen.
				Strukturierung und transparente Zielausrichtung	Die Unterrichtsschritte sind nachvollziehbar und klar strukturiert.

Schulqualität					
Schulkultur	2.77	1.01	-	Fortschreibung des Schulprogramms	In der Schule wird nachvollziehbar an der Umsetzung der Schwerpunkte des Schulprogramms gearbeitet.
Schulmanagement	3.23	0.67	0.69	Schulleitungshandeln und Schulgemeinschaft	Die Schulleiterin/der Schulleiter wird von den Lehrkräften in ihrem bzw. seinem Führungsverhalten anerkannt.
				Schulleitungshandeln und Qualitätsmanagement	Die Schulleiterin/der Schulleiter sorgt für die Entwicklung einer schulspezifischen Steuerungsstruktur zur Qualitätsentwicklung und -sicherung.
				Evaluation schulischer Entwicklungsvorhaben	Die Schule wählt aus ihren Entwicklungsvorhaben Schwerpunkte zur internen Evaluation aus.
Kooperation/ Koordination	3.31	0.52	0.52	Systematische Unterstützung, Förderung und Beratung	Die Schule stimmt sich im Hinblick auf Fördermaßnahmen mit internen und/oder externen Fachleuten ab.
				Beteiligung der SuS und der Eltern	Die Schüler/innen beteiligen sich aktiv an der Schulentwicklung.
				Arbeits- u. Kommunikationskultur im Kollegium	Es gibt klare Teamstrukturen, in die eine bedeutsame Anzahl des Kollegiums eingebunden ist.
Personalentwicklung	2.84	0.85	-	Personalentwicklung und Personaleinsatz	Strategien zur Professionalisierung von Mitarbeiter/innen werden in konkreten Maßnahmen umgesetzt.

Anmerkungen. Die Qualitätsindikatoren der Unterrichts- und Schulqualität werden auf einer 4-stufigen Bewertungsskala von 1=„schwach ausgeprägt“, 2=„eher schwach ausgeprägt“, 3=„eher stark ausgeprägt“, 4=„stark ausgeprägt“ durch die Schulinspektion eingeschätzt

4.3.2 Leistungen der Schülerinnen und Schüler

Als Indikator der Schulleistung einer Grundschule wurden die Ergebnisse der Schülerinnen und Schüler bei den Vergleichsarbeiten der dritten Jahrgangsstufe (VERA 3) genutzt. Die Vergleichsarbeiten basieren auf den deutschen Bildungsstandards, die Kompetenzen in verschiedenen Domänen (zum Beispiel Deutsch und Mathematik) überprüfen. Die Teilnahme an den Vergleichsarbeiten war für alle öffentlichen Berliner Grundschulen verpflichtend. Um einen Indikator der Schulleistung zu ermitteln, wurde zunächst der Mittelwert der Testergebnisse in Mathematik und Deutsch für jeden Schüler/jede Schülerin berechnet (Korrelationen der beiden Testwerte lagen auf Schülerebene zwischen $r = .85$ und $r = .91$). Die gemittelten individuellen Leistungswerte wurden über alle Schülerinnen und Schüler an einer Schule aggregiert, um für jede Schule einen Schulleistungswert zu erhalten. Schließlich wurden die aggregierten Schulleistungswerte für jedes Schuljahr jeweils z-standardisiert mit Mittelwert $M = 0$ und $SD = 1$. Da die Grundschulen in einem bestimmten Jahr zufällig von der Schulinspektion ausgewählt werden und nicht alle Grundschulen im gleichen Jahr inspiziert werden, erlaubt die z-Standardisierung einen Vergleich der Schulleistungen über die Jahre hinweg. Ein Schulleistungswert von 0 bedeutet beispielsweise, dass das mittlere Leistungsniveau der Schülerinnen und Schüler an einer bestimmten Grundschule im Vergleich zu anderen Grundschulen in Berlin durchschnittlich ausgefallen ist. Ein Schulleistungswert von 1 bedeutet, dass die mittlere Schulleistung dieser Schule eine Standardabweichung über dem Durchschnitt aller anderen Berliner Grundschulen lag.

4.3.3 Zusammensetzung der Schülerschaft

Als Indikatoren der Zusammensetzung der Schülerschaft einer Schule dienen zwei Berlinspezifische Merkmale: die nicht-deutsche Herkunftssprache (ndH) und die Lernmittelzuzahlungsbefreiung (lmb). Die nicht-deutsche Herkunftssprache dient als Indikator des Migrationshintergrunds, die aus der amtlichen Statistik stammt und von 0% -100% reicht. 0% bedeutet, dass kein Schüler und keine Schülerin an der Schule einen Migrationshintergrund hat; 100% bedeutet, dass alle Schülerinnen und Schüler der Schule einen Migrationshintergrund haben. Die Lernmittelzuzahlungsbefreiung dient als Indikator für den sozioökonomischen

Hintergrund (SES) der Schülerinnen und Schüler. Dieses Merkmal entstammt ebenso der amtlichen Statistik und erfragt, ob die Familie von der Zuzahlung der Lernmittel für die Schule befreit wurde. Eltern müssen dafür nachweisen, dass sie finanzielle staatliche Unterstützung bekommen. Damit hohe Werte einen hohen SES abbilden, wurde ein Indikator erstellt, der den sozioökonomischen Hintergrund der Schule wie folgt darstellt: 100% minus den schulspezifischen prozentualen Anteil der Schülerinnen und Schüler, die von der Lernmittelzuzahlung befreit wurden. Daher impliziert ein Wert des sozioökomischen Hintergrunds von 0%, dass alle Schülerinnen und Schüler dieser Schule von der Lernmittelzuzahlung befreit wurden, während ein Wert von 100% bedeutet, dass an der Schule keine Schülerinnen und Schüler von der Lernmittelzuzahlung befreit wurde. In anderen Studien zeigte sich dieser Berlinspezifische Indikator als hoch korreliert mit anderen etablierten Indikatoren für SES (bspw. der Highest International Socio-Economic Index of Occupational Status [HISEI]; Maaz et al. 2016).

4.4 Vorgehen/Analysen

Die Forschungsfrage (a) zur Weiterentwicklung der Unterrichts- und Schulqualität der Schulen mit Entwicklungsbedarf erfolgte durch einen Vergleich der Werte der Regelinspektion von den 10 Schulen, die im Rahmen der zweiten Runde der Schulinspektion zweimal evaluiert wurden. Dafür werden standardisierte Mittelwertsdifferenzen (Cohens *d*) für alle Unterrichts- und Schulqualitätsindikatoren wie folgt berechnet: Mittelwertsdifferenz ($M_{Nachinspektion} - M_{Regelinspektion}$) geteilt durch Standardabweichung des Indikators bezogen auf die Schulinspektionsdaten aller Berliner Grundschulen.

Die Beantwortung der Forschungsfragen zur Entwicklung (b) der Schulleistungen und (c) der Schülerzusammensetzung in Folge der Diagnose Schule mit Entwicklungsbedarf basierte auf den Daten von allen 333 Berliner Grundschulen. Für die Analysen wurden mittels lavaan (Rosseel 2012) drei Strukturgleichungsmodelle zur Vorhersage von (1) der Schulleistung, (2) SES der Schülerschaft und (3) dem schulspezifischen Anteil der Kinder mit Migrationshintergrund spezifiziert (s. Abb. 2-4). In allen drei Modellen wurden diese Variablen ein Jahr nach der Schulinspektion, zwei Jahre nach der Schulinspektion, und drei Jahre nach

der Schulinspektion vorhergesagt. Als Prädiktor diente eine Indikatorvariable, die angab, ob eine Schule mit Entwicklungsbedarf eingeordnet wurde (kodiert als 1) oder nicht (kodiert als 0). Zusätzlich wurden als weitere Kovariaten die im Jahr der Schulinspektion ermittelte Schulleistung, SES und Migrationshintergrund im Modell kontrolliert. Der Anteil fehlender Werte lag zwischen 0 % und 53 %, durchschnittlich bei 12.6 %[8] (siehe Fallzahlen Tab. 2). Zur Schätzung aller Modellparameter wurde daher das „Full Information Maximum Likelihood"-Verfahren in lavaan genutzt (Enders 2010).

5. Ergebnisse

Ein Vergleich der Mittelwertsdifferenzen wies auf differenzielle Veränderungsprozesse der Schul- und Unterrichtsqualität (Forschungsfrage a) hin (Tab. 4). Die stärksten positiven Veränderungen zeigten sich für die kognitive Aktivierung (d=0.5) und konstruktive Unterstützung (d=0.67). Nach Cohen (1988) sind diese Veränderungen als mittlere Effekte einzuordnen. Im Bereich der Klassenführung war keine bedeutsame Veränderung festzustellen. Mit Blick auf die Schulqualität verbesserten sich die Kooperation/Koordination (d=0.26) und der Bereich Personal (d=0.36) leicht (kleiner Effekt). Bei den anderen beiden Qualitätsdimensionen (Schulkultur und –management) zeigten sich keine Veränderungen.

[8] Grund dafür ist, dass die Anzahl der verfügbaren Daten der Schulen mit jedem Jahr nach der Schulinspektion geringer werden, da diese dann für einen Teil der Schulen über das Ende der zweiten Welle der Schulinspektion hinausreichen und daher nicht zur Verfügung standen.

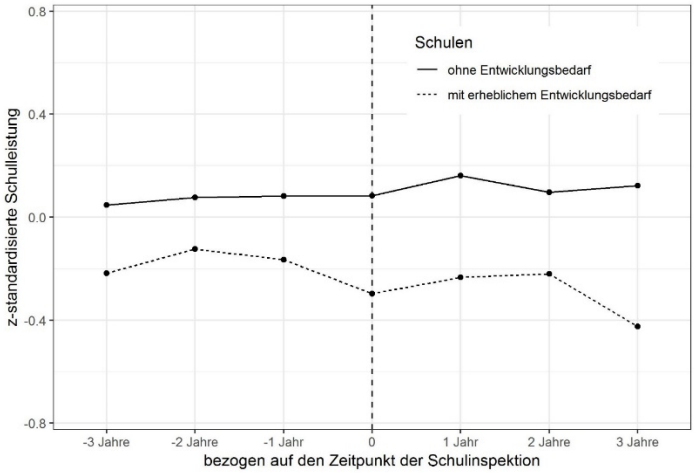
Tabelle 4 Forschungsfrage a: Veränderung der Evaluationsergebnisse bei nachinspizierten Schulen (N=10)

	Regel-inspektion *M (SD)*	Nach-inspektion *M (SD)*	Mittelwerts-differenz (KI)[9]	Standardisierte Mittelwerts-differenz (Cohens *d*)
		Unterrichtsqualität		
Klassenführung	3.71 (0.15)	3.72 (0.17)	0.01 (-0.13; 0,14)	0.05
Kognitive Aktivierung	2.28 (0.26)	2.40 (0.29)	0.12 (-0.21; 0.45)	0.50
Konstruktive Unterstützung	2.89 (0.21)	3.01 (0.30)	0.12 (-0.19; 0.43)	0.67
		Schulqualität		
Schulkultur	2.20 (1.14)	2.30 (1.16)	0.10 (-1.09; 1.29)	0.10
Schulmanagement	2.42 (0.93)	2.47 (1.03)	0.05 (-1.03; 1.13)	0.07
Kooperation/ Koordination	2.60 (0.54)	2.73 (0.58)	0.13 (-0.50; 0.76)	0.26
Personal	1.90 (0.99)	2.20 (1.03)	0.30 (-0.71; 1.31)	0.36

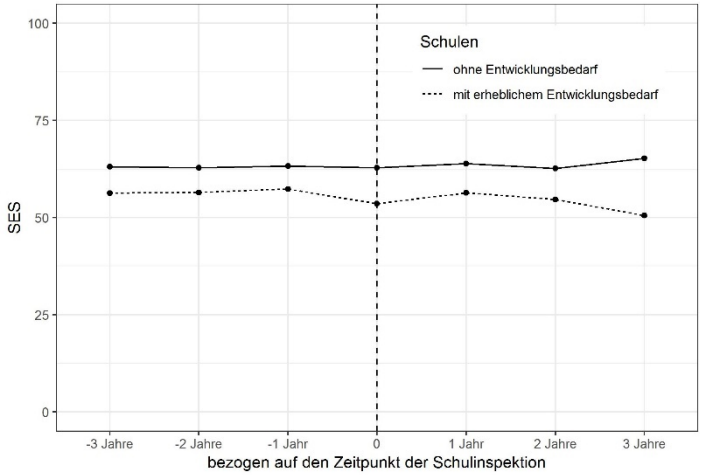
Anmerkungen. Standardisierte Mittelwertsdifferenz Cohens *d*: Mittelwertsdifferenz ($M_{Nachinspektion} - M_{Regelinspektion}$) geteilt durch Standardabweichung des Indikators bezogen auf die Inspektionsdaten aller Berliner Grundschulen. Nach Cohen (1988) steht d=0.20 für einen kleinen, d=0.50 für einen mittleren und d=0.80 für einen großen Effekt. KI = 95%-Konfidenzintervall.

[9] Die Konfidenzintervalle der Mittelwertsdifferenzen sind in der aktuell eingereichten Version des Manuskripts nicht vorhanden, sondern wurden nachträglich für die Dissertation hinzugefügt.

a. Entwicklung der Schulleistung



b. Entwicklung der Zusammensetzung der Schülerschaft: SES



c. Entwicklung der Zusammensetzung der Schülerschaft: Migrationshintergrund

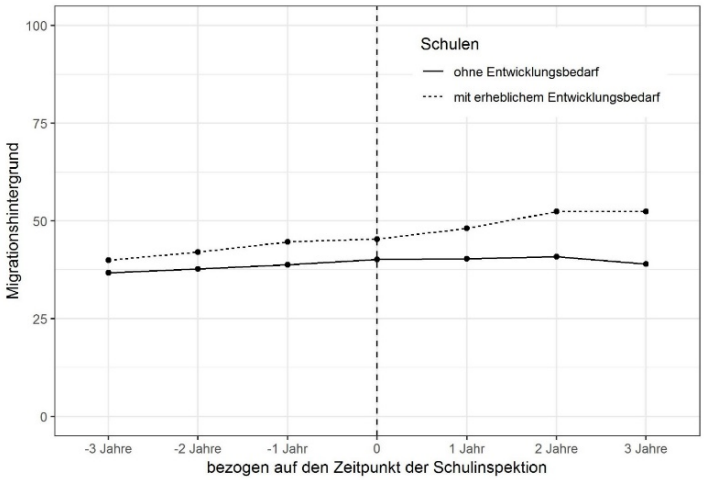


*Abbildung 1 Entwicklung der Schulen vor/nach der Schulinspektion: (a) Schulleistung (mittlere Schülerleistung aggregiert auf Schulebene), (b) SES (Prozentualer Anteil der Schüler*innen ohne Lernmittelzuzahlungsbefreiung) und (c) Migrationshintergrund (Prozentualer Anteil der Schüler*innen mit nicht-deutscher Herkunftssprache)*

Abb. 1a zeigt die mittlere Leistung von Schulen mit bzw. ohne Entwicklungsbedarf. Es zeigt sich, dass Schulen mit Entwicklungsbedarf schon vor der Schulinspektion im Mittel niedrigere Leistungen erzielten als Schulen ohne Entwicklungsbedarf. Dieser Trend setzte sich auch nach der Schulinspektion fort. Die Leistungen von Schulen mit Entwicklungsbedarf sind zudem *im* Jahr der Schulinspektion niedriger als zuvor. Weiter zeigte sich, dass sich auch bereits vor der Schulinspektion die Schulen mit bzw. ohne Entwicklungsbedarf in der Zusammensetzung der Schülerschaft im Mittel unterschieden: Der mittlere SES war an Schulen mit Entwicklungsbedarf niedriger (Abb. 1b), der Anteil von Kindern mit Migrationshintergrund war höher (Abb. 1c). Auch diese Mittelwerttrends setzten sich nach der Schulinspektion fort.

Bei der Frage nach der möglichen Wirkung der Diagnose Schule mit Entwicklungsbedarf wurden die vor der Schulinspektion bestehenden Unterschiede (hinsichtlich Leistung, SES und Anteil von Kindern mit Migrationshintergrund) zwischen Schulen mit bzw. ohne Entwicklungsbedarf in den spezifizierten Strukturgleichungsmodellen statistisch kontrolliert. Bezüglich der Entwicklung der Schulleistung (Forschungsfrage b) zeigten sich keine signifikanten Unterschiede der Schulen mit Entwicklungsbedarf gegenüber den Schulen ohne Entwicklungsbedarf (Abb. 2). Dies gilt sowohl für die Schulleistungen ein Jahr, zwei Jahre und drei Jahre nach der Schulinspektion. Die Differenzen der jeweiligen Regressionsgewichte sind nicht signifikant unterschiedlich (Tab. 5).

Tabelle 5 Forschungsfrage b und c, standardisierte Regressionskoeffizienten aus Strukturgleichungsmodellen

Prädiktoren aus dem Jahr der Schul-inspektion	Beta 1 Jahr nach SI [KI]	Beta 2 Jahre nach SI [KI]	Beta 3 Jahre nach SI [KI]	Differenz Beta SI+3- SI+2 [KI]	Differenz Beta SI+1- SI+3 [KI]	Differenz Beta SI+1- SI+2 [KI]
			AV1: Schulleistung			
Schule mit EB	-0.01 [-0.07; 0.05]	0.02 [-0.04; 0.09]	0.00 [-0.09; 0.08]	-0.03 [-0.11; 0.06]	0.00 [-0.09; 0.08]	-0.03 [-0.10; 0.04]
SES	0.31*** [0.18; 0.44]	0.26*** [0.11; 0.41]	0.24** [0.04; 0.44]	-0.02 [-0.22; 0.18]	0.07 [-0.13; 0.27]	-0.05 [-0.10; 0.21]
Migrations-hintergrund	-0.08 [-0.18; 0.02]	0.02 [-0.10; 0.14]	-0.04 [-0.19; 0.12]	-0.05 [-0.21; 0.10]	-0.05 [-0.20; 0.11]	-0.10 [-0.22; 0.02]
Leistung	0.53*** [0.44; 0.62]	0.64*** [0.53; 0.74]	0.58*** [0.44; 0.73	-0.05 [-0.20; 0.09]	-0.06 [-0.20; 0.09]	-0.11 [-0.22; 0.00]
	R^2 = 74%	R^2 = 71%	R^2 = 66%			
			AV2: SES			
Schule mit EB	0.01 [-0.02; 0.04]	0.02 [-0.02; 0.05]	-0.02 [-0.05; 0.02]	-0.03* [-0.07; 0.00]	0.03 [-0.01; 0.07]	0.00 [-0.04; 0.03]
SES	0.90*** [0.85; 0.95]	0.84*** [0.78; 0.91]	0.87*** [0.79; 0.96]	0.03 [-0.04; 0.11]	0.02 [-0.06; 0.11]	0.06 [0.00; 0.12]
Migrations-hintergrund	-0.03 [-0.08; 0.01]	-0.05 [-0.10; 0.01]	-0.01 [-0.08; 0.06]	0.04 [-0.02; 0.10]	-0.03 [-0.10; 0.04]	0.01 [-0.04; 0.06]
Leistung	0.06** [0.02; 0.11]	0.11*** [0.06; 0.16]	0.10*** [0.03; 0.17]	-0.01 [-0.07; 0.05]	-0.04 [-0.11; 0.03]	-0.05 [-0.10; 0.00]
	R^2 = 95%	R^2 = 93%	R^2 = 93%			

	AV3: Migrationshintergrund					
Schule mit EB	0.02* [0.00; 0.04]	0.02 [0.00; 0.04]	0.05* [0.01; 0.08]	0.02 [-0.01; 0.05]	-0.03 [-0.06; 0.01]	0.00 [-0.02; 0.02]
SES	0.00 [-0.04; 0.03]	0.05 [0.00; 0.10]	-0.01 [-0.09; 0.07]	-0.06 [-0.13; 0.01]	0.01 [-0.07; 0.09]	-0.05* [-0.09; -0.01]
Migrations-hintergrund	0.99*** [0.97; 1.01]	0.98*** [0.95; 1.01]	0.93*** [0.87; 0.98]	-0.05* [-0.10; 0.00]	0.06* [0.01; 0.11]	0.01 [-0.02; 0.04]
Leistung	0.00 [-0.03; 0.03]	-0.06*** [-0.10; -0.02]	-0.04 [-0.11; 0.02]	0.02 [-0.04; 0.08]	0.04 [-0.02; 0.10]	0.06*** [0.03; 0.09]
	R^2 =98%	R^2 =97%	R^2 =94%			

Anmerkungen. Schule mit EB: Schule mit Entwicklungsbedarf (kodiert=1), Schule ohne Entwicklungsbedarf (kodiert = 0). KI= 95%-Konfidenzintervall. SI = Schulinspektion. AV = Abhängige Variable. ***$p < .001$. **$p < .01$. *$p < .05$.

