technische universität dortmund

fakultät rehabilitationswissenschaften

Entwicklung und Erforschung
inklusiver Bildungsprozesse

Masterarbeit

Veränderungen im Direct Behavior Rating (DBR) über die Zeit

Eine Pilotierung mit fünf Messzeitpunkten in Grund- und Gesamtschulen

vorgelegt von

Sarah Hisker

Master – Lehramt für sonderpädagogische Förderung

LABG 2009

Betreuende:	Prof. Dr. Markus Gebhardt
	Prof. Dr. Jörg-Tobias Kuhn
Ausgegeben am:	11.04.2018
Eingereicht am:	25.07.2018

I. Inhaltsverzeichnis

II. Abkürzungsverzeichnis

ADHS	Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung
ANOVA	analysis of variance
α	Cronbach's Alpha
CBCL	Child Behavior Checklist
CBM	curriculum-based measurement; Curriculumbasierte Messung
DBR	Direct Behavior Rating
DBRC	Direct Behavior Report Cards
df	Anzahl der Freiheitsgrade
DVB	Direkte Verhaltensbeurteilung
bzw.	beziehungsweise
ebd.	ebenda
EP	Skala EP
EP11	Item „Wirkt besorgt, betrübt oder bedrückt"
EP12	Item „Wirkt ängstlich/ fürchtet sich"
EP13	Item „Wirkt nervös"
EXT	Skala Externalisierende Verhaltensweisen
FSP ESE	Förderschwerpunkt emotionale und soziale Entwicklung
FSP GG	Förderschwerpunkt geistige Entwicklung
FSP LE	Förderschwerpunkt Lernen
FSP SB	Förderschwerpunkt Sprache
GT	Generalisierbarkeitstheorie
HY	Skala HY
HY08	Item „Zappelt, ist (motorisch) unruhig/ überaktiv"
HY09	Item „Bricht Aufgaben häufig früh ab"
HY10	Item „Lässt sich leicht ablenken"
INT	Skala Internalisierende Verhaltensweisen
i.d.R.	in der Regel
KMK	Kultusministerkonferenz
KTT	Klassische Testtheorie
LEVUMI	Onlineplattform LEVUMI (**Le**rn**v**erla**u**fs-**M**oni**t**oring)
M	Mittelwert
MCAR	Missing completly at random
MIS	Multi-Item-Skalen
MZ	Messzeitpunkt

N	Gräße der Gesamtstichprobe
n	Größe der Teilstichprobe
PS	Skala Prosoziales Verhalten
PS14	Item „Verhält sich anderen gegenüber rücksichtsvoll“
PS15	Item „Verhält sich anderen gegenüber hilfsbereit“
PS16	Item „Verhält sich in Partner- und Gruppensituationen kooperativ“
r	Korrelationskoeffizient
r_i	Trennschärfe
RTI	Response-to-Intervention Ansatz
SD	Standardabweichung
SDO	Systematic Direct Observarion
SDQ	Strengths and Difficulties Questionnaires
SIS	Single-Item-Skalen
sog.	sogenannte
SV	Skala Schulbezogenes Verhalten
SV01	Item „Meldet sich im Unterricht“
SV02	Item „Hält sich an Gesprächsregeln“
SV03	Item „Richtet Aufmerksamkeit/Konzentration auf die Bearbeitung der Aufgabe“
SV04	Item „Arbeitet ruhig am Platz und verweigert nicht die Mitarbeit“
u.a.	unter anderem
VP	Skala Verhaltensprobleme
VP05	Item „Verhält sich wütend und aufbrausend, hat eine geringe Frustrationstoleranz“
VP06	Item „Missachtet Regeln und hört nicht auf die Lehrkraft“
VP07	Item „Streitet sich mit Mitschüler_innen/provoziert durch eigenes Verhalten seine Mitschüler_innen“
VPG	Skala Verhaltensprobleme mit Gleichaltrigen
VPG17	Item „Arbeitet/spielt meist oder lieber alleine“
VPG18	Item „Wird von Mitschüler_innen gehänselt oder geärgert, lässt sich provozieren“
VPG 19	Item „Arbeitet/spielt häufiger mit Erwachsenen als mit Mitschüler_innen“
z.B.	zum Beispiel

III. Gestaltung der Arbeit

Die vorliegende Arbeit stützt sich in ihrer formalen Gestaltung auf den „Leitfaden zur Abfassung von Prüfungsarbeiten im Fachgebiet Rehabilitationspsychologie / Psychologische Diagnostik“ von Heinrich Tröster (2018). Tabellarische Darstellungen und Abbildungen wurden ebenfalls in Anlehnung an diesen Leitfaden gestaltet.

Unter Berücksichtigung einer gendergerechten Schreibweise werden zum einen genderneutrale Formulierungen herangezogen und zum anderen wird das „Binnen-I“ als kombinierte Formulierung innerhalb eines Wortes verwendet.

Der beigefügte Datenträger (CD-ROM) umfasst sowohl die vorliegende Arbeit in Form eines PDF-Dokumentes als auch die SPSS-Datenmaske und Syntax. Des Weiteren findet sich auf dem Datenträger ein PDF-Dokument, welches alle relevanten Anlagen zusammenführt.

Die der Arbeit zugrundliegende Pilotierungsstudie wurde als Gemeinschaftsprojekt in Zusammenarbeit mit Anna Sauerland durchgeführt. Im Anschluss an die Durchführung der Studie wurden zwei eigenständige Arbeiten erstellt.

IV. Abstract

In der vorliegenden Studie wurde der Einsatz einer Direct Behavior Ratingskala im inklusiven schulischen Setting der Grund- und Gesamtschulen erprobt. Bei der Ratingskala handelt es sich um ein verlaufsdiagnostisches Instrument, welches zur Erfassung schulrelevanter Verhaltensbereiche eingesetzt werden kann. Verlaufsdiagnostische Instrumente können einen wichtigen Beitrag zur Umsetzung schulischer Inklusion bzw. inklusiver Konzepte in der allgemeinbildenden Schule leisten. An neun Grund- und Gesamtschulen setzten 39 Lehrkräfte die Ratingskala ein und beurteilten zu fünf Messzeitpunkten das Verhalten von insgesamt 205 SchülerInnen der Jahrgangsstufen 1 bis 10. In der vorliegenden Arbeit wird zum einen überprüft, inwieweit das Instrument den psychometrischen Eigenschaften der Statusdiagnostik entspricht und zum anderen erhoben, ob es sich zur Erfassung von Schülerverhalten im Verlauf eignet. Zur Überprüfung der Reliabilität und Skalierung wird in der vorliegenden Arbeit die interne Konsistenz der Skalen und die Itemtrennschärfe nach Cronbach’s Alpha berechnet. Um die Änderungssensibilität der Ratingskala über die fünf Messzeitpunkte zu untersuchen, wird die Produkt-Moment-Korrelation nach Pearson bestimmt. Des Weiteren werden mittels Experteninterviews die Nebengütekriterien der Ökonomie und Nützlichkeit für einen Einsatz im inklusiven Handlungsfeld überprüft. Die Auswertung der Ergebnisse zeigt, dass es sich bei der vorliegenden Ratingskala um eine reliable und änderrungssensible Ratingskala handelt, die im inklusiven Setting der Grund- und Gesamtschulen ökonomisch eingesetzt werden kann und eine hohe Nützlichkeit aufweist. Die Bildung einer gültigen Verrech-

nungsvorschrift ist für drei der vier psychometrischen Skalen möglich. Eine deskriptive Auswertung der Kennwerte zeigt, dass die Grund- und GesamtschülerInnen im Verlauf zunehmend angemessenere Verhaltensweisen in allen Verhaltensbereichen zeigten. Hinsichtlich der internalisierenden Verhaltensschwierigkeiten und dem prosozialen Verhalten unterschieden sich die Verhaltensveränderungen der beiden Schülergruppen signifikant voneinander. Die Skala bot den Lehrkräften eine angemessene Skalenbreite zur Beurteilung der Verhaltensveränderungen. Die Ratingskala erscheint unter den gegebenen Messbedingungen geeignet, Verhaltensveränderungen im Verlauf zu erfassen.