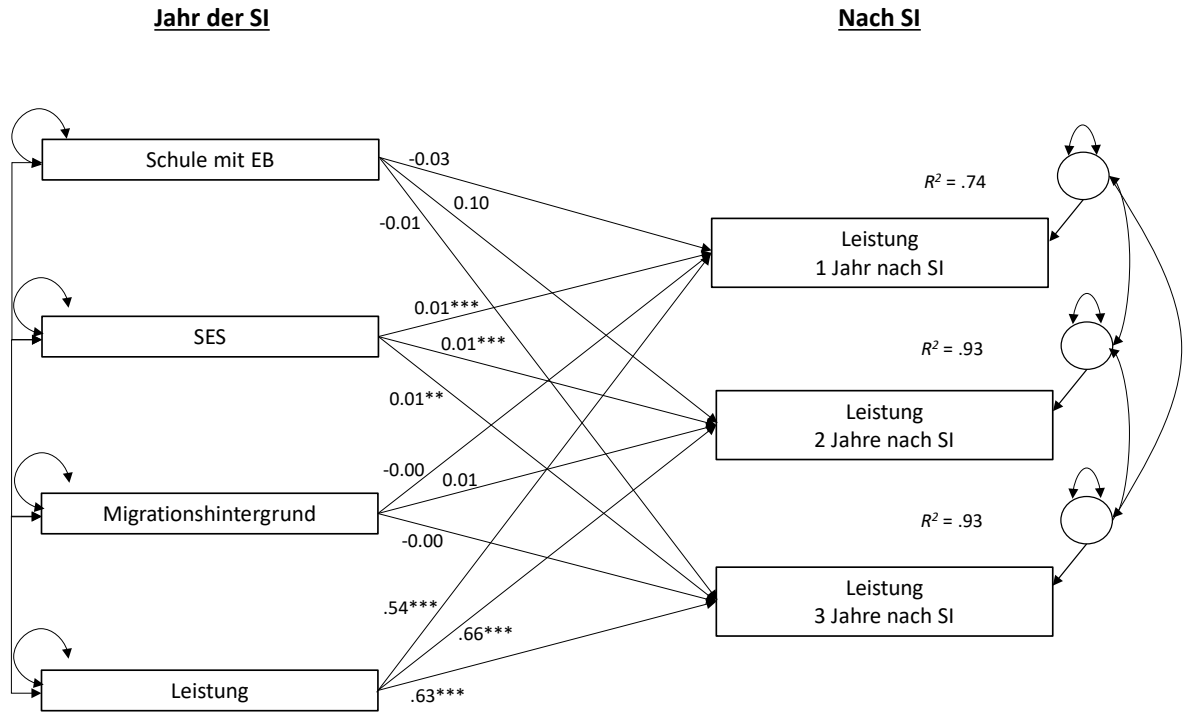


Abbildung 2 Strukturgleichungsmodell mit abhängigen Variablen Leistung (unstandardisierte Koeffizienten)
Anmerkungen. Schule mit EB: Schule mit Entwicklungsbedarf (kodiert=1), Schule ohne Entwicklungsbedarf (kodiert = 0). ***p < .001. **p < .01. *p < .05.

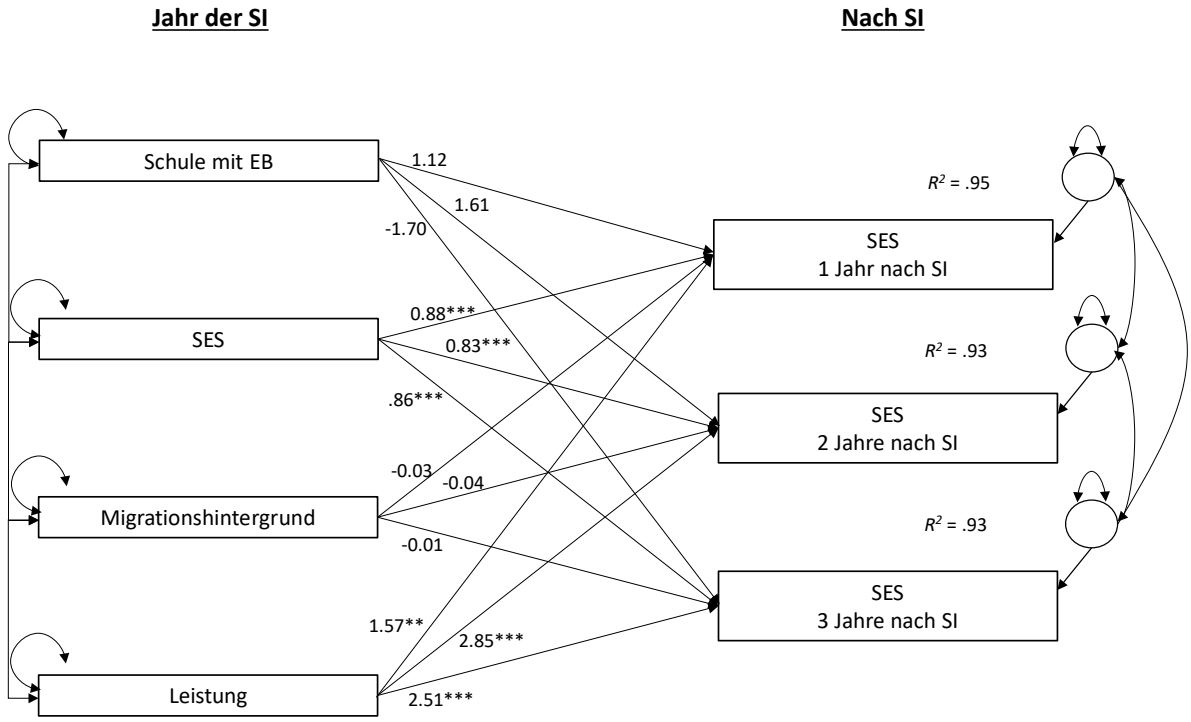


Abbildung 3 Strukturgleichungsmodell mit abhängigen Variablen SES (unstandardisierte Koeffizienten)
Anmerkungen. Schule mit EB: Schule mit Entwicklungsbedarf (kodiert=1), Schule ohne Entwicklungsbedarf (kodiert = 0). ***p < .001. **p < .01. *p < .05.

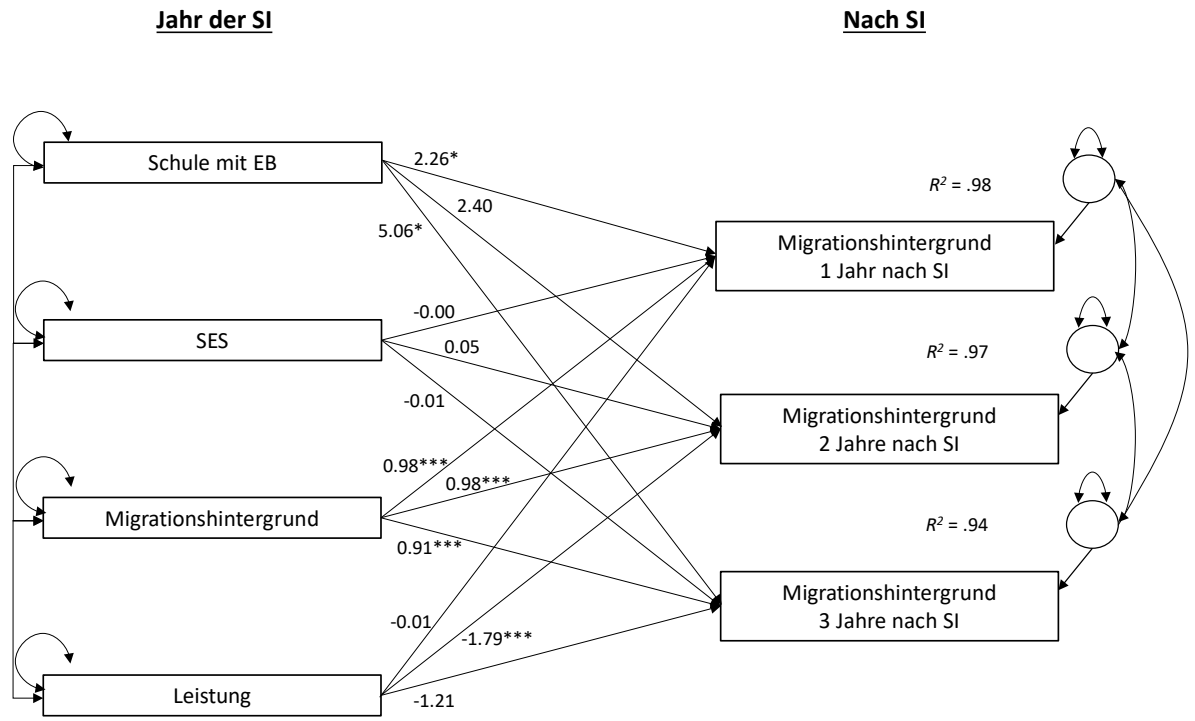


Abbildung 4 Strukturgleichungsmodell mit abhängigen Variablen Migrationshintergrund (unstandardisierte Koeffizienten)
Anmerkungen. Schule mit EB: Schule mit Entwicklungsbedarf (kodiert=1), Schule ohne Entwicklungsbedarf (kodiert = 0). ***$p < .001$. **$p < .01$. *$p < .05$.

Zur Entwicklung des SES der Schulen (Abb. 3) zeigt sich ebenfalls kein Unterschied zwischen Schulen mit und ohne Entwicklungsbedarf. Auch hier gilt dies für alle Zeitpunkte (ein Jahr, zwei Jahre und drei Jahre nach der Schulinspektion). Die Differenz der Regressionsgewichte bei Schulen mit diagnostiziertem Entwicklungsbedarf zum Zeitpunkt drei Jahre und zwei Jahre nach der Schulinspektion unterscheidet sich jedoch signifikant (diff = -0.03, $p < .05$), was darauf hinweist, dass der SES im Vergleich des Zeitpunkts zwei Jahre zu drei Jahre nach der Schulinspektion abnahm (Tab. 5).

Bezüglich der Entwicklung des Migrationshintergrunds der Schulen (Abb. 4) zeigen sich signifikante Unterschiede: bei Schulen mit Entwicklungsbedarf steigt der Anteil des Migrationshintergrunds der Schülerschaft nach der Schulinspektion an. Dieser Effekt wird signifikant zum Messzeitpunkt ein Jahr und drei Jahre nach der Schulinspektion. Die Regressionsgewichte unterscheiden sich dabei nicht signifikant zwischen den Messzeitpunkten (Tab. 5), was auf die zeitliche Stabilität dieses Zusammenhangs hinweist.

6. Diskussion

Die Schulinspektion evaluiert Schulen um Maßnahmen auszulösen, die zu einer Verbesserung der Schul- und Unterrichtsqualität sowie der Leistungen der Schülerinnen und Schüler beitragen sollen. Gleichzeitig sind durch die Veröffentlichung der Ergebnisse auch nicht-intendierte Nebenwirkungen möglich, zum Beispiel ein verändertes Schulwahlverhalten der Eltern in Folge der Ergebnisse und damit verbunden eine Veränderung der Zusammensetzung der Schülerschaft. Von besonderer Relevanz sind hierbei diejenigen Schulen für die im Rahmen der Schulinspektion Entwicklungsbedarf festgestellt wurde. Für diese Gruppe von Schulen ist national aber auch international wenig bekannt. In dieser Studie wurde daher erstmals untersucht, wie sich die Schul- und Unterrichtsqualität, die Schulleistung und die Zusammensetzung der Schülerschaft (SES und Anteil von Kindern mit Migrationshintergrund) von Schulen mit Entwicklungsbedarf in Folge der Schulinspektion verändert.

6.1 Die Entwicklung der Berliner Grundschulen mit Entwicklungsbedarf nach der Schulinspektion

Mit der Diagnose Schule mit Entwicklungsbedarf ist die Vorstellung verbunden, dass sich die Identifizierung dieser Schulen durch die Rückmeldung und durch die daraus folgende Unterstützung positiv auf die weitere Entwicklung der Schulen auswirkt. Der Forschungsstand zu Folgen der Schulinspektion im Allgemeinen lässt keine großen Effekte erwarten, die Ergebnisse zur Entwicklung von Schulen mit Entwicklungsbedarf aus Großbritannien und den Niederlanden verweisen jedoch auf tendenziell positive Effekte (Matthews und Sammons 2005; Ehren und Shackleton 2016).

Bezüglich der Entwicklung der Schul- und Unterrichtsqualität (a) konnten in dieser Studie positive Entwicklungen aufgezeigt werden, bspw. in der kognitiven Aktivierung und der konstruktiven Unterstützung. In der Klassenführung oder im Schulmanagement zeigten sich dagegen keine Verbesserungen. Insgesamt kann auf Basis der vorliegenden 10 Schulen mit Entwicklungsbedarf jedoch nur eingeschränkt von einer Verbesserung der Schul- und Unterrichtsqualität gesprochen werden. Weiterhin ist zu bedenken, dass die Schulen möglicherweise unterschiedlich mit den Schulinspektionsergebnissen umgegangen sind, je

nach Typ der Schule (aktiv, reaktiv, (selbst)-zufrieden, aktiv unzufrieden, passiv unzufrieden, Wurster und Gärtner 2013) und somit eine differenzielle Entwicklung stattfinden könnte, die sich nicht in einem Gesamtmittel aller Schulen mit Entwicklungsbedarf abbilden lässt.

Die vorliegenden Ergebnisse zeigen, dass sich die Leistungen der Schülerinnen und Schüler (b) an Schulen mit Entwicklungsbedarf nicht positiver entwickeln als an allen anderen Grundschulen. Die Diagnose Schule mit Entwicklungsbedarf kann die Entwicklung der Schulleistung für die drei Jahre nach der Schulinspektion nicht signifikant vorhersagen. Eine Ursache dafür kann darin liegen, dass alle Schulen ein Feedback durch die Schulinspektion bekommen, welches Veränderungen auslösen könnte und sich Schulen mit Entwicklungsbedarf nicht deutlich in der weiteren Entwicklung von anderen Schulen unterscheiden. Eine weitere Möglichkeit ist, dass die Entwicklung der Schulleistungen auch nach drei Jahren noch nicht sichtbar wird, sondern es eine längere Zeit dauert, bis sich entsprechende Veränderungsmechanismen in der Schule sichtbar auswirken. Es muss auch bedacht werden, ob die positiven Erwartungen realistisch waren: möglicherweise hat allein die Diagnose zu wenig Effekt und die verbundenen Unterstützungsangebote werden von den Schulen zu wenig genutzt. Dies bringt die Frage hervor, welche Unterstützungsmaßnahmen für eine positive Entwicklung relevant sind. Es gibt bislang einige Hinweise (Muijs et al. 2004; Hallinger und Heck 2011) wie „failing schools“/ Schulen mit großen Schwierigkeiten positive Veränderungsprozesse initiieren können, eine systematische Evaluation dessen gibt es jedoch nicht.

Die Entwicklung der Schülerzusammensetzung (c) wurde in dieser Studie betrachtet, um Hinweise darauf zu erlangen, inwieweit die Diagnose Schule mit Entwicklungsbedarf nichtintendierte Nebeneffekte im Bereich der elterlichen Schulwahl nach sich zieht: die Annahme, dass Eltern mit höherem SES bei der Schulwahl eher Schulen meiden, die Entwicklungsbedarf aufweisen und sich dadurch die Schülerzusammensetzung verändern würde, konnte hier nicht bestätigt werden. Es gibt jedoch den Hinweis, dass sich der SES mittelfristig verringert, da sich die Differenz der Regressionsgewichte der Schulen mit Entwicklungsbedarf von Jahr 2 auf Jahr 3 nach der Schulinspektion unterscheidet.

Bezüglich des Migrationshintergrunds der Schulen zeigten Schulen mit Entwicklungsbedarf jedoch einen signifikanten Anstieg um 2.3% (ein Jahr später) bis 5.1% (3 Jahre nach der Schulinspektion) pro Schule. Auf die Schülerinnen und Schüler der ersten Jahrgangsstufe bezogen, bedeutet das, dass bei durchschnittlich 50 Kindern pro Jahrgangsstufe 1, zwischen 9 (1 Jahr später) und 20 Kindern (3 Jahre später) mehr als vorher Migrationshintergrund aufweisen. Dies lässt vermuten, dass sich durch die Diagnose die Schülerschaft ändert und insbesondere Eltern ohne Migrationshintergrund andere Schulen bevorzugen. Dieser Befund geht mit Ergebnissen zur elterlichen Schulwahl einher, dass Eltern sich je nach Migrationshintergrund in ihrer Schulwahl unterscheiden (bspw. Riedel et al. 2010). Möglich ist, dass die öffentliche Verfügbarkeit der Schulinspektionsberichte sowie des Anteils der Schülerschaft mit Migrationshintergrund dazu beiträgt. Es könnte so zu einer stärkeren Segregation der Schülerschaft kommen und Schwierigkeiten an diesen Schulen beispielsweise hinsichtlich der Leistungsentwicklung zusätzlich erhöhen (van Ewijk und Sleegers 2010; Mickelson et al. 2013). Diesen Schwierigkeiten müsste die Schule dann zusätzlich begegnen, was wiederum Ressourcen der Schule in Anspruch nimmt. Diese Entwicklung der Schülerschaft sollte daher in den Unterstützungsangeboten an diese Schulen berücksichtigt werden.

6.2 Grenzen der Studie

Zur Untersuchung der Entwicklung von Schulen mit Entwicklungsbedarf analysierte diese Studie erstmals einen umfangreichen Datensatz aus verschiedenen Quellen zu allen öffentlichen Berliner Grundschulen. Dennoch bestehen Grenzen der Aussagekraft der vorliegenden Studie. (1) Möglicherweise werden positive Entwicklungsprozesse durch die hier verwendeten Variablen nicht sichtbar, da sich andere Outcome-Variablen verändert haben: beispielsweise könnten sich das Wohlbefinden der Schülerinnen und Schüler sowie der Lehrkräfte durch angestoßene Prozesse nach der Schulinspektion verbessert haben, die Leistungen bleiben aber (noch) unverändert. Zwei weitere Punkte schließen sich hier an: Ab wann kann man von stattfindender Entwicklung sprechen? Und wann ist der richtige Zeitpunkt um die Entwicklung zu messen? Als Zeitpunkt für die weiteren Entwicklungen haben wir drei

Messzeitpunkte nach der Schulinspektion gewählt (ein Jahr, zwei Jahre, drei Jahre danach). (2) Die Betrachtung der weiteren Entwicklung der Unterrichts- und Schulqualität (a) kann leider nur für die Schulen mit Entwicklungsbedarf erfolgen (N = 10), die in der zweiten Welle der Schulinspektion bereits nachinspiziert wurden. Zudem kann hier nicht mit Schulen ohne Entwicklungsbedarf verglichen werden, da dort erst in der nächsten Regelinspektion (fünf Jahre später) wieder Daten erhoben werden. (3) Es liegen keine konkreten Angaben aus den Schulen vor, welche Unterstützungsleistungen sie nach der Schulinspektion nutzten und für wie wirksam sie die Unterstützung wahrnahmen. (4) Schließlich wird in dieser Studie ausschließlich die Situation in Berlin dargestellt. Inwiefern die Ergebnisse auf andere Bundesländer generalisiert werden können, ist durch die größere Konkurrenz zwischen Schulen in einem Stadtstaat, sowie durch die öffentliche Verfügbarkeit der Schulinspektionsberichte in Berlin eine offene Forschungsfrage.

6.3 Schlussfolgerungen

Mit Hilfe der Kombination verschiedener Daten zu 333 Berliner Grundschulen aus den Jahren 2010-2017 konnte erstmals die Entwicklung von Schulen mit Entwicklungsbedarf untersucht werden. Erwartete positive Wirkungen der Schulinspektion haben sich teilweise eingestellt, es kam jedoch auch zu unerwarteten Nebeneffekten.

Nach der Theorie, dass zusätzlicher Druck durch die Schulinspektion positive Entwicklungen nach sich zieht (bspw. Kotthoff und Böttcher 2009), müsste man vermehrte Handlungen im Anschluss an die Schulinspektion erwarten, die sich auch auf die Leistungen der Schülerinnen und Schüler auswirken. Es zeigte sich hierbei, dass es positive Entwicklungen im Bereich der Unterrichtsqualität gibt, dies jedoch nur an einer sehr geringen Stichprobe untersucht werden konnte. Dass sich die Schülerleistung an Schulen mit Entwicklungsbedarf im Anschluss an eine Schulinspektion besser entwickelt als an anderen Schulen kann nicht nachgewiesen werden. Stattdessen wurden Hinweise auf eine zunehmende Segregation zwischen Schulen sichtbar: an Schulen mit diagnostiziertem Entwicklungsbedarf nahm der Anteil an Schülerinnen und Schülern mit Migrationshintergrund zu. Dies kann darauf

hinweisen, dass Eltern ohne Migrationshintergrund Schulen mit Entwicklungsbedarf eher meiden.