1. Einleitung

Die Weiterentwicklung des deutschen Schulsystems hin zu einem inklusiven Schulsystem, wie sie derzeitig basierend auf der Ratifizierung der UN-Behindertenrechtskonvention in Deutschland umgesetzt wird, führt zum einen zu neuen Bildungschancen für SchülerInnen mit Förderbedarfen jeglicher Art und zum anderen zu neuen Herausforderungen für die Lehrkräfte. Eine Herausforderung stellt die zunehmende Heterogenität der SchülerInnen einer Klasse dar. Insbesondere die Anzahl der SchülerInnen mit einem sonderpädagogischen Förderbedarf in der emotionalen und sozialen Entwicklung nimmt zu und damit auch die Anzahl der SchülerInnen mit herausfordernden Verhaltensweisen (Autorengruppe Bildungsberichterstattung, 2018). Die Grund- und Sekundarschullehrkräfte stehen unter diesen Bedingungen vor der Herausforderung inklusive Konzepte in der allgemeinbildenden Schule zu etablieren, die traditionell leistungsorientierte Konzepte und Selektions- sowie Homogenitätspraktiken fokussiert (Amrhein, 2015). Es gilt exkludierende Strukturen und Prozesse zu reduzieren und einen gemeinsamen Unterricht zu initiieren (Hennemann, Ricking & Huber, 2018). Ein Blick auf Länder, die inklusive Konzepte fortschrittlich umsetzen, zeigt das Potential gestufter Fördersysteme (wie z.B, dem Response-to-Intervention Ansatz (RTI)), die auf eine Prävention von Lern- und Entwicklungsschwierigkeiten und eine individualisierte Förderung abzielen (Huber & Rietz, 2015). Der Aspekt der Prävention ist dabei von zentraler Bedeutung. Präventive Maßnahmen sollten eingesetzt werden, um der Entstehung von Verhaltensstörungen und negativen Entwicklungstendenzen entgegenwirken zu können, denn bestehende oder manifestierte Verhaltensstörungen sind deutlich schwieriger zu behandeln. Da die allgemeinbildende Schule für alle SchülerInnen zugänglich ist, stellt sie das wichtigste Präventionssetting dar (Beelmann, 2008). SchülerInnen mit bestehenden Verhaltensstörungen oder potentiell gefährdete SchülerInnen können identifiziert werden und es können präventive Fördermaßnahmen eingeleitet werden. Diesbezüglich sind Instrumente notwendig, die es den Lehrkräften ermöglichen im Sinne einer evidenzbasierten pädagogischen Praxis das Verhalten der SchülerInnen im Verlauf zu erfassen. Traditionelle statusdiagnostische Instrumente sind diesbezüglich nur bedingt geeignet, da sie Verhaltensausprägungen lediglich punktuell erheben können. Das Auftreten bestimmter Verhaltensweisen hängt aber im besonderen Maße von spezifischen Faktoren, wie z.B. den Situationsbedingungen oder den Interaktionen der SchülerInnen mit Lehrkräften und MitschülerInnen ab. Entsprechende Ereignisse können zu rapiden Verhaltensveränderungen führen (Dever, Dowdy, Raines & Carnazzo, 2015). Dementsprechend sind Instrumente notwendig, die Verhaltensveränderungen im Verlauf erheben können. Bislang stand vornehmlich die Lernverlaufsdiagnostik, also die verlaufsdiagnostische Erhebung akademischer Leistungen, im Vordergrund der Forschung und der pädagogischen Praxis (Voß & Gebhardt, 2017b). Zahlreiche Forschungsarbeiten konnten diesbezüglich aufzeigen, dass sich kontinuierliche Rückmeldun-

gen zu Lern- und Entwicklungsständen der SchülerInnen positiv auf die Auswahl und Anpassung eingesetzter Fördermaßnahmen sowie das unterrichtliche Handeln der Lehrkräfte auswirkt, was wiederum einen positiven Einfluss auf die Entwicklung der SchülerInnen hat (vgl. z.B. Hattie, Beywl & Zierer, 2013; Kingston & Nash, 2011). Neben der Lernverlaufsdiagnostik findet zunehmend die Verhaltensverlaufsdiagnostik Berücksichtigung in der Forschung (Voß & Gebhardt, 2017a). Verhaltensstörungen stellen im schulischen Setting insbesondere in Form von externalisierenden, nach außen gerichteten Verhaltensstörungen eine Belastung für Lehrkräfte und Mitschüler dar und können einen negativen Einfluss auf die schulische Entwicklung der SchülerInnen haben. Sie bedürfen aufwändiger Präventions- und Fördermaßnahmen, die auf Grundlage verlaufsdiagnostischer Instrumente frühzeitig evaluiert werden und bei mangelnder Passung zwischen den Maßnahmen und den individuellen Bedürfnissen zeitnah modifiziert werden können (Huber & Rietz, 2015). Verlaufsdiagnostische Instrumente bieten dementsprechend ein großes Potential für die Arbeit mit SchülerInnen mit Verhaltensauffälligkeiten und unter dem Aspekt der Prävention ein großes Potential für die Arbeit mit SchülerInnen, die potentiell von Verhaltensauffälligkeiten gefährdet sind. Im englischen Sprachraum wurde diesbezüglich ein neuer vielversprechender Ansatz entwickelt, das Direct Behavior (DBR) (Christ, Riley-Tillman & Chafouleas, 2009). Es zielt darauf ab, Verhalten direkt, also unmittelbar nach dem Auftreten zu erfassen und mithilfe von Ratingskalen zu beurteilen, sodass ein ökonomischer Einsatz im schulischen Setting möglich ist. Die Eignung von DBR-Instrumenten für verlaufsdiagnostische Zwecke konnte sowohl international als auch national nachgewiesen werden (vgl. z.B. Casale, Grosche, Volpe & Hennemann, 2017; Christ et al., 2009).