Literaturverzeichnis

Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement, 26*(1), 32–56.

Böhm-Kasper, O., Selders, O., & Lambrecht, M. (2016). Schulinspektion und Schulentwicklung – Ergebnisse der quantitativen Schulleitungsbefragung. In Arbeitsgruppe Schulinspektion (Hrsg.), *Educational governance. Band 25: Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten* (1. Aufl. 2016, S. 1–50). Wiesbaden: Springer VS.

Brimblecombe, N., Shaw, M., & Ormston, M. (1996). Teachers' intention to change practice as a result of Ofsted school inspections. *Educational Management & Administration, 24*(4), 339–354.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9), 1045–1057.

Clausen, M. (2006). Warum wählen Sie genau diese Schule? Eine inhaltsanalytische Untersuchung elterlicher Begründungen der Wahl der Einzelschule innerhalb eines Bildungsgangs. *Zeitschrift für Pädagogik, 52*(1), 69–90.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Hillsdale NJ: Erlbaum.

Cullingford, C. (1999). *An Inspector Calls*. London: Routledge.

Dedering, K., Katenbrink, N., Schaffer, G., & Wischer, B. (2016). „Veränderung unter Druck". In Arbeitsgruppe Schulinspektion (Hrsg.), *Educational governance. Band 25: Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten* (1. Aufl. 2016, S. 201–226). Wiesbaden: Springer VS.

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein, & E. Terhart (Hrsg.), *Zeitschrift für Pädagogik. Beiheft. Bd. 41: Qualität und Qualitätssicherung im Bildungsbereich; Schule, Sozialpädagogik, Hochschule* (S. 73–92). Weinheim: Beltz.

Döbert, H., & Dedering, K. (2008). Externe Evaluation von Schulen in Deutschland. Die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede. In H. Döbert (Hrsg.), *Externe Evaluation von Schulen. Historische, rechtliche und vergleichende Aspekte* (S. 63–151). 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 relationship between school inspections, school characteristics and school improvement. *British Journal of Educational Studies, 56*(2), 205–227.

Ehren, M. C. M., & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability, 28*(4), 299–321.

Ehren, M. C.M., & Swanborn, M. S.L. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement, 23*(2), 257–280.

Enders, C. K. (2010). *Methodology in the social sciences: Applied missing data analysis*. New York: Guilford Press.

Faubert, V. (2009). School evaluation: Current practices in OECD countries and a literature review. *OECD Educational Working Papers 42*.

Fincke, G., & Lange, S. (2012). *Segregation an Grundschulen: Der Einfluss der elterlichen Schulwahl*. Berlin. https://www.svr-migration.de/wp-content/uploads/2014/11/Segregation_an_Grundschulen_SVR-FB_WEB.pdf.

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

Gärtner, H., & Pant, H. A. (2011). How valid are school inspections? Problems and strategies for validating processes and results. *Studies in Educational Evaluation, 37*(2-3), 85–93.

Gärtner, H., Wurster, S., & Pant, H. A. (2014). The effect of school inspections on school improvement. *School Effectiveness and School Improvement, 25*(4), 489–508.

Hallinger, P., & Heck, R. H. (2011). Exploring the journey of school improvement: classifying and analyzing patterns of change in school improvement processes and learning outcomes. *School Effectiveness and School Improvement, 22*(1), 1–27.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297–327.

Jurczok, A. (2019). *Schulwahl unter „gleichwertigen" Einzelschulen*. Wiesbaden: Springer Fachmedien Wiesbaden.

Kemethofer, D., & Helm, C. (2017). Effekte durch Rechenschaftsdruck im Kontext von Schulinspektionen: Ein Vergleich von Schweden und Österreich. In M. Pietsch & I. Hosenfeld (Hrsg.), *Empirische Pädagogik. 31. Jahrgang, Heft 2 (2017): Inspektionsbasierte Schul- und Unterrichtsentwicklung* (S. 138–157). Landau in der Pfalz: Verlag Empirische Pädagogik e. V.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (Hrsg.), *TIMSS – Impulse für Schule und Unterricht* (S. 43–57). Bonn.

Kotthoff, H.-G., & Böttcher, W. (2009). Neue Formen der „Schulinspektion": Wirkungshoffnungen und Wirksamkeit im Spiegel empirischer Bildungsforschung. In H. Altrichter (Hrsg.), *Educational governance. Bd. 7: Neue Steuerung im Schulsystem. Ein Handbuch* (1. Aufl., S. 295–325). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kunter, M., Baumert, J., Blum, W., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Landwehr, N. (2011). Wirkungen und Wirksamkeit der externen Schulevaluation. In C. Quesel (Hrsg.), *Wirkungen und Wirksamkeit der externen Schulevaluation* (S. 35–70). Bern: Hep der Bildungsverl.

Lee, J. (2006). Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education, 81*(4), 43–64.

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

Maaz, K., Böse, S., & Neumann, M. (2016). *BONUS-Studie. Wissenschaftliche Begleitung und Evaluation des Bonus-Programms zur Unterstützung von Schulen in schwieriger Lage in Berlin*. Zwischenbericht über die erste Schulleiterbefragung aus dem Schuljahr 2013/2014. Zugegriffen: 21. Juni 2017.

Matthews, P., & Sammons, P. (2005). Survival of the weakest: the differential improvement of schools causing concern in England. *London Review of Education, 3*(2), 159–176.

Mickelson, R. A., Bottia, M. C., & Lambert, R. (2013). Effects of school racial composition on K–12 mathematics outcomes. *Review of Educational Research, 83*(1), 121–158.

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.

Noreisch, K. (2007). School catchment area evasion: the case of Berlin, Germany. *Journal of Education Policy, 22*(1), 69–90.

Penninckx, M., Vanhoof, J., Maeyer, S. de, & van Petegem, P. (2016). Explaining effects and side effects of school inspections: a path analysis. *School Effectiveness and School Improvement, 27*(3), 333–347.

Perryman, J. (2010). Improvement after inspection. *Improving Schools, 13*(2), 182–196.

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.

Preuße, D., Pohl, J., & Gärtner, H. (2019). *Wahrgenommene Auswirkungen der Schulinspektion aus Sicht von Schulleitungen und Schulaufsicht in Berlin*. https://www.isq-bb.de/wordpress/wp-content/uploads/2019/09/Ergebnisbericht_Berlin_2019-07.pdf. Zugegriffen: 1. Oktober 2019.

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

Riedel, A., Schneider, K., Schuchart, C., & Weishaupt, H. (2010). School choice in German primary schools: How binding are school districts? *Journal for Educational Research Online / Journal für Bildungsforschung Online, 2*(1), 94–120.

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.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2).

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1. ed.). Oxford, New York: Pergamon.

Senatsverwaltung für Bildung, Jugend und Familie (2017). *Berliner Schulwegweiser. Wohin nach der Grundschule? Schuljahr 2018/2019*. Berlin. 11.02.2020.

Senatsverwaltung für Bildung, Jugend und Wissenschaft (2012). *Berliner Schule. Zweite Runde Schulinspektion in Berlin*. Berlin.

Senatsverwaltung für Bildung, Jugend und Wissenschaft (2013). *Handlungsrahmen Schulqualität in Berlin. Qualitätsbereiche und Qualitätsmerkmale*. Berlin.

Senatsverwaltung für Bildung, Jugend und Wissenschaft (2014). *7 Jahre Schulinspektion in Berlin*. Berlin.

Senatsverwaltung für Bildung, Wissenschaft und Forschung (2009). *Bildung für Berlin. Handbuch Schulinspektion*. Berlin.

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

Sommer, N. (2010). Schulen mit „gravierenden Mängeln“. In W. Böttcher (Hrsg.), *Evaluation, Bildung und Gesellschaft. Steuerungsinstrumente zwischen Anspruch und Wirklichkeit* (S. 209–227). Münster u.a: Waxmann.

van Ewijk, R., & Sleegers, P.J.C. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational research review, 5*, 134–150.

Wurster, S., & Gärtner, H. (2013). Schulen im Umgang mit Schulinspektion und deren Ergebnissen. *Zeitschrift für Pädagogik, 59*(3), 425–445.

Studie 3 - Online Supplement

Deskriptive Merkmale der Berliner Grundschulen getrennt nach Schulen mit erheblichem Entwicklungsbedarf (N = 22) und ohne Entwicklungsbedarf (N = 311)

	Schulen mit erheblichem Entwicklungsbedarf					Schulen ohne erheblichen Entwicklungsbedarf				
	N	*M*	*SD*	Min	Max	*N*	*M*	*SD*	Min	Max
Schulleistung[a] im Jahr der SI	22	-0.30	0.87	-1.79	1.46	311	0.08	0.97	-2.36	2.23
Schulleistung[a] ein Jahr nach SI	18	-0.23	0.87	-1.85	1.27	250	0.16	0.97	-2.56	2.57
Schulleistung[a] zwei Jahre nach SI	16	-0.22	0.81	-1.51	1.62	207	0.10	1.00	-2.37	2.08
Schulleistung[a] drei Jahre nach SI	12	-0.43	0.95	-1.70	0.97	144	0.12	1.01	-2.45	1.98
Migrationshintergrund (%) im Jahr der SI	22	45.34	21.70	5.07	83.48	311	40.19	28.48	0.31	98.59
Migrationshintergrund (%) ein Jahr nach SI	18	48.11	23.56	5.26	88.43	250	40.35	27.69	0.00	95.70
Migrationshintergrund (%) zwei Jahre nach SI	16	52.46	22.74	4.65	86.63	207	40.83	28.00	0.27	96.26
Migrationshintergrund (%) drei Jahre nach SI	12	52.45	23.59	4.27	86.74	144	38.94	27.35	0.61	96.11
SES (%) im Jahr der SI	22	53.57	20.37	11.97	86.05	311	62.86	25.52	1.21	99.74
SES (%) ein Jahr nach SI	18	56.33	18.67	14.05	82.90	250	63.90	24.84	8.02	97.81
SES (%) zwei Jahre nach SI	16	54.68	17.23	11.92	84.54	207	62.65	25.28	8.05	98.68
SES (%) drei Jahre nach SI	12	50.55	18.48	20.69	80.57	144	65.25	24.83	9.33	99.55

	Schulen mit erheblichem Entwicklungsbedarf					Schulen ohne erheblichen Entwicklungsbedarf				
	N	*M*	*SD*	Min	Max	*N*	*M*	*SD*	Min	Max
Unterrichtsqualität										
Effiziente Klassenführung	22	3.69	0.16	3.27	3.88	311	3.79	0.12	3.24	3.99
Kognitive Aktivierung	22	2.22	0.22	1.95	2.71	311	2.53	0.23	1.96	3.14
Konstruktive Unterstützung	22	2.86	0.18	2.60	3.26	311	3.12	0.16	2.69	3.61
Schulqualität										
Schulkultur	22	1.64	0.95	1.00	4.00	311	2.85	0.96	1.00	4.00
Schulmanagement	22	2.08	0.78	1.00	4.00	311	3.31	0.59	1.50	4.00
Kooperation/Koordination	22	2.39	0.55	1.00	4.00	311	3.38	0.45	2.00	4.00
Personal	22	1.59	0.80	1.00	4.00	309	2.93	0.78	1.00	4.00

Anmerkungen. Min = Minimum. Max = Maximum. SES (%): 100% – schul-spezifischer Anteil der Schülerinnen und Schüler, die lernmittelzuzahlungsbefreit sind (d.h. ein SES-Wert von 0% steht für ein sehr niedriges Niveau des SES der Schülerschaft dieser Schule). SI = Schulinspektion. Die Qualitätsindikatoren der Unterrichts- und Schulqualität werden auf einer 4-stufigen Bewertungsskala von 1=„schwach ausgeprägt“, 2=„eher schwach ausgeprägt“, 3=„eher stark ausgeprägt“, 4=„stark ausgeprägt“ durch die Schulinspektion eingeschätzt.

[a] z-standardisierte mittlere Schulleistung

5 Gesamtdiskussion

Diese Dissertation hat das Ziel, die Variabilität schulischer Lernumwelten zu beleuchten, indem sie Unterrichts- und Schulqualität sowie Kontextmerkmale und Leistungen von Schulen in den Blick nimmt. Es wurden zwei übergreifende Forschungsfragen entwickelt, die in drei empirischen Studien überprüft wurden:

(1) Wie unterschiedlich sind Schulen in ihrer Unterrichtsqualität? (Studie 1)

(2) Unterscheiden sich schulische Prozessmerkmale und Outputs je nach Kontext?
   a) Kontext im Hinblick auf Schülerzusammensetzung (Komposition; Studie 2)
   b) Kontext im Hinblick auf bildungspolitische Intervention (Studie 3)

In diesem Kapitel sollen die Ergebnisse dieser Arbeit diskutiert werden. Dafür werden zunächst die zentralen Ergebnisse der drei empirischen Studien hinsichtlich der Forschungsfragen zusammengefasst (Kapitel 5.1). Anschließend werden die Ergebnisse studienübergreifend in Bezug auf das Gesamtmodell (siehe Abb. 1) dieser Arbeit reflektiert (Kapitel 5.2). Daraufhin werden Limitationen der Arbeit beleuchtet (Kapitel 5.3), bevor Implikationen aus den Ergebnissen für die Forschung, die Bildungspolitik und die Praxis gezogen werden (Kapitel 5.4). Abschließend wird ein studienübergreifendes Fazit gezogen (Kapitel 5.5).

5.1 Zusammenfassung zentraler Befunde der Studien

Studie 1 adressierte die erste Forschungsfrage zu Unterschieden der Schulen in ihrer Unterrichtsqualität, indem statistische Kennwerte zur (1) Übereinstimmung, (2) Variabilität und (3) Reliabilität für zahlreiche Merkmale von Unterrichtsqualität betrachtet wurden. Dafür wurden Schülerurteile zu mehreren Merkmalen der effizienten Klassenführung, der kognitiven Aktivierung und der konstruktiven Unterstützung (Klieme et al., 2001; Kunter et al., 2011) untersucht. Um eine systematische Übersicht hinsichtlich der Übereinstimmung innerhalb der Schulen, der Unterschiede zwischen Schulen und der Reliabilität von aggregierten Urteilen der Schülerinnen und Schüler zu erlangen, wurden drei Indizes für eine umfassende Datenbasis (über 1.3 Millionen Schülerinnen und Schüler der internationalen PISA-Erhebungen) berechnet (vgl. bspw. Bliese, 2000; James et al., 1984; Lüdtke et al., 2006): (1) das Maß $r_{WG(J)}$ zur

Übereinstimmungsmessung der Schülerurteile innerhalb der Schulen, (2) die Varianzaufklärung mit Hilfe von Intraklassenkorrelationen ICC(1) und (3) die Reliabilität der aggregierten Schülerurteile mittels der ICC(2). Die Ergebnisse der Studie zeigen, dass (1) die Übereinstimmung der Urteile von Schülerinnen und Schülern in den meisten Merkmalen der Unterrichtsqualität moderat bis stark ausfiel. Es konnte weiterhin gezeigt werden, dass (2) sich Merkmale der Unterrichtsqualität aus Sicht der Schülerschaft systematisch zwischen Schulen unterscheiden, wenngleich die Unterschiede zwischen den Schulen nach bestehenden Konventionen zur Beurteilung der ICC(1) als „klein" bis „mittel" zu bezeichnen sind (LeBreton & Senter, 2008). Zudem wurden empirische Referenzwerte für den spezifischen Anwendungskontext von Merkmalen der Unterrichtsqualität auf Schulebene abgeleitet: Eine ICC(1)= 0.03 spricht für „kleine", eine ICC(1)= 0.08 für „mittlere" und eine ICC(1)= 0.19 für „große" Unterschiede zwischen den Schulen hinsichtlich der Unterrichtsqualität aus Sicht der Schülerinnen und Schüler. Zur Reliabilität der aggregierten Urteile von Schülerinnen und Schülern auf Schulebene konnte (3) festgestellt werden, dass diese in einem Großteil der PISA-Länder unter dem als „akzeptabel" geltenden Grenzwert (LeBreton & Senter, 2008) lag. Als zusammenfassende Antwort auf Forschungsfrage 1 kann somit gesagt werden, dass Schulen sich in ihrer Unterrichtsqualität (je nach betrachtetem Merkmal) in kleinem bis mittlerem Maße unterscheiden. Studie 1 leistet damit einen wichtigen Beitrag zur systematischen Beschreibung der Variabilität schulischer Lernumwelten.

Studie 2 untersuchte die Zusammenhänge von Kompositionsmerkmalen der Schülerschaft (Kontext) mit schulischen Prozessmerkmalen und Leistungsergebnissen (Output) (Forschungsfrage 2a). Um diese genau zu analysieren, wurden mit Hilfe einer Kombination verschiedener Datenquellen (amtliche Statistik, Schulinspektionsdaten, VERA-Daten) von Berliner Grundschulen die Forschungsfragen nach den Zusammenhängen (1) zwischen Komposition und Unterrichtsqualität, (2) zwischen Komposition und Schulqualität und (3) zwischen Komposition und späterer Schulleistung, vermittelt durch Unterrichts- und Schulqualität beantwortet. Zum Zusammenhang (1) zwischen Kompositionsmerkmalen der Schülerschaft und Unterrichtsqualität konnte gezeigt werden, dass bei zwei der vier gemeinsam untersuchten Kompositionsmerkmale (mittlerer SES, mittlere Schulleistung) positive

Zusammenhänge mit Unterrichtsqualität bestehen: Je höher der mittlere SES der Schulen, umso höher war auch die Klassenführung und kognitive Aktivierung. War die mittlere Schulleistung höher, zeigten die Schulen eine höhere kognitive Aktivierung und konstruktive Unterstützung. Zwischen (2) den Kompositionsmerkmalen der Schülerschaft und der Schulqualität konnte kaum ein Zusammenhang festgestellt werden. (3) Die Mediation des Zusammenhangs der Kompositionsmerkmale der Schülerschaft mit der späteren Schulleistung durch Unterrichts- und Schulqualität konnte nur teilweise nachgewiesen werden. Lediglich ein indirekter Effekt auf die spätere Schulleistung war signifikant: Schulen mit hohem SES zeigten eine bessere Klassenführung, welche wiederum zu besseren Schulleistungen führte. Weder die Merkmale kognitive Aktivierung, konstruktive Unterstützung noch eines der Merkmale der Schulqualität konnten jedoch den Zusammenhang zwischen Kompositionsmerkmalen der Schülerschaft und späterer Schulleistung mediieren. Die Forschungsfrage 2a kann daher zusammenfassend wie folgt beantwortet werden: Schulische Prozessmerkmale, im Sinne von Unterrichtsqualität und Output im Sinne von Schulleistungen, unterscheiden sich in dieser Studie zwischen den Schulen abhängig von der Komposition ihrer Schülerschaft. Die Mediation der Effekte von Komposition auf den Output durch die schulischen Prozesse konnte dabei nur zu einem geringen Teil gezeigt werden.

Studie 3 untersuchte die Folgen einer bildungspolitischen Intervention (Kontext) auf spätere schulische Prozessmerkmale und Leistungsergebnisse (Output) (Forschungsfrage 2b). Dabei wurde als Kontextbedingung die bildungspolitische Intervention verstanden, einem Teil der Schulen im Zuge der Evaluation durch die Schulinspektion einen „erheblichen Entwicklungsbedarf“ (Senatsverwaltung für Bildung, Wissenschaft und Forschung, 2009) zu diagnostizieren. Wie auch in Forschungsfrage 2a ging es dabei darum, Unterschiede der Schulen bezüglich der Entwicklung der schulischen Prozessmerkmale und Outputs im Zusammenhang mit Kontextmerkmalen zu betrachten. In dieser Studie wurde dafür ebenso auf Basis der Kombination verschiedener Datenquellen (amtliche Statistik, Schulinspektionsdaten, VERA-Daten) von Berliner Grundschulen die Entwicklung (1) der Unterrichts- und Schulqualität, (2) der mittleren Schulleistungen und (3) der Zusammensetzung der Schülerschaft von Schulen mit erheblichem Entwicklungsbedarf untersucht. Dabei konnte (1)

eine geringe Verbesserung der Unterrichts- und Schulqualität dieser speziellen Gruppe von Schulen gezeigt werden: Bezüglich der Unterrichtsqualität verbesserten sich die kognitive Aktivierung und konstruktive Unterstützung in den Schulen; bezüglich der Schulqualität verbesserte sich die Kooperation und Koordination und die Personalentwicklung leicht. Zudem wurde gezeigt, dass sich (2) die mittleren Schulleistungen im Vergleich zu den Schulen ohne diagnostiziertem Entwicklungsbedarf durch diese bildungspolitische Intervention nicht signifikant verbessern. (3) Bezüglich eines potentiellen nicht-intendierten Nebeneffekts der Schulinspektionsdiagnose „Schule mit erheblichem Entwicklungsbedarf" konnte folgendes festgestellt werden: Die Schülerschaft veränderte sich infolgedessen hinsichtlich des mittleren SES nicht signifikant. Es gab jedoch Hinweise darauf, dass die Anzahl an Schülerinnen und Schülern mit Migrationshintergrund potentiell steigt, wenn die Schule als erheblich entwicklungsbedürftig diagnostiziert wurde. Zusammenfassend kann festgehalten werden, dass sich die schulischen Prozessmerkmale im Sinne der Unterrichts- und Schulqualität abhängig von der bildungspolitischen Intervention als Kontextbedingung leicht verbessern, der Output im Sinne der Schulleistung sich nicht verbessert, es jedoch Anzeichen für eine Veränderung der Zusammensetzung der Schülerschaft gibt.

5.2 Studienübergreifende Reflexion

In diesem Kapitel werden wichtige Ergebnisse der Studien hinsichtlich studienübergreifender Aspekte zur Variabilität schulischer Lernumwelten bezüglich der zwei identifizierten Leitthemen reflektiert und in den Forschungsstand eingeordnet (siehe Kapitel 3): (1) Schule als relevante Analyseeinheit sowie (2) die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität als auch für die Schulleistung.

5.2.1 Schule als relevante Analyseeinheit

In allen drei Studien ist die Schulebene bzw. die Relevanz der Einzelschule zentral. Diese Relevanz ergibt sich aus verschiedenen Forschungssträngen wie der Schuleffektivitätsforschung, der Schulentwicklungsforschung, aber auch durch empirische Methoden wie der Untersuchung der Varianzaufklärung auf verschiedenen Ebenen, wie sie

beispielsweise im Zuge der Analyse der PISA-Daten (OECD, 2005; Kapitel 1) angewendet wird. Dabei besteht Konsens zumindest dahingehend, dass nach langer Streitigkeit die Einzelschule durchaus eine gewisse Rolle für die Leistung von Schülerinnen und Schülern spielt (Brookover et al., 1979; Coleman et al., 1966; Mortimore et al., 1988; Teddlie & Stringfield, 1993). Dies kann beispielsweise daran gezeigt werden, dass die Varianz in Schülerleistungen und Unterrichtsmerkmalen unter anderem durch Merkmale der Einzelschule erklärt werden kann (Baumert et al., 2006b; Klieme & Rakoczy, 2003; OECD, 2005; Wurster & Gärtner, 2013; Wurster & Feldhoff, 2019). Nichtsdestotrotz besteht noch keine Einigkeit darüber, welche Bedeutung dem Einfluss der Einzelschule zukommt. So werden beispielsweise die Ergebnisse der Varianzaufklärung durch die Einzelschule weiterhin ganz unterschiedlich interpretiert: Ist 10% Varianzaufklärung eines Unterrichtsmerkmals auf Schulebene beispielsweise als „hoch“ oder „niedrig“ zu werten? Um eine Einschätzung treffen zu können, ob eine gewisse Varianzaufklärung als „groß“, „mittel“ oder „klein“ bewertet werden kann, benötigt es eine systematische Übersicht über typische empirisch gefundene Ausprägungen dieser Werte (siehe bspw. für Korrelationen Bosco, Aguinis, Singh, Field & Pierce, 2015; Hemphill, 2003). Solch eine Systematisierung wurde für Merkmale der Unterrichtsqualität mit Hilfe der umfangreichen Übersichten unter anderem zur Höhe von Intraklassenkorrelationen (ICC (1)) in Studie 1 dieser Dissertation geliefert (vgl. Kapitel 4.1; Ergebniszusammenfassung Studie 1 in Kapitel 5.1). Zudem zeigte sich durch die moderate bis starke Übereinstimmung der Schülerurteile zur Unterrichtsqualität innerhalb der Schulen sowie durch die Variabilität zwischen den Schulen eine „schulspezifische Unterrichtsqualität“ (Studie 1). Das bedeutet, dass Schulen eine zu einem gewissen Grad einheitliche Unterrichtsqualität besitzen, was den Grundgedanken von Fend empirisch unterstützt, dass Schulen pädagogische Handlungseinheiten sind (vgl. bspw. Fend, 1988). Dieses Ergebnis ist auch aus methodischer Sicht interessant: Da es für die Bestimmung von Stichprobengrößen bei cluster-randomisierten Studien notwendig ist, zu wissen, wie groß Unterschiede zwischen den Clustern hinsichtlich des Zielkonstrukts sind, können diese Ergebnisse als empirische Referenzwerte für die Planung von Interventionsstudien auf Schulebene dienen (vgl. Kapitel 4.1, Diskussion Studie 1).

Unter anderem durch die Ergebnisse aus Studie 1 legitimiert, können Unterrichtsqualitätsmerkmale auf Schulebene auch in Zusammenhang mit anderen Merkmalen auf Schulebene (bspw. mit aggregierten Schülerleistungen) analysiert werden, so wie dies in Studie 2 und 3 erfolgte (vgl. die akzeptable Präzision bei der Analyse von aggregierten Daten auf Schulebene vs. Individualebene bei Jacob, Goddard & Kim, 2014). Dort wurden ausgehend von der inhaltlichen Prämisse, dass die Einzelschule als Gesamtkomplex auf Kontextmerkmale reagiert und diese gemeinsam mit den schulinternen Prozessmerkmalen den schulischen Output bedingen, alle interessierenden Merkmale aggregiert auf Schulebene betrachtet. Dabei zeigten sich interessante Ergebnisse, wie beispielsweise die Zusammenhänge von Kontextmerkmalen mit Unterrichtsqualitätsmerkmalen und Leistungen. Betrachtet man die Schulqualitätsmerkmale der Einzelschulen (Schulkultur, Schulmanagement, Kooperation und Koordination, Personal), zeigte sich in den Studien 2 und 3, dass diese weniger auf den Kontext (sei es als Form der Komposition in Studie 2 oder der Intervention in Studie 3) reagieren als die auf Schulebene aggregierten Merkmale der Unterrichtsqualität. Dieses Ergebnis bestätigt damit die Annahme, dass die proximalen Merkmale (des Unterrichts) im Gegensatz zu den distalen Bedingungen des Lernens (durch die Schule) besonders relevant für das Lernen von Schülerinnen und Schülern sind (vgl. bspw. Ditton, 2000; Kunter & Baumert, 2008). Dies wird damit begründet, dass diese Merkmale *näher* am tatsächlichen Lehr-Lerngeschehen liegen.

Eine weitere Erkenntnis aus den Studien dieser Dissertation ist, dass sich Schulen faktisch als Gesamtsystem mit bestimmten Rahmenbedingungen, also Kontextmerkmalen (sei es Komposition oder Intervention) auseinandersetzen müssen. Beispielsweise scheint es unrealistisch, dass einzelne Lehrkräfte auf Anreize der Schulinspektion, die die gesamte Schule betreffen, deutlich anders reagieren als andere Lehrkräfte und die Schulebene keine Rolle spielt. Es ist zwar vorstellbar, dass Lehrkräfte mit bestimmten schulischen Kontextmerkmalen unterschiedlich umgehen (bspw. mit Heterogenität im Unterricht), möglicherweise würde sich solch eine größere Variabilität im Umgang mit Kontextmerkmalen innerhalb der Schule aber auch gerade durch eine zu wenig vorhandene oder wenig wirksame Strategie auf der Schulebene (also bspw. durch die Schulleitung und Schulkonferenzen) ergeben. Auch die Ressourcenzuweisung erfolgt für die gesamte Schule und nicht auf Ebene der einzelnen

Lehrkräfte oder Klassen. Obwohl die Schule sich also als Gesamtsystem mit bestimmten Kontextmerkmalen auseinandersetzen muss, wird in der aktuellen erziehungswissenschaftlichen Forschung ein vermehrtes Augenmerk auf die Rolle der einzelnen Lehrkraft für die Leistungsentwicklung von Schülerinnen und Schülern gelegt („Auf den Lehrer kommt es an“ - Hattie, 2018). Die Fokussierung auf Kompetenzen der einzelnen Lehrkraft könnte dazu führen, dass die Zusammenarbeit der Lehrkräfte und die gemeinsame Verantwortung des Kollegiums der Schule aus dem Blick gerät. Dadurch verliert die Schulebene möglicherweise etwas an Bedeutung, obwohl Entwicklungen wie die hin zur Outputsteuerung und Maßnahmen der Schulentwicklung die Relevanz der Einzelschule praktisch aktuell macht.

Insgesamt kann festgehalten werden, dass in der Forschung die Relevanz der Einzelschule bzw. die Schule als Gesamtsystem mit ihren spezifischen Kontextmerkmalen wenig wahrgenommen wird, wobei vieles dafürspricht, sie zu beachten. Dies zeigen auch die Ergebnisse zu Unterschieden der Unterrichtsqualität zwischen den Schulen (Studie 1). Darüber hinaus ist die Beachtung der Einzelschule auch aus praktischer Perspektive höchstrelevant, da Schulen politisch als Gesamtsysteme behandelt werden, beispielsweise durch die Schulinspektion. Dies bedeutet jedoch nicht, dass Merkmale schulischer Lernumwelten stets nur auf Schulebene aggregiert betrachtet werden sollten. Es ist nach wie vor sinnvoll, Schulen auf mehreren Ebenen zu denken und auch zu analysieren: Beispielsweise Unterrichtsqualität als Gesamtmerkmal der Schule *und* Unterschiede auf Klassenebene zu berücksichtigen.

5.2.2 Die Bedeutung des Schulkontexts für Unterrichts-/Schulqualität sowie Schulleistung

Einen weiteren Erkenntnisgewinn leistet diese Dissertation hinsichtlich der Bedeutung des Schulkontexts (im Sinne von Komposition und Intervention) für Prozessmerkmale und Output von Schulen. Als Bezugspunkt diente dabei das Kontext-Input-Prozess-Output-Modell aus der Schuleffektivitätsforschung (Scheerens, 1990), dass um Aspekte der Lehr-Lernforschung (Kunter & Voss, 2011) und der Schulqualitätsforschung (Ditton, 2000; Scheerens, 1990) erweitert wurde und in dem die drei Studien der Dissertation eingeordnet wurden. Diese

Dissertation trägt in zweierlei Hinsicht dazu bei, etwas über die Bedeutung des Schulkontexts von Einzelschulen zu erfahren: Erstens wird das Zusammenspiel unterschiedlicher Qualitätsfaktoren von Schule umfassend analysiert, indem *mehrere* mögliche Verbindungen (Kontext-Prozess und Kontext-Output) mit *mehreren* Merkmalen (bspw. Kompositionsmerkmale: SES, Migrationshintergrund, Leistung) untersucht werden. Zweitens wird der Zusammenhang des Schulkontexts (Komposition und bildungspolitische Intervention) mit Unterrichts- und Schulqualität sowie mit der Schulleistung analysiert.

Zusammenspiel mehrerer Qualitätsfaktoren mit Kontext

Es existieren, wie in Kapitel 2 beschrieben, bisher zahlreiche Studien, die bestimmte Zusammenhänge zwischen Qualitätsfaktoren des Kontext-Input-Prozess-Output-Modells untersucht haben (vgl. bspw. Dumont et al., 2013; Kunter et al., 2009; Scheerens & Bosker, 1997). Dabei wurden häufig einzelne Merkmale herausgegriffen: Die Verbindung Kontext-Prozess wurde beispielsweise durch Studien zur *leistungsbezogenen* Zusammensetzung der Schülerschaft (als Kontextmerkmal) in Zusammenhang mit Unterrichtsmerkmalen (als Prozessmerkmal) untersucht (Decristan, Fauth, Kunter, Büttner & Klieme, 2017; Donaldson, LeChasseur & Mayer, 2017; Dreeben & Barr, 1988). Dabei wurde also lediglich *ein Merkmal* (leistungsbezogene Zusammensetzung der Schülerschaft) ausgewählt und *eine Verbindung* (Kontext-Prozess) fokussiert. Als Beispiele, in denen mehrere Verbindungen oder mehrere Merkmale analysiert wurden, können die folgenden genannt werden: Liu, Van Damme, Gielen & Van Den Noortgate, 2015; Opdenakker & Van Damme, 2001; Rjosk et al., 2014. Zu diesem Forschungsdesiderat der umfassenderen Analyse mit *mehreren* untersuchten Verbindungen oder *mehreren* Merkmalen schulischer Qualität können Studie 2 und 3 dieser Dissertation einen Beitrag leisten: In Studie 2 wird die Verbindung zwischen Kontext und Prozess durch mehrere Schülerkompositionsmerkmale (Kontext) mit mehreren Merkmalen der Unterrichts- und Schulqualität (Prozess) sowie zusätzlich Mediationseffekte dieser für die Schulleistung (Output) betrachtet (Verbindung Kontext-Prozess-Output). In Studie 3 wird der Zusammenhang zwischen Kontext und Prozess durch die Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ (Kontext) mit mehreren Merkmalen der Unterrichts- und

Schulqualität (Prozess) ein bis drei Jahre später untersucht. Zudem wird die Verbindung Kontext-Output durch die Folgen der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ mit den Leistungen (Output) sowie mit der Komposition der Schülerschaft (Output, bzw. neuer Kontext der Schule zu einem späteren Zeitpunkt) analysiert (siehe Abb. 1). Damit wird angestrebt, möglichst umfassend Zusammenhänge zwischen schulischen Kontextmerkmalen und Prozessmerkmalen sowie dem Output zu untersuchen. Es zeigt sich in der übergreifenden Betrachtung, dass sich durch die Berücksichtigung verschiedener Faktoren durchaus Unterschiede zwischen den Zusammenhängen mit Prozessmerkmalen und dem Output zeigen. Die Berücksichtigung mehrerer Kompositionsmerkmale in Studie 2 zeigt beispielsweise, dass sich diese in ihren Zusammenhängen mit den Prozessmerkmalen unterscheiden: Beispielsweise korreliert der SES der Schülerschaft signifikant positiv und der Migrationshintergrund negativ mit den Unterrichtsqualitätsmerkmalen (Prozess). Die mittlere Schulleistung korreliert positiv und die Leistungsheterogenität korreliert so gut wie gar nicht mit den Unterrichtsqualitätsmerkmalen (Prozess). Bei der Betrachtung mehrerer Prozessmerkmale in Studie 2 (Merkmale der Unterrichts- und Schulqualität), zeigt sich der bereits beschriebene Befund, dass ein Zusammenhang zwischen Komposition und Unterrichtsqualität vorhanden ist, jedoch kaum ein Zusammenhang mit der Schulqualität besteht. Die Berücksichtigung mehrerer Zusammenhänge in Studie 3 (Kontext-Prozess und Kontext-Output) zeigt, dass leicht positive Effekte des Kontextmerkmals Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ auf Prozessmerkmale (Unterrichtsqualität ebenfalls mehr als Schulqualität) zu verzeichnen sind, sich wiederum kein Effekt der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ auf den Output (Schulleistung) zeigt. Tendenziell negative Effekte dieses Kontexts zeigen sich für ein besonderes Outputmerkmal, das in Studie 3 noch zusätzlich betrachtet wurde: die Entwicklung der Schülerkomposition. Der Anteil der Schülerinnen und Schüler mit Migrationshintergrund erhöht sich hierbei in den Jahren nach der Schulinspektionsdiagnose, was darauf hindeutet, dass Familien ohne Migrationshintergrund die Schulen mit erheblichem Entwicklungsbedarf eher meiden. Die Komposition der Schülerschaft (hier als Outputmerkmal) kann gleichzeitig als neues Kontextmerkmal der Schule interpretiert werden. Zunächst kann an

dieser Stelle festgehalten werden, dass die Berücksichtigung mehrerer Faktoren und Zusammenhänge ein umfassenderes Bild des Gesamtkomplexes Schule ermöglicht.

An dieser Stelle soll auf zirkuläre Bezüge zwischen Output- und Kontextmerkmalen in Studie 2 und 3 eingegangen werden. In Studie 3 wurde die Entwicklung der Schülerkomposition als abhängige Variable betrachtet. Das heißt, es wurde analysiert, ob sich die Zusammensetzung der Schülerschaft durch die Schulinspektionsdiagnose „Schule mit erheblichem Entwicklungsbedarf" ändert. Die Zusammensetzung der Schülerschaft wurde hier also als *Output* betrachtet, wobei ebendiese Zusammensetzung gleichzeitig ein Merkmal des neuen *Kontexts* der Schule ein paar Jahre später darstellt. Auf diese Weise verbindet das Modell in Anlehnung an die Schuleffektivitätsforschung (Scheerens, 1990) auch Elemente der Schulentwicklung, da der Output von heute, den Kontext der Zukunft bedingt. Ein Außen-vor-Lassen der Entwicklungsprozesse von Schulen wird häufig an der Schuleffektivitätsforschung kritisiert (vgl. bspw. die zusammenfassende Beschreibung von Bischof, 2014). Dennoch gibt es einige Versuche, wie beispielsweise durch das Modell von Creemers und Kyriakides (2006), das „Dynamic model of educational effectiveness", welches auf Effektivität als dynamischen Prozess und auf die Veränderbarkeit der schulischen Faktoren sowie nichtlinearen Beziehungen zwischen den verschiedenen Faktoren der Schule fokussiert. Es geht jedoch weniger auf zirkuläre Bezüge zwischen Output und Kontext ein, wie sie hier beschrieben wurden. Es kann festgehalten werden, dass hier einzelne Elemente der Schulentwicklung mit dem Kontext-Input-Prozess-Output-Modell der Schuleffektivitätsforschung (Scheerens, 1990) verknüpft wurden, eine umfassende Verknüpfung der beiden theoretischen Forschungstraditionen jedoch sehr komplex ist und ein weiteres Forschungsdesiderat darstellt (vgl. zur theoretischen und empirischen Verknüpfung Bischof, 2014).

Einfluss des Schulkontexts im Sinne von Komposition und Intervention

Der in dieser Dissertation thematisierte Zusammenhang des Schulkontexts mit den Prozessmerkmalen und dem Output kann Erkenntnisse zur Variabilität schulischer Lernumwelten beitragen. Der Kontext ist hierbei unterteilt in Komposition und Intervention. Dabei ist Komposition ein typischer, viel untersuchter Kontextfaktor, während die Intervention

(bspw. durch bildungspolitische Maßnahmen wie die Schulinspektion) zumindest in Verbindung mit dem Modell der Schuleffektivitätsforschung (Scheerens, 1990) kaum untersucht wurde. Wie bereits betont, wurden Kompositionseffekte bisher häufig in Zusammenhang mit Leistungen von Schülerinnen und Schülern ausgewertet (Verbindung Kontext-Output, vgl. bspw. Dumont et al., 2013; Mickelson, Bottia & Lambert, 2013; van Ewijk & Sleegers, 2010). Was bisher sehr wenig im Fokus von empirischen Studien stand, ist Komposition in Verbindung mit Prozessmerkmalen (Verbindung Kontext-Prozess, siehe Literaturübersicht in Appendix Studie 2). Zum anderen Teil, der Intervention durch Maßnahmen der Bildungspolitik zur Schulentwicklung, gibt es einige Untersuchungen zu Effekten von Schulinspektion allgemein (vgl. bspw. Brimblecombe et al., 1996; Ehren & Visscher, 2008; Gärtner et al., 2009; siehe Übersicht in Studie 2), während die besondere Situation der Schulen, die als entwicklungsbedürftig diagnostiziert werden, bisher jedoch international und national sehr wenig fokussiert wurde.

Zur Bedeutung des Kontexts kann bezüglich der Komposition der Schülerschaft als Erkenntnis dieser Dissertation festgehalten werden, dass neben Zusammenhängen mit dem Output (Schulleistung) auch signifikante Zusammenhänge verschiedener Kompositionsmerkmale (insbesondere SES und Leistung) mit Prozessmerkmalen (Unterrichtsqualitätsmerkmale Klassenführung und kognitive Aktivierung) nachgewiesen werden konnten. Die Mediation des Zusammenhangs von Komposition und Output (Schulleistung) durch die Prozessmerkmale, konnte jedoch nur bedingt gezeigt werden. Mögliche Gründe dafür wurden in Studie 2 diskutiert.