Im deutschsprachigen Raum liegen vergleichsweise wenige evaluierte Instrumente vor (Voß & Gebhardt, 2017a). Im Rahmen des Forschungsprojektes „LEVUMI“, einer Onlineplattform zur Lernverlaufsdiagnostik der Technischen Universität Dortmund (Prof. Dr. Markus Gebhardt), der Europa-Universität Flensburg (Prof. Dr. Kirsten Diehl) und der Universität Kiel (Prof. Dr. Andreas Mühling) wurde deswegen eine Direct Behavior Ratingskala entwickelt, welche zur Erfassung schulbezogenen Verhaltens, externalisierender Verhaltensweisen (Verhaltensprobleme und Hyperaktivität), internalisierende Verhaltensweisen (Emotionale Probleme und Verhaltensprobleme mit Gleichaltrigen) und prosoziales Verhaltens eingesetzt werden soll. Im Zuge ihrer Masterarbeit entwickelte Sauerland (i.D.) die Ratingskala qualitativ weiter, woraufhin im Rahmen der vorliegenden Pilotierungsstudie eine Erprobung der überarbeiteten Ratingskala im inklusiven Setting der Grund- und Gesamtschulen durchgeführt wurde.

In der vorliegenden Arbeit soll überprüft werden, ob es sich bei der entwickelten Ratingskala um eine reliable und änderungssensible Ratingskala handelt, die Verhaltensveränderungen der SchülerInnen im Verlauf erfassen kann. Desweitern soll untersucht werden, inwieweit ein Einsatz der Ratingskala im inklusiven Setting der Grund- und Gesamtschule möglich ist. Eine

Evaluation der Ratingskala hinsichtlich der Nebengütekriterien Ökonomie und Nützlichkeit kann dabei wichtige Anhaltspunkte liefern.

Zur Einführung in die Thematik werden im Kapitel 2 zunächst schulrelevante Verhaltensweisen unter Berücksichtigung der schulischen Inklusion dargelegt. Im Anschluss daran wird die Thematik der Verlaufsdiagnostik (Kapitel 3) unter Berücksichtigung der zwei Ansätze Lernverlaufsdiagnostik (Kapitel 3.1) und Verhaltensverlaufsdiagnostik (Kapitel 3.2) vorgestellt, bevor abschließend zwei Methoden der Verhaltensdiagnostik erläutert werden, die systematische direkte Verhaltensbeobachtung (Kapitel 3.2.1) und die Verhaltensbeurteilung mit Ratingskalen (Kapitel 3.2.2). Im Anschluss daran wird im Kapitel 4 das Screeninginstrument Strengths and Difficulties Questionnaire (SDQ) vorgestellt. Als eine Kombination der systematischen direkten Verhaltensbeobachtung und der Verhaltensbeurteilung mit Ratingskalen wird im darauffolgenden Kapitel 5 die neue vielversprechende verlaufsdiagnostische Methode des Direct Behavior Ratings präsentiert. Kapitel 6 umfasst den empirischen Teil dieser Arbeit. Zu Beginn werden wird die Fragestellungen dieser Arbeit aufgeführt (Kapitel 6.1) und zur Einordnung der Studie das übergeordnete Forschungsprojekt LEVUMI dargelegt (Kapitel 6.2). Kapitel 6.3 umfasst das methodische Vorgehen. Es wird das Studiendesign (Kapitel 6.3.1), die Operationalisierung (Kapitel 6.3.2) sowie der Aufbau der Erhebungsmethoden (Kapitel 6.3.3), die Durchführung der Studie im schulischen Setting (Kapitel 6.3.4) und die Stichprobe (Kapitel 6.3.5) erläutert. Im Anschluss daran findet sich im Kapitel 6.4 die Darstellung der Ergebnisse der fünf Forschungsfragen. Im Kapitel 6.5 werden die Ergebnisse diskutiert und in den aktuellen Forschungsstand eingeordnet. Abschließend werden im Kapitel 7 die Forschungsergebnisse dieser Studie zusammengefasst und es wird ein Ausblick auf zukünftige Forschungsarbeiten gegeben.

2. Verhalten in der inklusiven Schule

Im Jahr 2009 hat Deutschland die UN-Behindertenrechtskonvention ratifiziert. Gemäß des Artikels 24 muss der deutsche Staat sicherstellen, dass kein Schüler und keine Schülerin vom allgemeinen Bildungssystem ausgeschlossen wird (Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderungen, 2017). Der Grundgedanke dieses Übereinkommens ist die „Inklusion“ (ebd.). Der Begriff der Inklusion beschreibt im schulischen Setting zum einen die Platzierung eines Kindes mit sonderpädagogischem Förderbedarf in der Regelschule und zum anderen ein pädagogisches Konzept (Gebhardt, Sälzer & Tretter, 2014). Dieses Konzept berücksichtigt nicht nur das Kind mit sonderpädagogischen Förderbedarf, sondern alle Kinder einer Klasse mit besonderen pädagogischen Bedürfnissen (Sander, 2006). Es stellt eine Erweiterung und Optimierung des Konzeptes der Integration dar, wobei Integration die Aufnahme von SchülerInnen mit Behinderungen in allgemeine Schulen meint (ebd.). Die Entwicklungen des deutschen Schulsystems hin zu einem inklusiven Schulsystem bringen umfassende Veränderungen mit sich (Gebhardt, Diehl & Mühling, 2015). Gebhardt et al.

(2014) konnten aufzeigen, dass in allen Bundesländern mit geringen Unterschiedenen inklusive Konzepte implementiert und umgesetzt werden. Ziel dieser Entwicklungen ist das Angebot eines inklusiven Unterrichts für alle SchülerInnen (Gebhardt, 2015). Im Sinne der schulischen Inklusion sind exkludierende Bedingungen und Prozesse abzubauen und es ist ein gemeinsamer Unterricht von SchülerInnen mit und ohne Förderbedarf im Klassenverband umzusetzen (Hennemann et al., 2018; Koch & Textor, 2015). Des Weiteren sind gestufte Fördersysteme aufzubauen (Huber & Rietz, 2015). Gestufte Fördersysteme konnten sich bereits in vielen Ländern, die Inklusion fortschrittlich umsetzen, bewähren (ebd.). Sie unterstützten auf Basis verlaufsdiagnostischer Instrumente die Individualisierung der Förderung der SchülerInnen (ebd.). Ein Beispiel eines solchen gestuften Fördersystems stellt der RTI-Ansatz dar, welcher darauf abzielt eine Passung zwischen den Bedürfnissen der SchülerInnen und den individuellen Fördermaßnahmen herzustellen (Casale, Hennemann, Huber & Grosche, 2015a; Huber & Rietz, 2015). Mithilfe verlaufsdiagnostischer Instrumente kann die individuelle Entwicklung der SchülerInnen erfasst werden und so die Effektivität (response) der individuellen Fördermaßnahmen (intervention) bestimmt werden (Casale et al., 2015a). Im Fokus inklusiver Konzepte steht nicht ein selektionspädagogischer, sondern ein inklusionspädagogischer Ansatz (Huber & Rietz, 2015). Ziel dieser Konzepte ist nicht die „Ettikettierung“ von SchülerInnen, sondern die Prävention von Lern- und Entwicklungsschwierigkeiten und eine individualisierte Förderung (ebd.). In diesem Zuge verändern sich die Anforderungen an den gemeinsamen Unterricht (Balt, Ehlert & Fritz, 2017). Individuelle Lern- und Entwicklungsverläufe sind stärker bei der Gestaltung des Unterrichts und der Planung einer Förderung zu berücksichtigen (ebd.). Die Gestaltung der schulischen Settings erfolgt dabei dimensional mit fließenden Übergängen und in Abhängigkeit zu verschiedenen Faktoren wie den Bedürfnissen der SchülerInnen oder den Anforderungen der Schulfächer (Gebhardt et al., 2014). Es ist in erster Linie ein gemeinsamer Unterricht aller SchülerInnen zu gewährleisten. Jedoch können auch Settings der äußeren Differenzierung gewählt werden, wie z.B. ein Einzelunterricht, sofern diese Settings allen SchülerInnen zur Verfügung stehen unabhängig des Leistungsstandes und sofern sie nicht mit einer Herabsetzung verbunden werden (Koch & Textor, 2015).