Zur Verbindung von Kontext im Sinne der Intervention durch die Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ mit Prozessmerkmalen sowie Outputs (Studie 3) kann folgendes festgehalten werden: Es zeigen sich durch die Intervention geringe Veränderungen für die Prozessmerkmale (Unterrichts- und Schulqualität) und keine Effekte für die Leistungen dieser Gruppe von Schulen. Es lassen sich jedoch nichtintendierte Nebeneffekte bezüglich der Entwicklung der Schülerschaft identifizieren: Der Anteil der Schülerinnen und Schüler mit Migrationshintergrund stieg infolge der Diagnose „erheblicher Entwicklungsbedarf“ signifikant an. Es kann angenommen werden, dass Familien ohne

Migrationshintergrund diese Schulen aufgrund der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ meiden und es so zu dieser Erhöhung kommt. Betrachtet man diesen Befund zusammen mit den Ergebnissen aus Studie 2, deutet sich eine Art „Abwärtsspirale“ an: An Schulen mit einer bestimmten Schülerschaft (niedriger SES, geringe Leistung) zeigt sich eine geringere Unterrichtsqualität (Studie 2), welche dann von der Schulinspektion festgestellt wird, was wiederum zu einer Diagnose der Schule als erheblich entwicklungsbedürftig führt. Diese Diagnose kann in der Folge zu einer Erhöhung des Anteils an Schülerinnen und Schülern mit Migrationshintergrund (Studie 3) und somit zu einer Verstärkung der Segregation an Schulen führen.

Insgesamt kann bezüglich der Bedeutung des Kontexts von Schulen resümiert werden, dass dieser zur Variabilität zwischen Schulen beiträgt. Es kommen also Unterschiede zwischen Schulen unter anderem durch ihre Kontextmerkmale zustande und daher sollte der Kontext der Schulen stärker bei der Analyse schulischer Prozessmerkmale und Outputs berücksichtigt werden (vgl. auch die Diskussion zur Kontextberücksichtigung in der Schuleffektivitätsforschung bei Bischof, 2014; Thrupp & Lupton, 2006). Diese Forderung nach der Berücksichtigung schulischer Kontextmerkmale in der Forschung zu Schuleffektivität und Schulentwicklung sowie für die Praxis und Politik, beispielsweise für die Schulinspektion, lässt sich auf Basis dieser Ergebnisse bestätigen.

5.3 Limitationen der Arbeit & Ausblick

In diesem Kapitel werden Limitationen dieser Arbeit dargestellt und es wird ein möglicher Ausblick gegeben.

Outputmerkmale

Als große Einschränkung dieser Dissertation soll zunächst genannt werden, dass bei Betrachtung des Outputs lediglich auf Leistungen fokussiert wurde. Neben Leistungen von Schülerinnen und Schülern existieren jedoch noch zahlreiche andere wichtige Outputmerkmale, die durch schulisches Lernen beeinflusst werden können. Dazu gehören beispielsweise sozio-emotionale Merkmale und die Persönlichkeitsentwicklung. Dies umfasst Merkmale, die

relevant sind für das Erreichen von Zielen, Regulieren von Emotionen und die Zusammenarbeit mit anderen Menschen (John & De Fruyt, 2015). Diese sind zentral für den Erfolg in Schule und Beruf (vgl. bspw. Durlak, Weissberg, Dymnicki, Taylor & Schellinger, 2011; National Research Council, 2012) und sollten daher neben Leistungen und kognitiven Fähigkeiten in zukünftiger Forschung als Outputmerkmale berücksichtigt werden. Zudem können Ergebnisse zu Zusammenhängen zwischen Prozessmerkmalen von Schulen und nicht-kognitiven Outputs (Brand, Felner, Shim, Seitsinger & Dumas, 2003; Davis, 2003; Niemiec & Ryan, 2009) weitere Ansatzpunkte für Schulentwicklungsprozesse bieten. In den USA wurde das Accountability-System diesbezüglich bereits angepasst. Alle Staaten haben bereits Standards für den Bereich der frühen Kindheit (Kindergarten) festgelegt, für den Schulbereich (K-12) bestehen noch größere Unterschiede im Fortschritt sowie im Inhalt der Entwicklung der Standards für nicht-kognitive Kompetenzen. Präzise Messungen der nicht-kognitiven Kompetenzen bedürfen jedoch noch weiterer Forschung (im Überblick siehe Eklund, Kilpatrick, Kilgus & Haider, 2018).

Fehlender Input

Eine weitere Limitation, die diskutiert werden muss, betrifft die Vernachlässigung von Inputmerkmalen in dem verwendeten Modell dieser Dissertation und den dazugehörigen Studien. Vor allem Merkmale der Lehrkräfte der Schule werden häufig als Input betrachtet (OECD, 2005; Scheerens, 1990; Scheerens & Bosker, 1997) und sind auch relevant für die innerschulisch stattfindenden Prozesse (Kunter et al., 2011). Dazu zählen beispielsweise Merkmale wie die Ausbildung, das Alter, Fachwissen, Motivation etc. der Lehrkräfte. Auf Basis der Ergebnisse, dass beispielsweise Motivation und fachdidaktische Kompetenz von Lehrkräften für Unterrichtsqualität ausschlaggebend sind (Kunter et al., 2009), kann vermutet werden, dass diese auch eine relevante Rolle für Variabilität zwischen Schulen in den Prozess- sowie Outputmerkmalen innehaben. In zukünftigen Studien wäre es daher interessant, detaillierter zu erfahren, ob bestimmte Merkmale der Lehrkräfte mit dem Umgang bestimmter Kontextmerkmale in Zusammenhang stehen: Kann also eine Lehrkraft mit stark ausgeprägter Motivation (Input) beispielsweise besser mit großer Heterogenität der Schülerschaft (Kontext)

umgehen und eine bessere Unterrichtsqualität (Prozess) gewährleisten als eine Lehrkraft mit geringer Motivation? In diesem Zusammenhang verweisen Vock und Gronostaj (2017) bezüglich des Unterrichts mit heterogenen Klassen auf die besondere Bedeutung folgender Eigenschaften der Lehrkräfte (Input): Fachwissen, diagnostische Kompetenz und förderliche Einstellungen, wie beispielsweise ohne Vorurteile zu unterrichten (vgl. im Überblick Vock & Gronostaj, 2017).

Zudem ist an dieser Stelle festzuhalten, dass die Unterscheidung von Kontext- und Inputmerkmalen von verschiedenen Autoren unterschiedlich vorgenommen wird (vgl. bspw. Ditton, 2000; Helmke, 2009; Scheerens, 1990). Orientiert am Kontext-Input-Prozess-Output-Modell aus Kapitel 2, das „achievement stimulants from higher administrative levels" (Scheerens, 1990) als Kontext einordnet, ist ein bildungspolitisches Vorgehen wie die Einteilung der Schulinspektion in erheblich entwicklungsbedürftige Schulen als Kontextfaktor zu werten. In anderen Modellen könnten bildungspolitische Vorgaben und Rahmenbedingungen jedoch auch als Input gelten.

Kontextmerkmale

Bezüglich des Kontexts sei darauf verwiesen, dass die in dieser Dissertation genutzten Kontextmerkmale (Komposition in Studie 2 und Intervention in Studie 3) nur Ausschnitte des Schulkontexts beleuchten und es durchaus noch zahlreiche weitere Merkmale gibt, die in Zusammenhang mit schulischen Prozessmerkmalen untersucht werden sollten. Dazu gehören beispielsweise regionale Bedingungen bzw. Bedingungen der schulischen Infrastruktur (Ainsworth, 2002; Garner & Raudenbush, 1991; Horr, 2015), die häufig in der Soziologie und sozialwissenschaftlichen Bildungsforschung Anwendung finden. Helbig (2010) untersuchte beispielsweise mit Informationen zur Sozialstruktur der Nachbarschaften von Schülerinnen und Schülern aus Berliner Grundschulen Zusammenhänge mit ihrer Kompetenzentwicklung. Hierbei zeigten sich vor allem positive Effekte von einer „günstigen" Sozialstruktur auf die Kompetenzentwicklung aber keine negativen Effekte bei einer sozial benachteiligten Sozialstruktur, zumindest nicht zusätzlich zu Kompositionseffekten (Helbig, 2010). Erkenntnisse über regionale Bedingungen von Schulen können somit den Blick über

Komposition hinaus erweitern und dazu beitragen, Variabilität schulischer Lernumwelten zu ergründen.

Zudem ist hinsichtlich der Betrachtung einer bildungspolitischen Intervention folgendes einschränkend zu benennen: Ergebnisse der Analyse der Spezialgruppe der als entwicklungsbedürftig diagnostizierten Schulen sind nicht auf alle Schulen generalisierbar. Um die Verbindung von bildungspolitischen Rahmenbedingungen als Kontext der Schule tiefergehend analysieren zu können, bedarf es weiterer Forschung zum Zusammenhang von Bildungspolitik und schulischen Prozessmerkmalen sowie deren Vermittlung auf Outputvariablen.

Variabilität der Unterrichtsqualität innerhalb von Schulen

Obwohl viel dafür spricht, auch die Einzelschule als relevante Handlungseinheit zu betrachten, soll in diesem Zusammenhang auf eine Einschränkung der Daten, die in dieser Dissertation verwendet wurden, eingegangen werden. Unterrichtsqualität wird aufgrund der Datenlage (Studie 1 – PISA-Daten, Studie 2 und 3 – Schulinspektionsdaten) nur auf der Ebene der gesamten Schule betrachtet und nicht auf Ebene der einzelnen Lehrkräfte bzw. Klassen. Wie wiederholt angesprochen, kann auf Basis dieser Daten dennoch eine sinnvolle Aussage zur Unterrichtsqualität gemacht werden. Davon abgesehen wäre es mit einer umfassenderen Datengrundlage möglich gewesen, sowohl die Schulebene als auch die Ebene der einzelnen Lehrkräfte zu untersuchen und dies mehrebenenanalytisch zu betrachten. Eine solch umfassende Datenbasis müsste dafür Daten zur Unterrichtsqualität sowie zur Zusammensetzung der Schülerschaft auf Klassenebene beinhalten, wie dies beispielsweise bei der DESI-Studie (Deutsch Englisch Schülerleistungen International) oder ausgewählten Startkohorten des Nationalen Bildungspanels (NEPS) der Fall ist. Für Studie 1 hätte mit einer solchen umfassenden Datenbasis der Varianzanteil, der auf den verschiedenen Ebenen aufgeklärt werden kann, berechnet werden können (wie bspw. bei Wurster & Feldhoff, 2019). Im Kontext von Studie 2 ist insbesondere ein unterschiedliches Vorgehen von Lehrkräften im Umgang mit der jeweiligen Schülerkomposition denkbar und daher wären Angaben zur Schülerkomposition auf Individual- bzw. Klassenebene hier hilfreich gewesen. Damit könnte

dem in Kapitel 5.2 formulierten Anspruch, der Schulebene zusätzlich zur Klassenebene Bedeutung beizumessen, besser nachgekommen werden. Andere Studien, die sich den Themen auf der Klassenebene widmen, zeigen jedoch überwiegend ähnliche Befunde im Vergleich zu denen dieser Dissertation, in Bezug auf Zusammenhänge zwischen Komposition und Unterrichtsqualität (vgl. bspw. Decristan et al., 2017; Rjosk et al., 2014) und in der Varianzaufklärung durch die Ebene der Schule (Wurster & Gärtner, 2013; Wurster & Feldhoff, 2019).

Stabilität Unterrichtsqualität

Als Grenze der Einordnung der Ergebnisse dieser Dissertation ist zu benennen, dass es wenige Erkenntnisse zur Stabilität schulischer Lernumwelten gibt. Die Vorstellung, dass Unterrichtsqualität zwischen Schulen variiert (oder auch innerhalb von Schulen variiert) basiert darauf, dass das Konstrukt Unterrichtsqualität eine gewisse Stabilität aufweist. Wäre dem nicht so, könnte es keine gemeinsame Unterrichtsqualität auf Schulebene geben, es würde noch nicht einmal *eine* Unterrichtsqualität einer einzelnen Lehrkraft geben. Da Unterricht immer auch ein Wechselspiel in der Interaktion mit den Schülerinnen und Schülern darstellt (siehe Angebot-Nutzungsmodell von Helmke, 2009), ist es jedoch gut vorstellbar, dass der Unterricht an einem bestimmten Tag mit einer bestimmten Lerngruppe kognitiv anregend, unterstützend und ohne Störungen verläuft, wohingegen der Unterricht der gleichen Lehrkraft an einem anderen Tag mit einer anderen Lerngruppe wenig unterstützend und weniger anregend verläuft. Dies würde für eine niedrige Stabilität des Konstrukts Unterrichtsqualität sprechen. Ergebnisse der Studie von Gärtner und Brunner (2018) unterstützen diese Vermutung. Gärtner und Brunner (2018) zeigen, dass die Messung bestimmter Merkmale der Unterrichtsqualität zwischen Klassen und Jahrgangsstufen variiert, hingegen weniger zwischen Schuljahren und unterschiedlichen Fächern variiert. Wenn Unterrichtsqualität schon bei der gleichen Lehrkraft ein Konstrukt darstellt, dass eine gewisse Instabilität aufweist, liegt nahe, dass Unterrichtsqualität dann auch an unterschiedlichen Schulen mit unterschiedlichen Schülerschaften variiert. Eine hohe Instabilität des Konstrukts Unterrichtsqualität würde dazu führen, dass die Ergebnisse dieser Dissertationsstudien verzerrt wären, da eine präzise Messung von Unterrichtsqualität mit nur

einmaliger Erhebung und spezifischen Rahmenbedingungen (bspw. Jahrgangsstufe, Schülerschaft, Fach, äußere Einflüsse etc.) nicht gewährleistet wäre. Andererseits zeigen Studien, die sich mit einer ressourceneffizienten Messung von Unterrichtsqualität mittels sehr kurzer Videoaufnahmen (ca. 30 Sekunden) beschäftigen – „Thin-Slices-Technik" –, dass hiermit durchaus eine hohe Güte der Unterrichtsqualitätsmessung (zumindest auf der Ebene der einzelnen Lehrkraft) erreicht werden kann (Begrich, Fauth, Kunter & Klieme, 2017). Dies könnte in der Art interpretiert werden, dass Unterrichtsqualität eine gewisse Stabilität aufweist, die dazu führt, dass die Güte der Messung trotz einmaliger Messung mit spezifischen Rahmenbedingungen hoch ist. Zudem ergab sich bei Wagner et al. (2016) in einem Zeitraum von drei Monaten moderate bis hohe Stabilität in Merkmalen der Unterrichtsqualität. Diese Befunde deuten wiederum auf eine gewisse Stabilität des Konstrukts Unterrichtsqualität hin. Verschiedene Studien haben sich in diesem Zusammenhang damit auseinandergesetzt, wie Unterrichtsqualität gemessen werden muss um reliabel und valide zu sein (vgl. bspw. Clausen, 2002; Praetorius, 2013). Offen bleibt, was dafür geleistet muss, um auf der Messung von Unterrichtsqualität eine weitreichende Diagnostik (wie bspw. im Rahmen der Schulinspektion) basieren zu können.

Ein anderer Aspekt der Stabilität von Unterrichts- und Schulqualität als Prozessmerkmale von Schulen betrifft die Frage, wie lange es dauert, bis sich Veränderungen in Schulen vollziehen. Wie schnell werden hier Änderungen, beispielsweise im Rahmen von Schulentwicklungsprozessen sichtbar? Erkenntnisse in diesem Bereich hätten dann auch Folgen für die Schulinspektion und die Idee einer „Nachinspektion" um Veränderungen zu evaluieren.

Schulinspektionsdaten für Bildungsforschung

Die Schulinspektionsdaten stellen eine umfangreiche und besondere Datenbasis dar, die jedoch in der Nutzung für die Bildungsforschung auch Nachteile mit sich bringt. Die Erhebung und Auswertung der Schulinspektionsdaten ist nicht auf die Nutzung für Zwecke der Bildungsforschung ausgerichtet, sodass dies auch mit Limitationen für diese Dissertation einhergeht. So ist hier zu nennen, dass die Basisdimensionen für Unterrichts- und Schulqualität (Klassenführung, kognitive Aktivierung, konstruktive Unterstützung, Schulkultur,

Schulmanagement, Kooperation und Koordination, Personal), wie sie in den Dissertationsstudien verwendet wurden, nur *eine* mögliche Variante darstellen, die in der Schulinspektion erhobenen Merkmale theoretisch einzuordnen. Da sich die Schulinspektion bzw. der zugrunde liegende Handlungsrahmen für Schulqualität (Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013) nicht aus der Lehr-Lernforschung begründet, wurde diese theoretische Einordnung im Rahmen der Dissertationsstudien vorgenommen und die Passung der erhobenen Konstrukte zu den Basisdimensionen war nicht immer eindeutig. Offen bleibt dabei, welche Dimensionen geeigneter sind um valide und vollumfänglich Unterrichtsgeschehen beschreiben zu können – die Indikatoren des Handlungsrahmens aus Schulpolitik und -verwaltung oder die theoretischen Dimensionen, wie sie beispielsweise in der COACTIV-Studie (Kunter et al., 2011) oder von Ditton (2000) erarbeitet wurden. Weiterhin bleibt unklar, inwieweit die Schulinspektion in ihrer Evaluation tatsächlich die „Tiefenstruktur" des Unterrichts messen kann (Kunter & Voss, 2011). Hierfür wären zusätzliche Analysen interessant, die die Qualitätsevaluation der Schulinspektion mit anderen Unterrichts- oder auch Schulqualitätsmessungen vergleichen, als Validierung der Schulinspektion.

Zudem muss hier einschränkend angemerkt werden, dass sich durch die Kombination der Schulinspektionsdaten mit Daten der amtlichen Statistik und Leistungsdaten aus Vergleichsarbeiten (Studien 2 und 3) kein optimaler Längsschnittdatensatz ergibt, der uneingeschränkt kausal interpretiert werden kann. Die Daten stammen aus unterschiedlichen Quellen mit unterschiedlichem Zweck und beziehen sich nicht exakt auf die gleiche Personengruppe: Es wurden Leistungsdaten des dritten Jahrgangs der Berliner Grundschulen aus den Vergleichsarbeiten (VERA 3) verwendet, während sich die Daten der Unterrichts- und Schulqualität aus der Schulinspektion sowie die Daten der amtlichen Statistik auf die gesamte Schule beziehen. Dennoch kann durch die zeitliche Abfolge, wie sie in Studie 2 und 3 beschrieben ist, von einer temporalen Ordnung ausgegangen werden, die sich nicht nur korrelativ interpretieren lässt.

Zur Güte von Schulinspektionsdaten zeigten wiederum einige Untersuchungen, dass die Übereinstimmung der Beobachtungen des Unterrichts durch externe Evaluatoren und Evaluatorinnen der Schulinspektion überwiegend als ausreichend beurteilt werden (Müller &

Pietsch, 2011; Pietsch & Tosana, 2008) und auch die Reliabilität zeigt sich überwiegend als angemessen (Wurster & Gärtner, 2013). Ein grundsätzlicher Vorteil der Schulinspektionsdaten liegt in der großen Breite der Daten, da sie eine Vollerhebung aller öffentlichen Schulen des Landes darstellen und zahlreiche Merkmale des Unterrichts und der Schule beinhalten. In einer gemeinsamen Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DgfE), der Gesellschaft für Empirische Bildungsforschung (GEBF) und der Gesellschaft für Fachdidaktik (GFD) zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten, wird daher auch darauf verwiesen, den Zugang zu Schulinspektionsberichten für wissenschaftliche Zwecke zu ermöglichen und die Verknüpfung dieser Daten mit anderen Forschungsdaten zu erleichtern (DGfE, GEBF, GFD, 2020). Auch das Potenzial, Daten der amtlichen Statistik für die Bildungsforschung zu nutzen, zeigt sich beispielsweise durch die Zuverlässigkeit der Daten und den Zugang zu einer Großzahl an Daten, die über einzelne Studien kaum erreicht werden kann (vgl. auch Artelt, Bug, Kleinert, Maaz & Runge, 2019; Fickermann & Weishaupt, 2019).

Wurster und Gärtner (2013) beschreiben weiterhin die Möglichkeit der Nutzung von Schulinspektionsdaten auch in Kombination mit anderen Daten, unter anderem für die Untersuchung von Zusammenhängen zwischen Kontextmerkmalen und Unterrichtsqualität sowie Fragen der Schuleffektivitätsforschung, also Zusammenhängen von Unterricht und Schule (Prozessmerkmalen) und Outputs, so wie dies in den Studien 2 und 3 dieser Dissertation erfolgte.

Generalisierbarkeit

Inwieweit sind die Befunde der Dissertationsstudie auf die deutsche Schullandschaft generalisierbar? Die Ergebnisse der Studie 1 sind aufgrund der internationalen, metaanalytischen Perspektive auch auf andere Länder und Schulbedingungen übertragbar. Da ausschließlich die 15-Jährigen Grundlage der Daten sind, wäre es jedoch zudem interessant, vergleichende Daten aus der Primarstufe nutzen zu können. Andere Studien deuten darauf hin, dass bereits Schülerinnen und Schüler in der Grundschule zuverlässige, hilfreiche Einschätzungen des Unterrichts geben können (Fauth et al., 2014; Gärtner, 2010), eine

systematische Überprüfung der Unterschiede zwischen Grundschulen steht jedoch noch aus. Hierfür könnten beispielsweise Daten der IGLU/PIRLS-Studie (Internationale Grunschul-Lese-Untersuchung/Progress in International Reading Literacy, für Ergebnisse zur Unterrichtsqualität siehe bspw. Hußmann et al., 2017) genutzt werden.