Die inklusive Beschulung von zunehmend heterogenen Schulklassen stellt für die Lehrkräfte der allgemeinbildenden Schulen eine große Herausforderung dar (Gebhardt et al., 2015a). Dabei ist zu berücksichtigen, dass mangelnde Kompetenzen und Erfahrungen der Lehrkräfte im Hinblick auf eine inklusive Beschulung von SchülerInnen zum Scheitern eines gemeinsamen Unterrichts führen können (Autorengruppe Bildungsberichterstattung, 2018). Mangelnde Kompetenzen können z.B. aufgrund einer marginalen Behandlung von Inklusionsthemen in der LehrerInnenausbildung vorliegen (Amrhein, 2015). Inklusion wird in nicht-sonderpädagogischen LehrerInnenausbildungen i.d.R. nur additiv in einzelnen Seminaren thematisiert

(ebd.). Für die TU Dortmund konnten Rütter und Lühn (2017) für den Studiengang Gymnasium/Gesamtschule lediglich eine Anzahl von 16 inklusiv-didaktischen Veranstaltungen an 368 Veranstaltungen des Studiengangs (4,4%) feststellen. Die Anzahl inklusiv-didaktischer Veranstaltungen im Studiengang Grundschullehramt liegt mit 17 Veranstaltungen an insgesamt 198 Veranstaltungen (8,6%) deutlich höher (ebd.). Die höchste Anzahl von inklusiv-didaktischen Veranstaltungen konnten die Autorinnen für den Studiengang der Sonderpädagogik feststellen: 21 inklusiv-didaktische Veranstaltungen an 270 gesamten Veranstaltungen (7,8%) (ebd.). Diese Befunde bestätigen, dass angehende Sekundarschullehrkräfte im geringeren Maße Zugang zu inklusiv-didaktischen Lehrinhalten haben. Die Lehrkräfte der allgemeinbildenden Schulen müssen im Hinblick auf die Inklusion bessere Ausbildungs- und Weiterbildungsmöglichkeiten erhalten, um die notwendigen sonderpädagogischen Kompetenzen erwerben zu können (Autorengruppe Bildungsberichterstattung, 2018). Dies ist insbesondere für Lehrkräfte von Schulformen mit geringen Inklusionsanteilen, wie z.B. der Gesamtschule, von Bedeutung (ebd.). Lehrkräfte der Sekundarschule sind primär mit Problemen auf der Umsetzungsebene konfrontiert, denn im Zuge einer inklusiven Weiterentwicklung des Sekundarschulsystems entsteht zwischen Grundgedanken der Inklusion und der Leistungsorientierung, der Selektionsfunktion sowie den Homogenitätspraktiken dieser Schulformen ein enormes Spannungsfeld (Amrhein, 2015). Eine inklusive Beschulung ist jedoch für alle SchülerInnen mit sonderpädagogischen Förderbedarf von großer Bedeutung, da sie im gemeinsamen Unterricht bessere schulische Leistungen zeigen als an Förderschulen oder in separierten Klassen (Gebhardt, 2015). Auch die schulischen Leistungen der SchülerInnen ohne sonderpädagogischen Förderbedarf werden nicht negativ beeinflusst (ebd.). Insgesamt weisen 7,1% der SchülerInnen einen diagnostischen Förderbedarf auf (Autorengruppe Bildungsberichterstattung, 2018). Von diesen SchülerInnen wurde im Schuljahr 2016/17 ein Anteil von 39% inklusiv beschult (ebd.). Dabei ist der Anteil der FörderschülerInnen in der Grundschule höher als in der Sekundarstufe (Autorengruppe Bildungsberichterstattung, 2018; Bertelsmann-Stiftung, 2015). Während der Inklusionsanteil in Grundschulen 46,9% beträgt, liegt dieser in der Gesamtschule lediglich bei 33,4% (Schuljahr 2013/14) (Bertelsmann-Stiftung, 2015). Entsprechende Ergebnisse wurden auch in dem Bildungsbericht zum Schuljahr 2016/17 berichtet: In 44% aller Klassen der vierten Jahrgangsstufe und in 17% aller Klassen der neunten Jahrgangsstufe werden SchülerInnen mit sonderpädagogischem Förderbedarf gemeinsam beschult (Autorengruppe Bildungsberichterstattung, 2018). FörderschülerInnen finden sich in der Sekundarstufe am häufigsten in Haupt- und Gesamtschulen (Bertelsmann-Stiftung, 2015).

Insgesamt nimmt die Heterogenität der SchülerInnen in allgemeinbildenden Schulklassen weiter zu. Im Vergleich zu den Vorjahren ist die Förderquote angestiegen (Autorengruppe Bildungsberichterstattung, 2018; Bertelsmann-Stiftung, 2015; Gebhardt et al., 2015a). Im För-

derschwerpunkt emotionale und soziale Entwicklung hat sich die Förderquote mehr als verdoppelt (Autorengruppe Bildungsberichterstattung, 2018). Tendenziell zeichnet sich für diese Schülergruppe eine positive Prognose der inklusiven Beschulung ab (Ellinger & Stein, 2012). Es ist jedoch zu berücksichtigen, dass nur wenige Forschungsbefunde vorliegen und die Schülerschaft eine große Heterogenität aufweist (ebd.). SchülerInnen mit einem Förderbedarf im Bereich der emotionalen und sozialen Entwicklung weisen häufig einen erheblichen Leistungsrückstand und psychische Verhaltensauffälligkeiten auf, die mit starken Ausprägungen im Verhalten einhergehen (Gebhardt, 2015; Link 2018). Psychische Auffälligkeiten, wie z.B. externalisierende Verhaltensstörungen, stellen große Herausforderungen und Belastungen für die Lehrkräfte, die MitschülerInnen und die Familien der betreffenden SchülerInnen dar (Hölling, Schlack, Petermann, Ravens-Sieberer & Mauz, 2014; Kern et al., 2015; Makarova, Herzog & Schönbächler, 2014; Volpe & Fabiano, 2013). Verhaltensstörungen belasten zudem die Betroffenen selbst (Hölling et al., 2014). Sie können die individuellen Lernerfolge und Entwicklungen der SchülerInnen und der MitschülerInnen gefährden sowie die Unterrichtsprozesse und –abläufe negativ beeinträchtigen (Casale, 2017; Forness, Kim & Walker, 2012).

Dem Begriff der Verhaltensstörung liegt keine eindeutige Bestimmung zugrunde (Hillenbrand, 2008a). Zahlreiche Begriffe wie „psychische Störungen“, „emotionale Störungen“ und „Verhaltensauffälligkeiten“ werden synonym verwendet (Hillenbrand, 2008a; Seitz & Stein, 2010). Von zentraler Bedeutung ist es die beiden Begriffe „Verhaltensstörung“ und „psychische Störung“ zu unterscheiden. Üblicherweise bezeichnet der erste Begriff der Verhaltensstörung allgemeine Auffälligkeiten im psychischen Bereich, die ohne Krankheitswert einhergehen. Der zweite Begriff wirrd bei Erkrankungen verwendet, die mithilfe internationaler Klassifikationssysteme diagnostiziert werden (Barkmann & Schulte-Markwort, 2007). Im schulischen Kontext ist der Begriff „Förderschwerpunkt emotionale und soziale Entwicklung“ zu finden (Hillenbrand, 2008a; Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2000). Er wird in den Empfehlungen zum Förderschwerpunkt emotionale und soziale Entwicklung der Kultusministerkonferenz (KMK, 2000) wie folgt umschrieben:

> „Sonderpädagogischer Förderbedarf ist bei Kindern und Jugendlichen mit Beeinträchtigungen der emotionalen und sozialen Entwicklung, des Erlebens und der Selbststeuerung anzunehmen, wenn sie in ihren Bildungs-, Lern- und Entwicklungsmöglichkeiten so eingeschränkt sind, dass sie im Unterricht der allgemeinen Schule auch mit Hilfe anderer Dienste nicht hinreichend gefördert werden können.“ (S. 10).

Dieser Begriff sollte aufgrund der hohen Relevanz für die schulische Praxis berücksichtigt werden (Hillenbrand, 2008b). Es wird ein sonderpädagogischer Förderbedarf in Orientierung an Kriterien des allgemeinbildenden Schulsystems zugewiesen (ebd.). Im Rahmen dieser Definition wird die enge Verknüpfung von emotionalem Erleben und sozialem Handeln deutlich

(Methner & Popp, 2017).

Verhaltensstörungen unterliegen vielschichtigen Wechselwirkungen zwischen Variablen des Kindes oder Jugendlichen und seiner Umwelt (Linderkamp & Grünke, 2007). Kontextuelle Faktoren stellen dabei sogenannte „setting events" dar, die die Auftretenswahrscheinlichkeit von spezifischen, problematischen Interaktionen zwischen Schülern, Lehrern und Gleichaltrigen potentiell erhöhen oder verringern können (Fox & Conroy, 1995, S. 130). Entsprechende Ereignisse können zu rapiden Veränderungen im Verhalten der SchülerInnen führen (Dever et al., 2015). Verhaltensstörungen entsprechen folglich Reaktionen auf bestimmte Situationen, Reize und Interaktionsformen und sind abhängig von den Normen und den gesellschaftlichen und persönlichen Erwartungen der Bezugspersonen (Linderkamp & Grünke, 2007). Des Weiteren ist es von Relevanz, inwieweit die betreffenden SchülerInnen selbst das Verhalten als angemessen oder unangemessen einschätzen (Linderkamp, 2007). Verhaltensweisen werden subjektiv und im Hinblick auf bestimmte schulische Settings als angemessen oder störend empfunden (Voß & Gebhardt, 2017a). Dementsprechend müssen Beurteilungen von Verhaltensweisen immer entwicklungsbezogen und unter Berücksichtigung des spezifischer schulischer Kontexte erfolgen (Linderkamp, 2007; Voß & Gebhardt, 2017a). Angestrebte Verhaltensziele werden im schulischen Setting nicht auf Basis schulorganisatorischer Regelungen, sondern mit Bezug zum jeweiligen situativen Kontext festgelegt (Voß & Gebhardt, 2017a). Die Verhaltensstörungen treten in vielfältiger Weise und in unterschiedlichen Erscheinungsformen auf (Casale et al., 2017). SchülerInnen mit Verhaltensstörungen stellen folglich eine heterogene Gruppe dar (Casale et al., 2017; Conroy, Stichter, Daunic & Haydon, 2008). Sie zeigen Verhaltensweisen entlang eines Kontinuums von extremen externalisierenden bis zu extremen internalisierenden Verhaltensstörungen (Lewis, 2016). Die Unterscheidung zwischen externalisierenden und internalisierenden Verhaltensstörungen basiert auf zahlreichen gesicherten empirischen Untersuchungen (Myschker & Stein, 2014). Unter externalisierendem Verhalten wird ein „aggressiv-ausagierendes" Verhalten verstanden (ebd., S. 58). Es ist nach außen, gegen die Umwelt gerichtet (ebd.). Internalisierendes Verhalten entspricht hingegen einem „ängstlich-gehemmten" Verhalten, welches selbstbeeinträchtigend wirkt (ebd., S. 58). SchülerInnen mit externalisierenden Verhaltensweisen zeigen u.a. aggressive, impulsive, und regelverletzende Verhaltensweisen sowie Hyperaktivität, Konzentrations- und Aufmerksamkeitsschwierigkeiten (ebd.). SchülerInnen mit internalisierenden Verhaltensweisen zeigen demgegenüber Ängstlichkeit, Gehemmtheit sowie psychosomatische Störungen. Sie sind traurig, freudlos und ziehen sich zurück (ebd.). Internalisierende Verhaltensstörungen fallen im Vergleich zu externalisierenden Verhaltensstörungen weniger auf und werden seltener als störend empfunden (Blumenthal, Hartke & Vrban, 2017). Internalisierende Störungen sollten dennoch nicht in ihrer Bedeutung für die Entwicklung von Kindern und Jugendlichen unterschätzt werden (Ihle & Esser, 2002; Myschker & Stein, 2014). Beide Verhaltensstörungen können mit

psychologischen Entwicklungsproblemen, wie Substanzkonsum, Delinquenz oder Gewaltbereitschaft einhergehen, die schulischen Leistungen der Kinder und Jugendlichen beeinträchtigen und sich negativ auf schulische Lernerfolge auswirken (DeVries, Gebhardt & Voß, 2017; Haller et al., 2016).

Zwei weitere Gruppen von Kindern und Jugendlichen mit Verhaltensstörungen, sind Kinder und Jugendliche mit sozial-unreifem und sozialisiert-delinquentem Verhalten (Myschker & Stein, 2014). Erstere verhalten sich nicht altersentsprechend, sind leicht ermüdbar und weisen Sprach- und Sprechstörungen auf (ebd.). Kinder und Jugendliche mit sozialisiert-delinquentem Verhalten handeln verantwortungslos und risikobereit, missachten Normen und weisen niedrige Hemmschwellen sowie Beziehungsstörungen auf (ebd.).

Grundsätzlich ist der Bereich des Verhaltens für allgemeine Lernerfolge von besonderer Bedeutung (Chafouleas, Riley-Tillman & Christ, 2009a; Linderkamp & Grünke, 2007). Im schulischen Setting sind insbesondere Verhaltensweisen von Relevanz, die den Lernerfolg der SchülerInnen unmittelbar beeinflussen (Chafouleas, Sanetti, Kilgus & Maggin, 2012a; Voß & Gebhardt, 2017a). Nachfolgend werden mit Bezug zu den klassifizierten Schülergruppen Verhaltensbereiche erläutert, die für das schulische Setting und diese Arbeit von hoher Relevanz sind: Schulbezogenes Verhalten, Verhaltensprobleme, Hyperaktivität, Emotionale Probleme, Prosoziales Verhalten und Verhaltensprobleme mit Gleichaltrigen. Dabei ist zu beachten, dass die verwendeten Verhaltensbereiche in der Forschungsliteratur nicht eindeutig definiert sind und folglich immer unter Berücksichtigung des Forschungs- und Anwendungsbereiches betrachtet werden müssen (Casale et al., 2015a; Jurkowski & Hänze 2014).