Die Ergebnisse von Studie 2 und 3 beziehen sich hingegen ausschließlich auf Grundschulen in Berlin, sind dabei aber nahezu als Vollerhebung zu betrachten. Auf andere Bundesländer (oder andere Länder international) sind die Ergebnisse daher ohne weitere Überprüfung nur begrenzt übertragbar. In den Studien 2 und 3 wurden dazu jedoch teils auch internationale Ergebnisse beschrieben, die die eigenen Ergebnisse unterstreichen, sodass hier eher davon ausgegangen werden kann, dass diese nicht nur für Berliner Grundschulen gelten. Dennoch wären weitere Studien, die die vorliegenden Befunde für andere Bundesländer replizieren und erweitern für eine übergreifende Einordnung der Ergebnisse hilfreich.

5.4 Implikationen

Auf Grundlage der Studien dieser Dissertation können verschiedene Implikationen im Hinblick auf die Variabilität schulischer Lernumwelten abgeleitet werden. Die Implikationen zu inhaltlichen und methodischen Aspekten betreffen die Bereiche Forschung, Politik und (Schul-)Praxis und sollen im Folgenden erläutert werden.

5.4.1 Für die Forschung

Berücksichtigung der Schule als relevante Analyseeinheit

Als ein wichtiges Ergebnis dieser Dissertation ist zu nennen, dass die Relevanz der Schule als Analyseeinheit durch die empirischen Studien bestätigt werden konnte. In aktuellen Forschungsthemen wird der Blick häufig auf Schulformen (Baumert et al., 2006b), einzelne Lehrkräfte, die Klassenebene und den Unterricht im Speziellen gelenkt (vgl. bspw. Hattie, 2018; Lipowsky, 2006). Dies sind relevante Faktoren, die zur Leistungs- und Persönlichkeitsentwicklung von Schülerinnen und Schülern beitragen und die auch in zukünftiger Forschung nicht vernachlässigt werden sollten. Dennoch ist auch die Schule als Gesamtkomplex ein relevanter Faktor (Fend, 1986; OECD, 2005), der in der Forschung, gerade

auch in Kombination mit den anderen genannten Analyseebenen genauer beachtet werden sollte. Auch durch die Relevanz, die die Einzelschule in der Praxis erfährt, beispielsweise durch die Schulinspektion und die Evaluation schulischer Entwicklung, finden Überlegungen zur Verbesserung von Schul- und Unterrichtsqualität, aber auch von Leistungen der Schülerinnen und Schüler auf Ebene der gesamten Schule statt (Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013). Durch eine starke Orientierung der aktuellen Forschungsthemen, hin zur Klassenebene oder zur Ebene der einzelnen Lehrkraft („Auf den Lehrer kommt es an" – Hattie, 2018), gerät die Bedeutung der Schule als Analyseeinheit möglicherweise etwas aus dem Blick. Es ist weitere Forschung dazu nötig, inwieweit durch Kontexteinflüsse (durch Komposition und Intervention) Variation in der Unterrichtsqualität innerhalb der Einzelschulen besteht, also ob bezüglich der Reaktion auf Kontextmerkmale die Klassenebene oder die Schulebene bedeutsamer ist. Beispielweise könnte dazu untersucht werden, durch welche Merkmale sich die Höhe der Varianz der Unterrichtsqualität innerhalb der Schule vorhersagen lässt: Welche Merkmale bedingen also, dass sich Schulen in ihrer Unterrichtsqualität ähnlicher sind als andere Schulen?

Studie 1 dieser Dissertation zeigt die Unterschiede, die hinsichtlich Unterrichtsqualität zwischen Schulen bestehen. Die Studienergebnisse verweisen darauf, dass es durchaus Unterrichtsmerkmale gibt, die durch „etwas Gemeinsames" der Schule geprägt werden wohingegen es andere Merkmale gibt, die stärker individuell beurteilt werden – dies kann bei der Erhebung von Unterrichtsqualität in zukünftigen Studien berücksichtigt werden. Mögliche Implikationen für die Bestimmung von Stichprobengrößen bei cluster-randomisierten Stichproben, die auf Schulebene in Studien zu Unterrichtsqualität gezogen werden, wurden bereits in Studie 1 diskutiert (vgl. auch Kapitel 5.2.1). Auch hinsichtlich der Unterschiede von Schulqualität, wäre es hilfreich, einen metaanalytischen Überblick über die Unterschiede zwischen Schulen zu erlangen. Denkbar wäre auch hier, in Anlehnung an das Vorgehen in Studie 1, zu berücksichtigen, wie stark die Einschätzungen der Schulqualität von verschiedenen Lehrkräften der Schule, oder Schülerinnen und Schülern der Schule oder externen beurteilenden Personen übereinstimmen.

Kontext-Input-Prozess-Output-Modell

Welche Implikationen lassen sich für die Schuleffektivitätsforschung ziehen? Eines der „Basis-Modelle" der Schuleffektivitätsforschung, das Kontext-Input-Prozess-Output-Modell (Scheerens, 1990) begründet sich zwar empirisch (Scheerens & Bosker, 1997), es wurde aber seitdem nicht systematisch überprüft. Obwohl diese Dissertation hier einen Beitrag leisten kann, wäre eine übergreifende systematische Überprüfung der verschiedenen Verbindungen, förderlich um schulische Prozesse besser zu verstehen.

Diese Dissertation verweist darauf, dass sich schulische Lernumwelten auch aufgrund ihrer jeweiligen Kontextbedingungen (Komposition oder Intervention) unterscheiden. Dies betrifft also die Verbindung Kontext-Prozess-Output. Da die Verbindung der Prozessmerkmale (Unterrichts- und Schulqualität) mit Kompositionsmerkmalen bisher kaum explizit untersucht wurde (siehe Studie 2, Appendix für einen Literaturüberblick), ist es relevant, dies in den Fokus zukünftiger Forschung zu stellen. Insbesondere hinsichtlich des Aspekts, dass Schülerinnen und Schüler, unabhängig von ihrem sozialen oder Migrationshintergrund, qualitativ hochwertige Bedingungen in der Schule zur Verfügung gestellt bekommen sollen, ist diese Verbindung (Kontext durch Komposition-Prozess) relevant und sollte auch in zukünftigen Studien weiter untersucht werden. Diesbezügliche Untersuchungen würden zudem dazu beitragen, die bisher eher unbekannten Mechanismen zu ergründen, die hinter Kompositionseffekten auf Leistungen durch die *black box* Schule stecken (Dumont et al., 2013; van Ewijk & Sleegers, 2010). Diesbezüglich konnte zwar Studie 2 einen Beitrag leisten, jedoch ergeben sich hieraus auch weitere Forschungsfragen, beispielsweise zur Mediation der Kompositionseffekte durch Unterrichts- und Schulqualität. Hier konnte nur ein geringer Mediationseffekt nachgewiesen werden, wobei dazu diskutiert wurde, dass dies auch methodische Gründe haben könnte. Es wäre sehr hilfreich, weitere Studien durchzuführen, die sich dieser Mediation widmen, auch um die Frage zu klären, ob die in dieser Dissertation verwendeten, verschiedenen Datenquellen (zu Unterrichts- und Schulqualität, Leistungen auf Schulebene) zu unpräzise in der Messung waren oder inwieweit Schulen eine bedeutsame Rolle bei der Verringerung von Kompositionseffekten auf Leistungen spielen können.

5.4.2 Für die Bildungspolitik

Steuerung zu Fairness und hoher Qualität

In der Schulverwaltung liegt der Fokus auf der Ebene der Einzelschule – was dazu führt, dass sich eine praktische Bedeutung dieser Ebene für die Schulen ergibt. Diese Dissertation stützt den Gedanken, dass dies eine relevante Analyseebene ist, beispielsweise dadurch, dass sich Unterschiede in der Unterrichtsqualität zwischen den Schulen zeigen (Studie 1) und dass der Schulkontext in Zusammenhang mit schulischen Prozessmerkmalen (Unterrichts- und Schulqualität) sowie dem Output (Schulleistung) steht (Studie 2 und Studie 3). Für die Bildungspolitik sollte nun insbesondere eine relevante Frage sein, wie die Relevanz des schulspezifischen Kontexts (hierbei im Sinne von Komposition) gesteuert werden kann. Es sollte im Interesse der Bildungspolitik sein, den Zusammenhang zwischen Kompositionsmerkmalen der Schülerschaft und Prozessmerkmalen (Ergebnis aus Studie 2) zu reduzieren um faire Lern- und Entwicklungsmöglichkeiten für alle Schülerinnen und Schüler zu gewährleisten. Auch wenn Studie 2 nur eine geringe Mediation des Zusammenhangs zwischen Kontext (im Sinne von Komposition) und der Schulleistung durch die Unterrichts- und Schulqualität nachweisen konnte, ist die praktische Relevanz dieser Prozessmerkmale für die Bildungspolitik groß: Denn auf dieser Ebene der Prozessmerkmale können bildungspolitische Steuerungsmechanismen (zumindest teilweise) Einfluss nehmen, nicht jedoch beim individuellen Hintergrund von Schülerinnen und Schülern. Die Variabilität der Prozessmerkmale (insbesondere Unterrichtsqualität) stellt sich folglich problematisch dar. Eine hohe Qualität bei geringer Variabilität wäre wünschenswert um faire Bedingungen für Schülerinnen und Schüler unabhängig ihres sozialen Hintergrunds oder Migrationshintergrunds zu schaffen. Um die Unterrichtsqualität unabhängig von ihrer Komposition der Schülerschaft zu erhöhen, sind aus bildungspolitischer Sicht verschiedene Aspekte denkbar: Beispielsweise könnte die Evaluation des Unterrichts in Schulen durch die Schulinspektion um die kontinuierliche Selbstevaluation ergänzt werden. Schülerinnen und Schüler (selbst in der Grundschule) können dabei hilfreiche und messgenaue Urteile der Unterrichtsqualität abgeben (vgl. bspw. Fauth et al., 2014). Die Evaluation bestehender Selbstevaluationsprojekte an Schulen, wie beispielsweise die der Nutzung des „Selbstevaluationsportals (SEP)“ in Berlin

und Brandenburg zeigen dabei die Bedeutung von Selbstevaluation für die Motivation der Lehrkräfte und für Veränderungen des Unterrichts durch Lehrkräfte (Gärtner & Vogt, 2013; Gärtner, 2014). Ein ähnliches Projekt stellt „EMU Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung“ dar, das im Auftrag der deutschen Kultusministerkonferenz entwickelt wurde (Helmke & Lenske, 2013). Durch die Ergänzung der externen Evaluation um die Selbstevaluation (interne Evaluation) wäre eine engmaschigere Überprüfung als Vorrausetzung zur Weiterentwicklung des Unterrichts sowie eine andere Form der Auseinandersetzung der Lehrkräfte mit ihrem Unterricht gegeben und die Erhöhung von Unterrichtsqualität an allen Schulen kann gefördert werden. Aus dieser Perspektive erscheint eine Verpflichtung zur Selbstevaluation hilfreich, so wie dies in Berlin der Fall ist (Abgeordnetenhaus Berlin, 2004, 2011). Zudem gibt es verschiedene mögliche Interventionen zur Verbesserung von Unterrichtsqualität, die sich als erfolgreich erwiesen haben. Die Nutzung dieser könnte durch die Bildungspolitik stärker initiiert und im Rahmen von Fortbildungen an Lehrkräfte weitergegeben werden (siehe auch Implikationen für die Lehrerbildung, Kapitel 5.4.3). Dazu gehören Interventionen zur Verbesserung der effizienten Klassenführung, die relevante Veränderungen im Wissen der Lehrkräfte über Klassenmanagement sowie im Verhalten der Lehrkräfte initiieren können (Piwowar, Thiel & Ophardt, 2013; Thiel, Ophardt & Piwowar, 2013). Zudem zählen dazu beispielsweise Materialien, die aus dem „SINUS“-Projekt entstanden, dass der Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts diente (vgl. bspw. Baptist & Raab, 2007; Dalehefte et al., 2014).

„Schwierige Schulen“ („failing schools“) unterstützen

Wie können Schulen mit „schwierigen“ Schülerschaften (also eines bestimmten Kontexts im Sinne einer Schülerkomposition, die sich vor allem durch niedriges SES und eine geringe mittlere Leistung auszeichnet) durch die Bildungspolitik unterstützt werden (vgl. im Überblick Manitius & Dobbelstein, 2017)? Diese Dissertation hat gezeigt, dass es möglicherweise eine Art Abwärtsspirale bei Schulen mit niedrigem SES und geringer mittlerer Leistung gibt: Ungünstige Kompositionsmerkmale hängen mit geringer ausgeprägter Unterrichts- und

Schulqualität zusammen (Studie 2), bei geringer Unterrichts- und Schulqualität erfolgt eine Einordnung als „schwierige“ Schule (in Berlin „Schule mit erheblichem Entwicklungsbedarf“) und dies trägt möglicherweise zu einer weiteren Segregation bei, da sich bestimmte Eltern (ohne Migrationshintergrund) bei der Schulwahl der darauffolgenden Jahre tendenziell gegen diese Schulen entscheiden (Studie 3). Daraus lassen sich Implikationen ableiten, die verschiedene Aspekte betreffen: (1) Verhindern/Verändern einer ungünstigen Komposition sowie die (2) Berücksichtigung und Umgang mit schwieriger Komposition der Schülerschaft.

(1) Inwiefern kann und sollte die Bildungspolitik Möglichkeiten der Regulierung der Zusammensetzung der Schülerschaft nutzen? Zu bedenken sind dabei Reaktionen der Familien, die sich einer „verpflichtenden“ Zuweisung zu einer bestimmten Schule durch andere Wege entziehen und Segregation dadurch noch stärker fördern (vgl. bspw. Helbig, Nikolai & Wrase, 2017; Riedel, Schneider, Schuchart & Weishaupt, 2010). Andererseits gilt es auch zu bedenken, welche Informationen von Seiten der Bildungspolitik für die Schulwahl zur Verfügung gestellt werden sollten. In Berlin werden beispielsweise Informationen und Eckdaten zu allen öffentlichen Schulen in einem Schulportrait online veröffentlicht (Senatsverwaltung für Bildung, Jugend und Familie: Berliner Schulen). Dazu gehören auch Angaben zum Anteil von Schülerinnen und Schülern mit Migrationshintergrund pro Schule sowie die verpflichtend öffentlich gestellten Berichte der Schulinspektion. Von verschiedenen Autoren wird diskutiert, dass diese Veröffentlichung möglicherweise zu einer verstärkten Segregation führen kann (Helbig & Nikolai, 2017; Jurczok & Lauterbach, 2014; van Ackeren, 2003).

(2) Wie kann die Komposition der Schülerschaft berücksichtigt werden? Dies betrifft die Frage, wie „failing schools“ mit schwierigen Bedingungen aufgrund ihrer Komposition der Schülerschaft angemessen unterstützt werden können. Obwohl verschiedene Faktoren schulischer Qualität identifiziert wurden, die wesentlich für die Entwicklung von „failing schools“ sind (vgl. bspw. Muijs, Harris, Chapman, Stoll & Russ, 2004), offenbart sich, dass es keine allgemeingültige Lösung für die Unterstützung von „failing schools“ zu geben scheint. Zu den häufig genutzten „Turnaround“-Methoden für US-amerikanische Schulen zählen laut Meyers und Murphy (2007) die folgenden: eine eigenständige

Schulentwicklungsplanung durch die Schule, Unterstützung der Schule durch Expertinnen und Experten, die Möglichkeit für Schülerinnen und Schüler, zusätzliche Nachhilfe in Anspruch zu nehmen, die Adoption eines umfassenden Schulmodells sowie die Rekonstitution der Schule.

Eine weitere Möglichkeit der Unterstützung der „failing schools“ könnte durch das Bereitstellen von zusätzlichen Ressourcen erfolgen. In Berlin wurde mit dem Bonusprogramm (Senatsverwaltung für Bildung, Jugend und Familie: Bonusprogramm) eine umfangreiche finanzielle Unterstützung von Schulen mit schwierigen Bedingungen durchgeführt, bei der alle Schulen mit einem Anteil von über 50 % der Schülerinnen und Schüler, die von der Zuzahlung der Lernmittel befreit sind, berechtigt wurden, Fördermittel zu erhalten (Senatsverwaltung für Bildung, Jugend und Familie: Bonusprogramm). Die Schulleitung kann diese Mittel hauptsächlich selbst verantworten. Bei der Evaluation des Bonusprogramms zeigten sich teils deutliche Verbesserungen „in den Bereichen des Schulklimas, des Schülerverhaltens (z.B. aggressives Verhalten/Verhaltensauffälligkeiten), der Außenwirkung der Schulen, der Motivation und Innovationsbereitschaft des Kollegiums, den Möglichkeiten zum Umgang mit der sozialen und leistungsbezogenen Heterogenität der Schülerinnen und Schüler und zur individuellen Förderung (etwa im Bereich der Sprachförderung) sowie der generellen Ausstattung der Schulen“ (Böse, Neumann & Maaz, 2018). Andere schulische Merkmale, wie Fehl- und Abbrecherquoten und Förderprognosen der Grundschulen zeigten keine Verbesserungen durch das Programm (Böse et al., 2018). Weiterhin zeigen Erfahrungen mit dem „Turnaround“-Projekt (Robert Bosch Stiftung GmbH, 2018) in Berlin, an dem drei Grundschulen und sieben Sekundarschulen teilnahmen, dass positive Entwicklung durch eine vielfältige Unterstützung durch das Projekt erzielt wurden. Dazu gehören Prozessbegleiter, Fortbildungsangebote, Hospitationen in anderen Schulen, finanzielle Ressourcen, gemeinsame Workshops und Zwischenbilanzgespräche mit der Schulaufsicht. Die Evaluation des Programms zeigt zwar positive Entwicklungen, macht jedoch aufgrund der Kombination des Programms mit anderen Maßnahmen (bspw. mit proSchul und dem Bonusprogramm) keine Aussage darüber, ob die Erfolge aufgrund des Programms zu Stande kamen (Robert Bosch Stiftung GmbH, 2018).

Nichtsdestotrotz deuten die Erfolge daraufhin, dass derartige Programme der finanziellen Schulunterstützung, in denen Schulen selbstständig entscheiden können und in engem Austausch mit der Schulverwaltung stehen, weiter zur Unterstützung von „failing schools“ genutzt werden könnten.

Schulinspektion

Es können verschiedene Implikationen für die Schulinspektion aus den Studien dieser Dissertation gezogen werden: Grundsätzlich wird der Ansatz der Schulinspektion durch Befunde aus Studie 1 gestützt, dass es ein Qualitätsprofil des Unterrichts der Schule gibt. Dennoch zeigten sich dabei auch Unterschiede zwischen Unterrichtsmerkmalen – manche weisen deutlich weniger Übereinstimmung und damit mehr Variabilität in den Urteilen der Schülerinnen und Schüler derselben Schule auf. Es wäre daher gewinnbringend neben Mittelwerten auch Indizes zur Variabilität zu nutzen (siehe Studie 1, Gärtner, 2010; Lüdtke et al., 2006). Da die Schulinspektion eine große Anzahl an Unterrichtsbesuchen in jeder Schule durchführt, erscheint es sinnvoll, diese Informationen über die Variabilität innerhalb der Schule *zusätzlich* zum Unterrichtsprofil der Schule zu nutzen und an die Schulen zurückzumelden. Insbesondere für den Umgang der Schulen mit den Schulinspektionsergebnissen, kann dies nützlich sein: So ist es vorstellbar, dass es für die Weiterarbeit der Schulen einen Unterschied macht, ob in der Folge sehr guter und sehr schlechter Bewertungen des Unterrichts beispielsweise ein mittleres Qualitätsniveau des Unterrichts an die Schule zurückgemeldet wird oder eine hohe Variabilität der Unterrichtsqualität in der gleichen Schule angesprochen wird.

Implikationen bezüglich der Schulinspektionsdiagnose als „Schulen mit erheblichem Entwicklungsbedarf“ (oder anderen Bezeichnungen, siehe Studie 3), wurden bereits in der Diskussion zu Studie 3 angesprochen und sollen hier nur sehr knapp wiederholt werden: Die Diagnose führt lediglich zu einer geringen Verbesserung der Unterrichts- und Schulqualität der Schulen mit Entwicklungsbedarf und trägt nicht zur Verbesserung der Schulleistung bei, wohingegen sich nichtintendierte Nebeneffekte bei der Zusammensetzung der Schülerschaft zeigen. Daraus ergibt sich, dass hier womöglich eine andere oder verstärkte Form der Unterstützung genutzt werden sollte und allein die Diagnose nicht den gewünschten Effekt

bringt. Beispielsweise wurde in Berlin die Beratungseinrichtung „proSchul“ in diesem Zusammenhang gegründet, es gibt jedoch keine Statistiken darüber, wie viele Schulen mit Entwicklungsbedarf diese nutzen oder Evaluationsergebnisse zu der Beratung durch proSchul.