Unter schulbezogenem Verhalten werden in dieser Studie Arbeits- und Sozialverhaltensweisen verstanden, die für den schulischen Kontext von Relevanz sind und ein erfolgreiches schulisches Lernen ermöglichen (DiPerna & Elliott, 2002; Henning, Schramm & Linderkamp, 2017; Lohbeck, Petermann & Petermann, 2015a). DiPerna und Elliott (2002) prägten diesbezüglich den Begriff „academic enablers“ (DiPerna & Elliott, 2002, S. 293). „Academic enablers“ sind Verhaltensweisen und Einstellungen, die Lernprozesse im Unterricht ermöglichen und unterstützen (z.B. Engagement, Motivation, Lernfertigkeiten) (ebd.). Beispiele für das Arbeits- und Sozialverhalten im schulischen Kontext sind die aktive Beteiligung am Unterricht, die konzentrierte Arbeit an Aufgaben und die Einhaltung von Regeln (Casale, Hennemann, Volpe, Briesch & Grosche, 2015c; Henning et al., 2017).

Der Bereich der Verhaltensprobleme wird in Anlehnung an Goodman et al. (2010) den externalisierenden Verhaltensstörungen zugeordnet. Dieser Bereich umfasst oppositionelle und aufsässige Verhaltensstörungen und greift nach außen gerichtete Verhaltensstörungen auf, die sich gegen die Lehrkraft und MitschülerInnen richten (Goodman, Lamping & Ploubidis, 2010) (Myschker & Stein, 2014).

Der Verhaltensbereich Hyperaktivität ist ebenfalls den externalisierenden Verhaltensstörungen zuzuordnen (Goodman et al., 2010). Dieser Bereich greift die „Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS)“ (Goodman et al., 2010; Steinhausen, 2016) auf. Die drei Leitsymptome dieser Störung sind Unaufmerksamkeit, Hyperaktivität und Impulsivität (Steinhausen, 2016).

Emotionale Probleme können als internalisierende Verhaltensstörungen klassifiziert werden (Goodman et al., 2010). Diese Verhaltensstörungen umfassen Angststörungen und depressive Störungen (ebd.). Symptome einer Angststörung können in zahlreichen Ausprägungen auftreten. Als Beispiele können Anspannung, Zittern, Erregtheit, Erstarren, Selbstzweifel, Besorgtheit, Stottern, Weinen oder Schreien genannt werden (Nevermann, 2008). Merkmale von depressiven Störungen sind u.a. reduzierte körperliche oder psychische Aktivitäten, Energielosigkeit, verlangsamte Sprache oder Denkprozesse sowie Konzentrationsprobleme (Reicher & Rossmann, 2008). Hier zeigt sich, dass sich Emotionen aus Ebenen zusammensetzen: Zum einen der subjektiven Ebene, den erlebten Gefühlen und zum anderen der objektiven beobachtbaren Ebene, dem Ausdruck bzw. der körperlichen Komponente (z.B. dem Gesichtsausdruck) (Hennemann, Hövel, Casale, Hagen & Fitting-Dahlmann, 2017).

Ein weiterer für diese Studie relevanter Verhaltensbereich ist das Prosoziale Verhalten. Konkrete prosoziale Verhaltensweisen operationalisieren hierbei das Konstrukt der sozialen Kompetenz (Beelmann & Raabe, 2007). Im Fokus stehen Fertigkeiten zur Bildung positiver Sozialbeziehungen, wie z.B. das Anbieten von Hilfeleistungen (ebd.). Prosoziale Verhaltensweisen basieren auf dem Aspekt der Freiwilligkeit und erfolgen unabhängig von jeglicher Aufgabenstellung (Bierhoff, 2010). DeVries, Rathmann & Gebhardt (2018) konnten für ausgeprägte prosoziale Verhaltensweisen einen leicht positiven Einfluss auf die Benotung feststellen (DeVries, Rathmann & Gebhardt, 2018).

Verhaltensprobleme mit Gleichaltrigen werden in dieser Arbeit ebenfalls in Anlehnung an Goodman et al. (2010) der Dimension der internalisierenden Verhaltensstörungen zugeordnet. Sie äußern sich in Störungen des Sozialkontaktes, wie der Ablehnung durch Gleichaltrige, in unreifem und erwachsenenabhängigen Sozialverhalten (Döpfner & Petermann, 2012; Lohbeck, Nitkowski, Petermann & Petermann, 2014a). Verhaltensprobleme mit Gleichaltrigen sind im schulischen Kontext von Relevanz, da sie einen starken negativen Einfluss auf akademische Leistungen haben (DeVries et al., 2018).

Aus entwicklungspsychologischer Sicht sind Verhaltensweisen im schulischen Setting vor dem Hintergrund der unterschiedlichen Bezugsrahmen der jeweiligen Schulstufen zu betrachten (Voß & Gebhardt, 2017a). GrundschülerInnen müssen angemessene schulbezogene Verhaltensweisen erst erlernen und einüben, hingegen werden diese Verhaltensweisen im Setting der Sekundarstufe bereits vorausgesetzt (ebd.). SchülerInnen der Sekundarstufe sind somit

seit mehreren Jahren mit den Anforderungen und Erwartungen des schulischen Settings konfrontiert (Chafouleas et al., 2010).

Rund 10-20% der Kinder und Jugendlichen sind in Deutschland von psychischen Auffälligkeiten betroffen (Robert Koch-Institut, 2018a; Klasen, Meyrose, Otto, Reiss & Ravens-Sieberer, 2017; Petermann, 2005; Barkmann & Schulte-Markwort, 2007; Hölling et al., 2014). Die unterschiedlichen Angaben zur Prävalenz sind auf die Vielfalt der verwendeten Erhebungsinstrumente, Klassifikationsverfahren sowie auf die verschiedenen Altersbereiche der jeweiligen Stichproben zurückzuführen (Hölling et al., 2014; Petermann, 2005) (Petermann, 2005). Die Prävalenz psychischer Störungen variiert zwischen den unterschiedlichen Altersstufen (Costello, Mustillo, Erkanli, Keeler & Angold, 2003; Hölling et al, 2014; Ravens-Sieberer et al., 2008). In der dritten BELLA-Welle der KiGGS-Studie konnte für die 7- bis 10-Jährigen ein Prävalenzanteil psychischer Auffälligkeiten von 19,8% festgestellt werden. Etwas höher fällt der Anteil (22%) bei den 11- bis 13-Jährigen aus (Klasen et al., 2017). Wiederum niedriger ist der Anteil mit 17,3% bei den 14- bis 17-Jährigen (ebd.). Erlangen SchülerInnen mit Verhaltensschwierigkeiten das Alter der Pubertät, sind sie doppelt belastet (Wettstein, Bryjová, Faßnacht & Jakob, 2011). Neben bereits vorliegenden Verhaltensstörungen müssen sie zunehmend komplexere Entwicklungsaufgaben bewältigen (Robert Koch-Institut, 2018b; Wettstein et al., 2011). Ferner konnte die KiGGS-Studie Unterschiede zwischen den Geschlechtern erheben. In der Altersgruppe der 7- bis 13-Jährigen weisen Jungen einen höheren Anteil psychischer Auffälligkeiten auf als Mädchen (Klasen et al., 2017). In der Altersgruppe der 14- bis 17-Jährigen ist hingegen für die Mädchen ein höherer Anteil feststellbar (ebd.). Psychische Auffälligkeiten entstehen bei Jungen vermehrt in der Entwicklungsphase vom Vorschulalter bis zu Ende der Grundschulzeit (Robert Koch-Institut, 2018a). Der Anteil psychisch auffälliger Jungen reduziert sich dann mit zunehmenden Alter (ebd.). Hingegen nimmt der Anteil der Mädchen mit persistierenden psychischen Auffälligkeiten mit zunehmender Zeit zu (ebd.). Psychische Auffälligkeiten entstehen bei Mädchen vermehrt ab dem Ende der Grundschule bis zum Ende des Jungendalters (ebd.). Insgesamt kann für Jungen eine höhere Prävalenzrate verzeichnet werden als für Mädchen und zudem weisen sie häufiger externalisierende Verhaltensauffälligkeiten auf (Costello et al., 2003; Ihle & Esser, 2002, 2008; Robert Koch-Institut, 2018a). Mädchen zeigen hingegen häufiger internalisierende Verhaltensweisen (Robert Koch-Institut, 2018a).

Eine einzelne Betrachtung der Verhaltensbereiche zeigt auf, dass emotionalen Probleme und Verhaltensprobleme mit Gleichaltrigen über drei Altersbereiche (7- bis 17-Jährigen) am geringsten ausgeprägt sind (Hölling et al., 2014). Erstere treten im Vergleich zur Adoleszenz labiler und kürzer auf (Robert Koch-Institut, 2018a; Steinhausen 2016). Verhaltensprobleme traten ebenfalls in schwacher Ausprägung auf, wohingegen Hyperaktivitätsprobleme mit deutlich höheren Werten einhergingen (Hölling et al., 2014). Letztere waren bei 14- bis 17-Jährigen

seltener zu beobachten als bei 7- bis 13-Jährigen (ebd.). Eine ähnliche Verteilung über die Altersspannen zeigt sich beim prosozialen Verhalten (ebd.). Die 7- bis 13-Jährigen zeigten ein ausgeprägteres prosoziales Verhalten als die 14- bis 17-Jährige (ebd.).

Die hohen Anteile von SchülerInnen mit psychischen Störungen gehen mit hohen Persistenzraten einher (Ihle & Esser, 2002; 2008). Es besteht somit das Risiko der Manifestation von Verhaltensstörungen und der Ausbildung zusätzlicher Beeinträchtigungen (Brezinka, 2003). Anfängliche Verhaltensstörungen sollten möglichst frühzeitig erfasst werden, damit ungünstigen Entwicklungsverläufen und Manifestationen durch angemessene Maßnahmen entgegengewirkt werden kann (Petermann & Lehmkuhl, 2010). Dabei gilt, dass es leichter ist negativen Entwicklungen entgegenzuwirken als manifestierte Störungen zu behandeln (Petermann & Lehmkuhl, 2010). Darüber hinaus sinken die Behandlungsaussichten für Kinder und Jugendliche mit zunehmenden Alter (Landscheidt, 2001). Diese Aspekte verdeutlichen die Bedeutsamkeit präventiven Handelns (Brezinka, 2003) . Ein konsequenter Einsatz von Präventionsmaßnahmen kann Risikofaktoren reduzieren (Hennemann et al., 2018). Neben Kontextfaktoren, wie den zuvor erläuterten „setting events“ stellen Erziehungsfaktoren (wie ineffektive Erziehungspraktiken), Faktoren des Kindes (wie ein schwieriges Temperament oder negative Emotionalität) und Peerfaktoren (wie die Ablehnung durch Gleichaltrige) Risikofaktoren dar (Scheithauer, Mehren & Petermann, 2003; Wiedebusch & Petermann, 2011). Solche Risikofaktoren führen nicht unausweichlich zur Entstehung psychischer Störungen (Fingerle, 2008). Spezifische genetische, neurobiologische, psychologische und soziale Risikofaktoren können jedoch die Auftretenswahrscheinlichkeit von emotionalen-sozialen Entwicklungsstörungen erhöhen (Fingerle, 2008; Wiedebusch & Petermann, 2011). Stehen den Risikofaktoren ausreichend personale und soziale Schutzfaktoren gegenüber, werden psychische Belastungen gemindert und der Entwicklungsverlauf der Kinder und Jugendlichen wird positiv beeinflusst (Fingerle, 2008). Als Schutzfaktoren sind beispielsweise sozial-emotionale Kompetenzen, wie das prosoziale Verhalten zu nennen (Beelmann & Raabe, 2007; Lohbeck, Petermann & Petermann, 2014b). Diese sollten möglichst frühzeitig gefördert werden (Voß & Gebhardt, 2017a; Wiedebusch & Petermann, 2011). Präventive Maßnahmen zielen darauf ab Lern- und Entwicklungsbarrieren zu reduzieren, die schulischen Leistungen der SchülerInnen positiv zu beeinflussen, die Ausbildung von erwünschten Handlungsmustern zu unterstützten und den SchülerInnen die bestmögliche schulische, berufliche und soziale Teilhabe zu ermöglichen (Hennemann et al., 2018; KMK, 2000). Die SchülerInnen sollen u.a. erlernen, ihr Verhalten angemessen zu steuern und diese (Steuerungs-)Fähigkeit langfristig zu stabilisieren (KMK, 2000). Verhaltensstörungen sollten nicht ausschließlich als Schwierigkeiten aufgefasst werden (Stein & Stein, 2014). Es sollte ebenfalls der zugrundeliegende Sinngehalt berücksichtigt werden, um sie in angemessener Weise in Fördermaßnahmen und Unterrichtsprozessen berück-

sichtigen zu können (ebd.). Der Schule kommt als Präventionssetting eine besondere Bedeutung zu, da sie für alle Kinder zugänglich ist und somit alle Kinder potentiell erreicht werden können (Brezinka, 2003). Beelmann (2008) bezeichnet die Schule sogar als das wichtigste Präventionssetting. Auch im Förderschwerpunkt emotionale und soziale Entwicklung soll eine „Förderung im Rahmen der Vorbeugung“ in den allgemeinbildenden Schulen in kooperativer Zusammenarbeit mit Förderschulen, Förderzentren und Beratungs- und Unterstützungsdiensten stattfinden (KMK, 2000, S. 3). Erfolgreich sind präventive Maßnahmen dann, wenn über verschiedene Kontexte eine regelmäßige Kommunikation erfolgt (Hartke, 2017; Petermann & Lehmkuhl, 2010). Als Grundlage dafür sollten Schülerdaten mithilfe verlaufsdiagnostischer Verfahren erhoben werden (Hartke, 2017; Voß & Gebhardt, 2017a). Diese ermöglichen die Abbildung von Entwicklungsverläufen und unter Berücksichtigung der individuellen Bezugsnorm Aussagen über den Erfolg eingesetzter präventiver Maßnahmen (Voß & Gebhardt, 2017a). Es zeigt sich insgesamt ein großer Bedarf an statusdiagnostischen und auch verlaufsdiagnostischen Instrumenten (ebd.). In den nachfolgenden Kapiteln werden verlaufsdiagnostische Ansätze in Abgrenzung zur Statusdiagnostik vorgestellt.