Ein weiterer Punkt, der für die Schulinspektion von Bedeutung ist, betrifft den, der Evaluation zugrundeliegenden, Handlungsrahmen zur Schulqualität. Wie bereits in den Limitationen dieser Arbeit diskutiert (Kapitel 5.3), entstammen die Indikatoren der Schulinspektion nur teilweise forschungstheoretischen Konzepten zu Unterrichts- und Schulqualität (siehe Kapitel 2). Eine Nutzung der empirischen Forschung zu Bedingungen guten Unterrichts und guter Schulqualität könnte jedoch dazu beitragen, die Evaluationskriterien der Schulinspektion inhaltlich zu schärfen, was wiederum auch für die schulische Weiterarbeit auf Basis dieser Ergebnisse hilfreich sein kann. Dazu sollte auch beachtet werden, welche „Messbedingungen“ bzw. Evaluationsstandards gegeben sein sollten, um darauf basierend relevante Entscheidungen treffen zu können. Des Weiteren soll an dieser Stelle erwähnt werden, dass die Indikatoren der Schulinspektion nicht nur outputorientiert angelegt, also nicht nur Merkmale betreffen sollten, die in Zusammenhang mit der Verbesserung schulischer Leistungen stehen, sondern auch Merkmale berücksichtigen sollten, die beispielsweise gesellschaftlich relevant sind. Als ein Beispiel kann hier die Berücksichtigung von Heterogenität der Schülerschaft (und somit auch der Kompositionsmerkmale der Schule) hervorgehoben werden: Das Schaffen von gleichen Chancen führt zwar nicht zwangsläufig zu besserem Output im Sinne der Leistung von Schülerinnen und Schülern, ist jedoch trotzdem ein hochrelevantes Ziel von Schulen und sollte sich daher auch in den Schulinspektionsindikatoren widerspiegeln.

Es existieren international diskutierte Bestrebungen der Schulinspektion, hin zu risikobasierten Inspektionen, die beinhalten, dass nicht mehr alle Schulen regelmäßig begutachtet werden, sondern lediglich die Schulen „mit hohem Risiko“ (Ehren & Shackleton, 2016; Timmermans, Wolf, Bosker & Doolaard, 2015). Diese Bestrebungen stehen eher im Widerspruch zu der zuvor genannten Empfehlung, da sie sich sehr häufig an Outputs wie Leistungsmaßen, oder an einer Kombination mit Kompositionsmerkmalen orientieren (Ehren & Shackleton, 2016; Timmermans et al., 2015) und zudem Unterrichts- und Schulqualität als

zeitlich stabile Konstrukte auf Schulebene angesehen werden müssten, was für einen solchen Zweck zunächst evaluiert werden müsste (vgl. 5.3 Stabilität von Unterrichtsqualität).

5.4.3 Für die Praxis

Schulleitungen und Schulen

Auch für Schulleitungen ergeben sich Implikationen aus dieser Dissertation. Das Ergebnis aus Studie 2, dass ein Zusammenhang zwischen Komposition und Prozessmerkmalen (insbesondere Unterrichtsqualität) der Schule besteht, legt eine Priorisierung der Verbesserung der Unterrichtsqualität nahe. Dies sollte eine hohe Priorität an allen Schulen einnehmen, erscheint aber insbesondere für Schulen mit „schwieriger" Schülerkomposition (insbesondere mit niedrigem SES, geringeren Leistungen) zentral (vgl. bspw. Chapman & Harris, 2004; Muijs et al., 2004). Wie bereits angesprochen, ergibt sich durch eine geringe Unterrichtsqualität eine Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf" und daraus folgend kommt es womöglich zur Vermeidung der Schule bei bestimmten Gruppen, was wiederum eine Veränderung der Komposition für die Schule bedingen kann (Studie 3). Durch eine Verbesserung der Unterrichtsqualität insbesondere an Schulen mit „schwieriger" Komposition könnte der Zusammenhang zwischen Komposition und Unterrichtsqualität verringert werden und zudem könnten womöglich Segregationsprozesse der Schülerschaft aufgehalten werden, indem die Schulen für alle Klientel ansprechend bleiben. Im Interesse der Schulen (insbesondere „failing schools") könnten gezielt Strategien entwickelt werden, die sich mit der eigenen Kompositionssituation auseinandersetzen und diese in den Konzepten ihres Unterrichts sowie im Schulalltag berücksichtigen. Solche Strategien zum Umgang auch mit einer heterogenen Schülerschaft könnten dann auch auf der Schulhomepage oder bei Informationsveranstaltungen der Schulen kommuniziert werden, um Segregationsprozesse bei der Schulwahl zu verhindern oder abzumindern (Böhm-Kasper, Selders & Lambrecht, 2016; Clausen, 2006; Jurczok, 2019). Eine besondere Rolle in diesen Strategien könnte dabei die Berücksichtigung einer heterogenen Schülerschaft haben. Maßnahmen zur Verbesserung des Klassenmanagements sind beispielsweise in diesem Kontext relevant, um die Lernzeit der Schülerinnen und Schüler zu erhöhen (Kunter & Voss, 2011). Zudem profitieren Schülerinnen

und Schüler mit Migrationshintergrund besonders von einer effektiven Klassenführung (Seiz, Decristan, Kunter & Baumert, 2016). Weiterhin können Konzepte zur inneren Differenzierung im Unterricht mittels selbstständigem und kooperativem Lernen oder eine Orientierung am Response-to-Intervention-Modell (RTI) eine Chance für heterogene Lerngruppen bieten (vgl. bspw. Vock & Gronostaj, 2017). Eine weitere Möglichkeit bietet die Auseinandersetzung mit Interventionen zur Verhinderung eines Stereotype-Threat-Effekts (vgl. bspw. BIM, 2017). Dieser beinhaltet, dass Schülerinnen und Schüler, die sich einer negativ stereotypisierten Gruppe angehörig fühlen, davon bedroht fühlen und gegebenenfalls schlechtere Leistungen erbringen bzw. weniger motiviert sind (vgl. bspw. Steele & Aronson, 1995).

Die effektive Leitung bzw. das gemeinsame Handeln und Planen als Schule kann insbesondere an Schulen mit „schwierigen Schülerschaften" eine Chance sein, um die Herausforderung nicht allein auf die Unterrichtsebene zu geben, auf der die Lehrkräfte dann einzeln damit umgehen müssen. Schulleitungen könnten es als Chance sehen, über „das Gemeinsame" der Schule bezüglich des Unterrichtsprofils (Studie 1), einen Einfluss auf den Unterricht der einzelnen Lehrkräfte ausüben zu können (vgl. bspw. Goddard, Goddard, Sook Kim & Miller, 2015; Hallinger & Heck, 1996). Zudem kann es auch hier hilfreich sein, Lehrkräfte dafür zu sensibilisieren, dass Unterricht sich an den Schulen je nach Komposition unterscheidet (Studie 2) und die eigene Rolle (eigenes Verhalten, eigene Einstellungen) zur Verbesserung der Chancen *aller* Schülerinnen und Schüler zu betonen (Dubberke, Kunter, McElvany, Brunner & Baumert, 2008; Trautmann & Wischer, 2011; Vock & Gronostaj, 2017). Diesbezüglich sollten die Erwartungen, die an die jeweilige Schülerschaft der einzelnen Schule herangetragen werden, thematisiert werden um eine Benachteiligung bestimmter Schülerinnen und Schüler zu reduzieren (BIM, 2017) sowie die kognitive Aktivierung im Unterricht zu erhöhen.

Weiterhin kann festgehalten werden, dass das Feedback der Schulinspektion an Schulen genutzt werden kann und bei offenem Umgang mit Kritik auch in hilfreiche Maßnahmen zur Verbesserung der Unterrichtsqualität umgesetzt werden kann (Studie 3). Zur Verbesserung von Unterrichtsqualität (Prozessqualität) kann es zudem hilfreich sein, interne Evaluationsmöglichkeiten zu nutzen (siehe Kapitel 5.4.2), sowie eine Kultur im Kollegium zu

schaffen, die es ermöglicht, mit Fehlern umzugehen (vgl. bspw. Spychiger, Oser, Hascher & Mahler, 1999). Dadurch könnten kollegiale Hospitationen und gemeinsames Weiterentwickeln und Reflektieren von Unterricht zur Normalität werden.

Lehrerbildung

Da die Variabilität der Prozessmerkmale (insbesondere Unterrichtsqualität) in Abhängigkeit der Schülerkomposition (Studie 2) ein Problem darstellt, sollte die Verbesserung der Unterrichtsqualität von Schulen auch für die Lehrerbildung eine besondere Rolle spielen. Dabei stellt insbesondere der Umgang mit Heterogenität im Unterricht ein zentrales Thema dar, das in der Lehrerbildung (in allen Phasen) verstärkt berücksichtigt werden sollte, um auf eine unterschiedliche Schülerschaft hinsichtlich ihrer Leistung, ihres sozialen Hintergrunds und ihres Migrationshintergrunds eingehen zu können. Unter Lehrerbildung kann dabei das Lehramtsstudium und der Vorbereitungsdienst verstanden werden (erste und zweite Phase der Lehrerbildung), in dem die fachliche, didaktische und bildungswissenschaftliche Wissensbasis erworben wird (KMK, 2019). Zunächst sollte dabei eine Sensibilisierung der Lehrkräfte für den Zusammenhang zwischen der Komposition der Schülerschaft und der Unterrichtsqualität stattfinden. Dies könnte im Rahmen von bildungswissenschaftlichen Teilen des Studiums erfolgen. Zudem kann die Berücksichtigung von heterogenen Schülerschaften sowohl in fachdidaktischen aber auch in bildungswissenschaftlichen Veranstaltungen des Studiums im Fokus stehen, so wie dies auch in den „Standards für die Lehrerbildung“ (KMK, 2019) gefordert wird. Anschließend an die Implikationen für die Bildungspolitik (Kapitel 5.4.2) und für Schulleitungen und Schulen (Kapitel 5.4.3), können diese Ansätze zur Verbesserung der Unterrichtsqualität (vgl. bspw. Baptist & Raab, 2007; Dalehefte et al., 2014; Piwowar et al., 2013; Thiel et al., 2013), aber auch zur Bedeutung der Einstellungen und Haltungen der Lehrkraft (Dubberke et al., 2008; Trautmann & Wischer, 2011; Vock & Gronostaj, 2017), die Reflexion der eigenen Erwartungen an die Schülerinnen und Schüler sowie effizientes Klassenmanagement bereits im Rahmen von bildungswissenschaftlichen Veranstaltungen der ersten und zweiten Phase der Lehrerbildung thematisiert werden. Beispielsweise können dafür Videosequenzen genutzt werden (Barth, 2017; Barth, Piwowar, Kumschick, Ophardt & Thiel,

2019). Zusätzlich kann auch im Bereich der Fort- und Weiterbildung (der dritten Phase der Lehrerbildung, KMK, 2019) der Umgang mit Heterogenität und die Steigerung von Unterrichtsqualität berücksichtigt werden.

5.5 Fazit

Ziel dieser Dissertation war es, die Variabilität schulischer Lernumwelten differenziert zu betrachten. Die Variabilität schulischer Lernumwelten beinhaltet unter anderem Unterschiede in der Unterrichts- und Schulqualität und das Zusammenspiel dieser Prozessmerkmale mit Kontextmerkmalen sowie die Auswirkungen auf Leistungen von Schülerinnen und Schülern. Um diese Aspekte zu beleuchten, wurden drei empirische Studien durchgeführt, die in Anlehnung an ein Modell der Schuleffektivitätsforschung (Scheerens, 1990), erweitert durch Elemente der Lehr-Lernforschung (Klieme et al., 2001; Kunter & Voss, 2011) eingeordnet wurden (Abb. 1).

Inhaltlich ergaben sich zwei übergeordnete Leitthemen, entlang derer sich die Erkenntnisse dieser Arbeit zusammenfassend verdeutlichen lassen. Dies sind (1) die Schule als relevante Analyseeinheit und (2) die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität sowie Schulleistung. Es zeigt sich, dass (1) die Schule als relevante eigenständige Handlungs- und Analyseeinheit gesehen werden kann, da innerhalb von Schulen weitestgehend Übereinstimmung hinsichtlich der Unterrichtsqualität besteht und sich Schulen in der Qualität ihres Unterrichts unterscheiden (Studie 1). Zudem reagieren Schulen als Gesamtkomplex auf Kontextmerkmale wie Komposition und Intervention (Studien 2 und 3), sodass Prozessmerkmale und Outputs der Schulen variieren. Diese Ergebnisse weisen auf (2) die Bedeutung des Schulkontexts für Unterrichts- und Schulqualität sowie für die Leistungen von Schülerinnen und Schülern hin. Hierbei wurden mehrere Verbindungen des vorgestellten Modells (Verbindung Kontext-Prozess *und* Kontext-Output) sowie mehrere Merkmale schulischer Qualität (wie bspw. mehrere Kompositionsmerkmale, mehrere Prozessmerkmale) betrachtet um eine möglichst umfassende Analyse zu ermöglichen: Beispielsweise zeigt sich dadurch, dass Zusammenhänge zwischen verschiedenen Kompositionsmerkmalen und Prozessmerkmalen unterschiedlich stark ausgeprägt sind und sich die bildungspolitische

Intervention durch die Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“ unterschiedlich auf Prozessmerkmale und Outputs auswirkt. Weiterhin verdeutlichen die Ergebnisse dieser Arbeit die Relevanz der Berücksichtigung des schulspezifischen Kontexts (Komposition und Intervention) für Prozessmerkmale wie Unterrichts- und Schulqualität sowie für schulspezifischen Output wie Leistungen: Es bestehen Zusammenhänge zwischen Kompositionsmerkmalen und Unterrichtsqualität, zudem wird der Zusammenhang zwischen Kompositionsmerkmalen und Leistungen durch ein Merkmal der Unterrichtsqualität (Klassenführung) mediiert (Studie 2). Außerdem zeigen sich durch die Schulinspektionsdiagnose „Schule mit erheblichem Entwicklungsbedarf“ (Intervention) geringe Verbesserungen der Unterrichts- und Schulqualität, keine Verbesserung der Schulleistung, jedoch nichtintendierte Nebeneffekte auf die Zusammensetzung der Schülerschaft (in Bezug auf den Migrationshintergrund von Schülerinnen und Schülern; Studie 3).

Aus den Ergebnissen der vorliegenden Dissertation ergaben sich Schlussfolgerungen für die Forschung zur weiteren Berücksichtigung der Schule als Analyseeinheit als Forschungsthema und weiterer Studien zur detaillierteren Überprüfung der Zusammenhänge zwischen Kontext- und Prozessmerkmalen. Weitere Implikationen zeigen sich für die Bildungspolitik, indem Ideen aufgezeigt werden, wie diese eine hohe Unterrichtsqualität und eine geringere Bedeutung des schulspezifischen Kontexts fördern kann und wie „schwierige“ Schulen unterstützt werden können. Zudem werden Implikationen für die Schulinspektion bezüglich der Variabilität des Unterrichts, der Schulinspektionsdiagnose „erheblicher Entwicklungsbedarf“, bezüglich der empirischen Grundlage des Handlungsrahmens zur Schulqualität und zu risikobasierten Inspektionen diskutiert. Weiterhin können mit Blick auf die Schulpraxis Schlussfolgerungen aus den Ergebnissen dieser Dissertation für Schulleitungen gezogen werden: Dies betrifft mögliche Schulstrategien zur Auseinandersetzung mit den spezifischen Kompositionsmerkmalen der Einzelschulen sowie zur Erhöhung der Unterrichtsqualität, insbesondere unter Berücksichtigung der Heterogenität im Unterricht. Schließlich werden daran anknüpfende Schlussfolgerungen auch für die Lehrerbildung hinsichtlich der benannten Aspekte gezogen.

Literaturverzeichnis

Abgeordnetenhaus Berlin. Schulgesetz, §9 Qualitätssicherung und Evaluation. Zugriff am 13.04.2020. Verfügbar unter http://gesetze.berlin.de/jportal/?quelle=jlink&query=SchulG+BE+TEIL+I&psml=bsbeprod.psml&max=true

Abgeordnetenhaus Berlin. Verordnung über schulische Qualitätssicherung und Evaluation, §6 Evaluation von Lehrkräften. Zugriff am 13.04.2020. Verfügbar unter http://gesetze.berlin.de/jportal/?quelle=jlink&query=EvalV+BE&psml=bsbeprod.psml&max=true

Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, *81*(1), 117–152.

Altrichter, H. & Maag Merki, K. (Hrsg.). (2010). *Handbuch neue Steuerung im Schulsystem.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Artelt, C., Bug, M., Kleinert, C., Maaz, K. & Runge, T. (2019). Nutzungspotenziale amtlicher Statistik in der Bildungsforschung. Ein Überblick zu Erreichtem, möglichen Chancen und anstehenden Herausforderungen. In D. Fickermann & H. Weishaupt (Hrsg.), *Bildungsforschung mit Daten der amtlichen Statistik* (Die Deutsche Schule. Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis - Beiheft, Bd. 14, 1. Auflage, S. 21–37). Münster: Waxmann.

Baptist, P. & Raab, D. (2007). *Auf dem Weg zu einem veränderten Mathematikunterricht.* Zentrum zur Förderung des mathematisch-naturwissenschaftlichen Unterrichts, Universität Bayreuth (Hrsg.). Bayreuth.

Barth, V. L. (2017). *Professionelle Wahrnehmung von Störungen im Unterricht*. Wiesbaden: Springer Fachmedien Wiesbaden.

Barth, V. L., Piwowar, V., Kumschick, I. R., Ophardt, D. & Thiel, F. (2019). The impact of direct instruction in a problem-based learning setting. Effects of a video-based training program to foster preservice teachers' professional vision of critical incidents in the classroom. *International Journal of Educational Research*, *95*, 1–12.

Baumert, J., Stanat, P. & Watermann, R. (Hrsg.). (2006a). *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., Stanat, P. & Watermann, R. (2006b). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (S. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.

Begrich, L., Fauth, B., Kunter, M. & Klieme, E. (2017). Wie informativ ist der erste Eindruck? Das Thin-Slices-Verfahren zur videobasierten Erfassung des Unterrichts. *Zeitschrift für Erziehungswissenschaft*, *20*(1), 23–47.

BIM. (2017). *Vielfalt im Klassenzimmer. Wie Lehrkräfte gute Leistung fördern können*. Berliner Institut für empirische Integrations- und Migrationsforschung (BIM)/Forschungsbereich beim Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR-Forschungsbereich). Berlin.

Bischof, L. M. (2014). *Schulentwicklung und Schuleffektivität* (Schulentwicklungsforschung, Bd. 1). Dissertation. Wiesbaden: Springer VS.

Bliese, P. (2000). Within-group agreement, non-independence, and reliability. Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Hrsg.), *Multilevel Theory, Research, and Methods in Organizations. Foundations, Extensions, and New Directions* (Frontiers of industrial and organizational psychology, S. 349–381). San Francisco: Jossey-Bass.

Böhm-Kasper, O., Selders, O. & Lambrecht, M. (2016). Schulinspektion und Schulentwicklung – Ergebnisse der quantitativen Schulleitungsbefragung. In Arbeitsgruppe Schulinspektion (Hrsg.), *Schulinspektion als Steuerungsimpuls? Ergebnisse aus Forschungsprojekten* (Educational governance, Band 25, 1. Aufl. 2016, S. 1–50). Wiesbaden: Springer VS.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G. & Pierce, C. A. (2015). Correlational effect size benchmarks. *The Journal of Applied Psychology*, *100*(2), 431–449.

Böse, S., Neumann, M. & Maaz, K. (Juni 2018). *BONUS-Studie. Wissenschaftliche Begleitung und Evaluation des Bonus-Programms zur Unterstützung von Schulen in schwieriger Lage in Berlin. Zweiter Ergebnisbericht über die Erhebungen aus den Schuljahren 2013/2014, 2015/2016 und 2016/2017*. Berlin.

Brand, S., Felner, R., Shim, M., Seitsinger, A. & Dumas, T. (2003). Middle school improvement and reform: Development and validation of a school-level assessment of climate, cultural pluralism, and school safety. *Journal of Educational Psychology*, *95*(3), 570–588.

Brimblecombe, N., Shaw, M. & Ormston, M. (1996). Teachers' intention to change practice as a result of Ofsted school inspections. *Educational Management & Administration*, *24*(4), 339–354.

Brookover, W. B., Beady, C. H., Flood, P. K., Schweitzer, J. H. & Wisenbaker, J. M. (1979). *School social systems and student achievement. Schools can make a difference*. New York: Praeger Publishers.

Chapman, C. & Harris, A. (2004). Improving schools in difficult and challenging contexts: strategies for improvement. *Educational Research*, *46*(3), 219–228.

Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 29). Münster: Waxmann.

Clausen, M. (2006). Warum wählen Sie genau diese Schule? Eine inhaltsanalytische Untersuchung elterlicher Begründungen der Wahl der Einzelschule innerhalb eines Bildungsgangs. *Zeitschrift für Pädagogik*, *52*(1), 69–90.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Hillsdale NJ: Erlbaum.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. et al. (1966). *Equality of Educational Opportunity*. Washington: Government Printing Office.

Creemers, B. P. M. & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*(3), 347–366.

Dalehefte, I. M., Wendt, H., Köller, O., Wagner, H., Pietsch, M., Döring, B. et al. (2014). Bilanz von neun Jahren SINUS an Grundschulen in Deutschland. Evaluation der mathematikbezogenen Daten im Rahmen von TIMSS 2011. *Zeitschrift für Pädagogik*, *60*(2), 245–263.

Davis, H. A. (2003). Conceptualizing the role and influence of student-teacher relationships on children's social and cognitive development. *Educational Psychologist*, *38*(4), 207–234.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M.-C. & Onghena, P. (2003). The effect of schools and classes on language achievement. *British Educational Research Journal*, *29*(6), 841–859.

Decristan, J., Fauth, B., Kunter, M., Büttner, G. & Klieme, E. (2017). The interplay between class heterogeneity and teaching quality in primary school. *International Journal of Educational Research*, *86*, 109–121.

Deutsches PISA-Konsortium (Hrsg.). (2003). *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Opladen: Leske + Budrich.

DGfE, GEBF, GFD (Hrsg.). (2020). *Empfehlungen zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten im Kontext erziehungs- und bildungswissenschaftlicher sowie fachdidaktischer Forschung. Gemeinsame Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), der Gesellschaft für Empirische Bildungsforschung (GEBF) und der Gesellschaft für Fachdidaktik (GFD) zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten in den Erziehungs- und Bildungswissenschaften und Fachdidaktiken.* Zugriff am 25.03.2020. Verfügbar unter https://www.dgfe.de/fileadmin/OrdnerRedakteure/Stellungnahmen/2020_03_Forschungsdatenmanagement.pdf

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein & E. Terhart (Hrsg.), *Qualität und Qualitätssicherung im Bildungsbereich; Schule, Sozialpädagogik, Hochschule* (Zeitschrift für Pädagogik. Beiheft, Bd. 41, S. 73–92). Weinheim: Beltz.

Ditton, H. & Müller, A. (2011). Schulqualität. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung* (S. 99–111). Wiesbaden: VS Verlag für Sozialwissenschaften.

Donaldson, M. L., LeChasseur, K. & Mayer, A. (2017). Tracking instructional quality across secondary mathematics and English Language Arts classes. *Journal of Educational Change*, *18*(2), 183–207.

Dreeben, R. & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education*, *61*(3), 129–142.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M. & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften. *Zeitschrift für Pädagogische Psychologie*, *22*(34), 193–206.

Dumont, H., Neumann, M., Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen. *Psychologie in Erziehung und Unterricht*, *60*(3), 163–183.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D. & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432.

Ehren, M. C. M. & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability*, *28*(4), 299–321.

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

Eklund, K., Kilpatrick, K. D., Kilgus, S. P. & Haider, A. (2018). A systematic review of state-level social–emotional learning standards: Implications for practice and research. *School Psychology Review*, *47*(3), 316–326.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Fend, H. (1977). *Schulklima. Soziale Einflussprozesse in der Schule.* Weinheim: Beltz.

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

Fend, H. (1988). Schulqualität. Die Wiederentdeckung der Schule als pädagogische Gestaltungsebene. *Neue Sammlung*, *28*(4), 537–547. Zugriff am 04.12.2019. Verfügbar unter https://www.pedocs.de/volltexte/2009/1629/pdf/Fend_Helmut_Schulqualitaet._Die_Wiederentdeckung_D_A.pdf

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität* (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Fickermann, D. & Weishaupt, H. (Hrsg.). (2019). *Bildungsforschung mit Daten der amtlichen Statistik* (Die Deutsche Schule. Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis - Beiheft, Bd. 14, 1. Auflage). Münster: Waxmann.

Garner, C. L. & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education*, *64*(4), 251.

Gärtner, H. (2010). Wie Schülerinnen und Schüler ihre Lernumwelt wahrnehmen. *Zeitschrift für Pädagogische Psychologie*, *24*(2), 111–122.

Gärtner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, *42*, 91–99.

Gärtner, H. & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(2), 159–182.

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

Gärtner, H. & Vogt, A. (2013). Selbstevaluation des Unterrichts: Wie Lehrkräfte Ergebnisse eines Schülerfeedback rezipieren. *Unterrichtswissenschaft*, *41*(3), 255–270.

Goddard, R., Goddard, Y., Sook Kim, E. & Miller, R. (2015). A theoretical and empirical analysis of the roles of instructional leadership, teacher collaboration, and collective efficacy beliefs in support of student learning. *American Journal of Education*, *121*(4), 501–530.

Haertel, G. D., Walberg, H. J. & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research*, *53*(1), 75–91.

Hallinger, P. & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly*, *32*(1), 5–44.

Harker, R. & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*(2), 177–199.

Hattie, J. (2018). *Lernen sichtbar machen* (4. unveränderte Auflage). Baltmannsweiler: Schneider Verlag Hohengehren GmbH.

Helbig, M. (2010). Neighborhood does matter! *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *62*(4), 655–679.

Helbig, M. & Nikolai, R. (2017). Ansturm auf „gute" Schulen? Die Auswirkungen der Veröffentlichung von Abiturnoten auf die Zusammensetzung von Schülerinnen und Schülern an Berliner Schulen. *Zeitschrift für Bildungsforschung*, *7*(2), 115–130.

Helbig, M., Nikolai, R. & Wrase, M. (2017). Privatschulen und die soziale Frage Wirkung rechtlicher Vorgaben zum Sonderungsverbot in den Bundesländern. *Leviathan*, *45*(3), 357–380.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (1. Aufl.). Seelze-Velber: Kallmeyer.

Helmke, A. & Lenske, G. (2013). Unterrichtsdiagnostik als Voraussetzung für Unterrichtsentwicklung. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *31*(2), 214–233. Verfügbar unter https://www.pedocs.de/volltexte/2017/13848/pdf/BZL_2013_2_214_233.pdf

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79.

Horr, A. (2015). Nachbarschaftseffekte. In C. Diehl, C. Hunkler & C. Kristen (Hrsg.), *Ethnische Ungleichheiten im Bildungsverlauf. Mechanismen, Befunde, Debatten* (EBL-Schweitzer, Online-ausg, S. 397–430). Wiesbaden: Springer Fachmedien Wiesbaden.

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M. et al. (Hrsg.). (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (1. Auflage). Münster: Waxmann.

Jacob, R. T., Goddard, R. D. & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions. *Educational Evaluation and Policy Analysis*, *36*(1), 44–66.

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), 85–98.

Jencks, C. (1972). *Inequality: A reassessment of the effect of family and schooling in America*: Basic Books, New York.

John, O. P. & De Fruyt, F. (2015). *Framework for the longitudinal study of social and emotional skills in cities.* (OECD, Hrsg.). Zugriff am 28.02.2020. Verfügbar unter http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/CERI/CD(2015)13&docLanguage=En

Jurczok, A. (2019). *Schulwahl unter „gleichwertigen" Einzelschulen*. Wiesbaden: Springer Fachmedien Wiesbaden.

Jurczok, A. & Lauterbach, W. (2014). Schulwahl von Eltern: Zur Geografie von Bildungschancen in benachteiligten städtischen Bildungsräumen. In P. A. Berger, C. Keller, A. Klärner & R. Neef (Hrsg.), *Urbane Ungleichheiten. Neue Entwicklungen zwischen Zentrum und Peripherie* (Sozialstrukturanalyse, S. 135–155). Wiesbaden: Springer Fachmedien Wiesbaden.

Klieme, E. & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 333–359). Opladen: Leske + Budrich.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS – Impulse für Schule und Unterricht* (S. 43–57). Bonn.

KMK. (2004). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung*. Veröffentlichungen der Kultusministerkonferenz: Luchterhand.

KMK. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Bonn: Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

KMK. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* Berlin: Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

KMK. (2019). *Standards für die Lehrerbildung: Bildungswissenschaften. (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2019)*. Berlin, Bonn.

Kotthoff, H.-G. & Böttcher, W. (2009). Neue Formen der „Schulinspektion": Wirkungshoffnungen und Wirksamkeit im Spiegel empirischer Bildungsforschung. In H.

Altrichter (Hrsg.), *Neue Steuerung im Schulsystem. Ein Handbuch* (Educational governance, Bd. 7, 1. Aufl., S. 295–325). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kunter, M. & Baumert, J. (2008). Schuleffekte. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 527–538). Hogrefe Verlag.

Kunter, M., Baumert, J., Blum, W. & Neubrand, M. (Hrsg.). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W. et al. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 502–520.

Kunter, M., Klusmann, U. & Baumert, J. (2009). Professionelle Kompetenz von Mathematiklehrkräften: Das COACTIV-Modell. Sonderdruck. In O. Zlatkin-Troitschanskaia, K. Beck, O. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrerprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (Beltz Bibliothek, S. 153–166). Weinheim: Beltz.

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 85–114). Münster: Waxmann.

Landwehr, N. (2011). Wirkungen und Wirksamkeit der externen Schulevaluation. In C. Quesel (Hrsg.), *Wirkungen und Wirksamkeit der externen Schulevaluation* (S. 35–70). Bern: Hep der Bildungsverl.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852.

Lenkeit, J. (2012). How effective are educational systems? A value-added approach to measure trends in PIRLS. *Journal for Educational Research Online*, *4*(2), 143–173.

Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda & E. Terhart (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern: Ausbildung und Beruf* (Zeitschrift für Pädagogik. Beiheft, Bd. 51). Weinheim: Beltz Verlag.

Liu, H., Van Damme, J., Gielen, S. & Van Den Noortgate, W. (2015). School processes mediate school compositional effects. Model specification and estimation. *British Educational Research Journal*, *41*(3), 423–447.

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.

Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch neue Steuerung im Schulsystem* (S. 145–170). Wiesbaden: VS Verlag für Sozialwissenschaften.

Manitius, V. & Dobbelstein, P. (Hrsg.). (2017). *Schulentwicklungsarbeit in herausfordernden Lagen* (Beiträge zur Schulentwicklung, 1. Auflage). Münster: Waxmann.

Matsumura, L. C., Garnier, H., Pascal, J. & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, *8*(3), 207–229.

Messner, R. (2016). Die Einzelschule als pädagogische Handlungseinheit und das Zusammenspiel der Handlungsebenen und institutionellen Akteure. In U. Steffens & T. Bargel (Hrsg.), *Schulqualität - Bilanz und Perspektiven. Grundlagen der Qualität von Schule 1* (Beiträge zur Schulentwicklung, 1. Auflage, S. 95–114). Münster: Waxmann.

Meyers, C. v. & Murphy, J. (2007). Turning around failing schools: An analysis. *Journal of School Leadership*, *17*(5), 631–659.

Mickelson, R. A., Bottia, M. C. & Lambert, R. (2013). Effects of school racial composition on K–12 mathematics outcomes. *Review of Educational Research*, *83*(1), 121–158.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. & Russell, E. (1988). *School matters.* Berkeley: University of California Press.

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.

Müller, S. & Pietsch, M. (2011). Was wir messen, wenn wir Unterrichtsqualität messen. Inter-Beurteilerübereinstimmung und -Reliabilität bei Unterrichtsbeobachtungen im Rahmen von Schulinspektion. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz aus empirischer Sicht* (S. 33–56). Waxmann Verlag GmbH.

National Research Council. (2012). *Education for Life and Work*. Washington, D.C.: National Academies Press.

Niemiec, C. P. & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom. *Theory and Research in Education*, *7*(2), 133–144.

OECD. (2002). *PISA 2000 Technical Report*. Paris: OECD Publishing. Zugriff am 11.04.2016. Verfügbar unter http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33688233.pdf

OECD. (2005). *School factors related to quality and equity. Results from PISA 2000*. Paris: OECD.

OECD. (2008). *Measuring improvements in learning outcomes. Best practices to assess the value-added of schools*. Paris: OECD.

OECD. (2010). *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)* (PISA). Paris: OECD Publishing.

OECD. (2013). *PISA 2012 Results: Excellence through Equity (Volume II). Giving Every Student the Chance to Succeed* (PISA). Paris: OECD Publishing.

Opdenakker, M.-C. & Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematics achievement. *British Educational Research Journal*, *27*(4), 407–432.

Ophardt, D. & Thiel, F. (2008). Klassenmanagement als Basisdimension der Unterrichtsqualität. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (Schule und Gesellschaft, Bd. 24, 2., vollst. überarb. Aufl., S. 259–282). Wiesbaden: VS, Verl. für Sozialwiss.

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. & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, *11*(3), 430–452.

Piwowar, V., Thiel, F. & Ophardt, D. (2013). Training inservice teachers' competencies in classroom management. A quasi-experimental study with teachers of secondary schools. *Teaching and Teacher Education*, *30*, 1–12.

Praetorius, A. K. (2013). *Messung von Unterrichtsqualität durch Ratings*: Waxmann Verlag GmbH.

Raudenbush, S. W. & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335.

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

Richter, D., Böhme, K., Becker, M., Pant, H. A. & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik*, *60*(2), 225–244.

Riedel, A., Schneider, K., Schuchart, C. & Weishaupt, H. (2010). School choice in German primary schools: How binding are school districts? *Journal for Educational Research Online / Journal Für Bildungsforschung Online*, *2*(1), 94–120.

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E. & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The mediating role of instructional quality. *Learning and Instruction*, *32*, 63–72.

Robert Bosch Stiftung GmbH. (n. d.). *Der Deutsche Schulpreis*. Zugriff am 24.03.2020. Verfügbar unter https://www.deutscher-schulpreis.de/

Robert Bosch Stiftung GmbH (Hrsg.). (2018). *Pilotprojekt „School Turnaround – Berliner Schulen starten durch". Zentrale Erkenntnisse und Empfehlungen der wissenschaftlichen Begleitstudie*. Zugriff am 17.03.2020. Verfügbar unter https://www.bosch-stiftung.de/sites/default/files/publications/pdf/2018-09/School_Turnaround_Begleitstudie.pdf

Rolff, H.-G. (2010). Schulentwicklung als Trias von Organisations-, Unterrichts- und Personalentwicklung. In *Handbuch Schulentwicklung. Theorie - Forschungsbefunde - Entwicklungsprozesse - Methodenrepertoire* (UTB, 8443 : Schulpädagogik, S. 29–36). Bad Heilbrunn: Klinkhardt.

Sammons, P., Hillmann, J. & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. Verfügbar unter http://files.eric.ed.gov/fulltext/ED389826.pdf

Saunders, L. (1999). A brief history of educational "Value-Added": How did we get to where we are? *School Effectiveness and School Improvement*, *10*(2), 233–256.

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

Scheerens, J. (2000). *Improving school effectiveness* (Fundamentals of educational planning, Bd. 68). Paris: Unesco, International Institute for Educational Planning.

Scheerens, J. (2004). *Review of school and instructional effectiveness research. The Paper commissioned for the EFA Global Monitoring Report 2005*

Scheerens, J. & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1. ed.). Oxford, New York: Pergamon.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499.

Seiz, J., Decristan, J., Kunter, M. & Baumert, J. (2016). Differenzielle Effekte von Klassenführung und Unterstützung für Schülerinnen und Schüler mit Migrationshintergrund. *Zeitschrift für Pädagogische Psychologie*, *30*(4), 237–249.

Senatsverwaltung für Bildung, Jugend und Familie. (n. d.a). *Berliner Schulen. Schulverzeichnis*. Zugriff am 18.03.2020. Verfügbar unter https://www.berlin.de/sen/bildung/schule/berliner-schulen/schulverzeichnis/

Senatsverwaltung für Bildung, Jugend und Familie. (n. d.b). *Bonus-Programm*. Zugriff am 19.03.2020. Verfügbar unter https://www.berlin.de/sen/bildung/unterstuetzung/bonus-programm/

Senatsverwaltung für Bildung, Jugend und Wissenschaft. (2013). *Handlungsrahmen Schulqualität in Berlin. Qualitätsbereiche und Qualitätsmerkmale*. Berlin.

Senatsverwaltung für Bildung, Wissenschaft und Forschung. (2009). *Bildung für Berlin. Handbuch Schulinspektion* (Senatsverwaltung für Bildung, Wissenschaft und Forschung, Hrsg.). Berlin.

Slavin, R. E. (1994). Quality, appropriateness, incentive, and time: A model of instructional effectiveness. *International Journal of Educational Research*, *21*(2), 141–157.

Spychiger, M., Oser, F., Hascher, T. & Mahler, F. (1999). Entwicklung einer Fehlerkultur in der Schule. In *Fehlerwelten - vom Fehlermachen und Lernen aus Fehlern* (S. 43–70).

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811.

Stringfield, S. (1994). A model of elementary school effects. In D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, E. C. Shaffer, S. Stringfield & C. Teddlie (Hrsg.), *Advances in school effectiveness research and practice* (S. 153–187). Oxford, England: Pergamon.

Teddlie, C. & Stringfield, S. (1993). *Schools make a difference. Lessons learned from a 10-year study of school effects*. New York: Teachers College Press.

Thiel, F., Hannover, B. & Pant, H. A. (2014). Nutzung und Effekte zentraler Abschlussprüfungen und standardbasierter Schulleistungstests als Instrumente der Neuen Steuerung. *Zeitschrift für Erziehungswissenschaft*, *17*(1), 3–6.

Thiel, F., Ophardt, D. & Piwowar, V. (Oktober 2013). *Kompetenzen des Klassenmanagements (KODEK). Entwicklung und Evaluation eines Fortbildungsprogramms für Lehrkräfte zum Klassenmanagement.* Freie Universität Berlin. Zugriff am 16.03.2020. Verfügbar unter https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/schulentwicklungsforschung/downloads/Abschlussbericht-KODEK.pdf

Thillmann, K. (2012). *Schulentwicklung und Schulorganisation. Eine empirische Untersuchung schulischer Organisationsgestaltung vor dem Hintergrund der Neuen Steuerung im Bildungssystem*. Dissertation. FU Berlin, Berlin.

Thrupp, M. & Lupton, R. (2006). Taking school contexts more seriously: The social justice challenge. *British Journal of Educational Studies*, *54*(3), 308–328.

Timmermans, A. C., Wolf, I. F. de, Bosker, R. J. & Doolaard, S. (2015). Risk-based educational accountability in Dutch primary education. *Educational Assessment, Evaluation and Accountability*, *27*(4), 323–346.

Trautmann, M. & Wischer, B. (2011). *Heterogenität in der Schule. Eine kritische Einführung* (Lehrbuch). Wiesbaden: VS Verlag für Sozialwissenschaften.

Van Ackeren, I. (2003). *Evaluation, Rückmeldung und Schulentwicklung. Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden* (1. Aufl.). Münster: Waxmann Verlag GmbH.

Van Ewijk, R. & Sleegers, P. J. C. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational research review*, *5*, 134–150.

Vock, M. & Gronostaj, A. (2017). *Umgang mit Heterogenität in Schule und Unterricht* (Schriftenreihe des Netzwerk Bildung, Bd. 40, 2. Auflage). Berlin: Friedrich-Ebert-Stiftung, Abt. Studienförderung.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B. & Trautwein, U. (2016). Student and teacher ratings of instructional quality. Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721.

Wang, M. C., Haertel, G. D. & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249–294.

Wilkinson, I. A. G. (2002). Introduction: peer influences on learning: where are they? *International Journal of Educational Research*, *37*(5), 395–401.

Wößmann, L. (2009). *School accountability, autonomy and choice around the world.* Cheltenham: Elgar.

Wurster, S. & Feldhoff, T. (2019). Schul- und Unterrichtsqualität aus der Mehrebenenperspektive: Ist die Schule oder die Klasse die relevante pädagogische Gestaltungseinheit? *Zeitschrift für Pädagogik*, *65*(1), 24–39.

Wurster, S. & Gärtner, H. (2013). Erfassung von Bildungsprozessen im Rahmen von Schulinspektion und deren potenzieller Nutzen für die empirische Bildungsforschung. *Unterrichtswissenschaft*, *41*(3), 217–236.

Erklärung

Hiermit versichere ich, dass ich die Dissertation „Variabilität schulischer Lernumwelten - Unterschiede in der Unterrichts- und Schulqualität zwischen Schulen und das Zusammenspiel mit Kontextmerkmalen und Schülerleistungen -“ selbstständig verfasst habe. Sämtliche Hilfsmittel, die ich verwendet habe, sind angegeben. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, April 2020

Marina Wenger

_____

Lebenslauf

Der Lebenslauf ist aus Gründen des Datenschutzes nicht enthalten.