3. Verlaufsdiagnostik

Mit der Weiterentwicklung des deutschen Schulsystems unter Berücksichtigung der Inklusion liegt der Schwerpunkt sonderpädagogischer Diagnostik nicht mehr wie zuvor auf der statusdiagnostischen Feststellung von Förderbedarfen oder der Zuordnung von Ressourcen, sondern vielmehr auf einer verlaufsdiagnostischen, lernbegleitenden Diagnostik, die Ableitungen und Evaluationen gezielter Fördermaßnahmen erlaubt (Voß & Gebhardt, 2017b). In der sonderpädagogischen Forschung und Praxis gewinnt damit die Thematik der Verlaufsdiagnostik zunehmend an Bedeutung (ebd.). Dies gilt insbesondere für den Bereich akademischer Leistungen, aber auch für den Bereich sprachlicher und emotional-sozialer Kompetenzen (ebd.). Dementsprechend sollte der Begriff der Verlaufsdiagnostik von dem Begriff der Lernverlaufsdiagnostik, wie er von Klauer (2011) geprägt wurde, unterschieden werden (ebd.). Die Verlaufsdiagnostik bzw. Prozessdiagnostik unterscheidet sich von der Statusdiagnostik hinsichtlich zentraler Aspekte (Casale et al., 2015a). Erstere zielt darauf ab basierend auf häufigen, kurzen Messungen enge Konstrukte bzw. Merkmale änderungssensitiv zu erfassen (Casale et al., 2015a; Grosche 2014). Sie liefert damit eine Grundlage für Analysen individueller Merkmalsänderungen und ermöglicht Aussagen über die Entwicklungsverläufe der SchülerInnen (Casale et al., 2015a). Die Statusdiagnostik hingegen wird einmalig bzw. selten im Sinne einer intensiven umfassenden Diagnostik zur Messung eines breit gefassten Konstrukts eingesetzt und dient der Beurteilung des „Ist-Zustandes“ (ebd., S. 38). Vergleiche erfolgen bei der Verlaufsdiagnostik unter Berücksichtigung der individuellen Bezugsnorm, bei der Statusdiagnostik unter Berücksichtigung der sozialen Bezugsnorm (Grosche, 2014). Im Zuge der Verlaufsdiagnostik erfolgt ein Monitoring betroffener Merkmale bzw. Problemfelder über längere Zeiträume (Voß &

Gebhardt, 2017a). Monitoring beschreibt dabei eine kontinuierliche Entwicklungsbeobachtung und -dokumentation, die Entscheidungen über die Angemessenheit von Fördermaßnahmen ermöglicht (ebd.). Auf Basis der Schülerdaten und Entwicklungsverläufe können dann Rückmeldungen für Lehrkräfte und SchülerInnen abgeleitet werden (ebd.). Methoden des Monitorings ermöglichen durch wiederholte Erhebungen von Schülerdaten die Abbildung von Entwicklungsverläufen (ebd.). Verlaufsdiagnostische Instrumente sollten somit änderungssensitiv sein, statusdiagnostische Instrumente hingegen änderungsresistent (Casale et al., 2015a; Grosche 2014). Instrumente der Verlaufsdiagnostik erscheinen geeignet, basierend auf der Abbildung der Entwicklungsverläufe einzelner SchülerInnen Fördermaßnahmen im Einzelfall zu überprüfen (Casale, 2017). Casale, Hennemann & Grosche (2015b) sehen hier die Möglichkeit zur Erweiterung der evidenzbasierten Praxis um das Konstrukt der „Evidenzbasierung im Einzelfall“ (S. 327). Im Sinne der Evidenzbasierung im Einzelfall sollen Schülerdaten engmaschig erhoben werden, um pädagogisch relevante Entscheidungen treffen zu können (Casale, 2017). Die Verlaufsdiagnostik ist im Allgemeinen für das Konzept der evidenzbasierten pädagogischen Praxis von zentraler Bedeutung (Casale et al., 2015b). Eine evidenzbasierte Pädagogik überprüft wissenschaftlich auf Basis empirischer Forschungsmethoden den Erfolg eingesetzter Maßnahmen (Hennemann et al., 2017). Verlaufsdiagnostische Verfahren ermöglichen dabei zeitnahe Aussagen hinsichtlich der Wirksamkeit dieser Fördermaßnahmen (Casale et al., 2015b). Folglich sind verlaufsdiagnostische Instrumente für gestufte Fördersysteme wie dem RTI-Ansatz von großer Bedeutung (ebd.). Der RTI-Ansatz zielt darauf ab eine Passung zwischen den individuellen Bedürfnissen und den individuellen Fördermaßnahmen der SchülerInnen herzustellen (Casale et al., 2015a). Mittels verlaufsdiagnostischer Methoden kann dann überprüft werden, ob die Fördermaßnahmen für die jeweiligen SchülerInnen geeignet sind und eine „Passung im Einzelfall“ vorliegt (Casale et al., 2015a, S. 37). Der Begriff „intervention“ entspricht hierbei den Fördermaßnahmen, der Begriff „response“ der Effektivität der Fördermaßnahmen, bzw. der individuellen Entwicklung der SchülerInnen (Casale et al., 2015a).

Volpe und Fabiano (2013) konnten am Beispiel der Daily Behavior Report Cards (DBRC) die Effektivität formativer Evaluation nachweisen. Bei den DBRC handelt es sich um eine systematischen Methode, die dazu dient den SchülerInnen ein häufiges Feedback bezüglich ihres Verhaltens zu geben (Volpe & Fabiano, 2013). Ein Vorteil von Methoden, die in regelmäßigen Zeitabständen (z.B. täglich) eingesetzt werden, ist, dass die Aufmerksamkeit der Lehrkräfte für angestrebte Verhaltensweisen aufrechterhalten wird (ebd.). Dies führt dazu, dass entsprechende Interventionsmaßnahmen routiniert und durchgängig eingesetzt werden (ebd.). Des Weiteren werden die Entwicklungen der SchülerInnen in zeitlicher Nähe zum Auftreten des Verhaltens dokumentiert, sodass zeitnahe Interventionen bei negativen Entwicklungstendenzen möglich sind (ebd.). Darüber hinaus können Methoden wie die DBRC die Kommunikation

zwischen am Förderprozess beteiligten Personen (z.B. Lehrer und Eltern) als nützliches und hilfreiches Werkzeug unterstützen (Volpe & Fabiano, 2013).

Fuchs (2004) empfiehlt für Forschungstätigkeiten zur Erhebung der Testgüte von verlaufsdiagnostischen Instrumenten Untersuchungen auf drei Forschungsstufen. Als erstes sollte die statusdiagnostische Testgüte der Verfahren überprüft werden (Stufe 1) (Fuchs, 2004). Es sollten die gängigen statusdiagnostischen Gütekriterien der Objektivität, Reliabilität und Validität nachgewiesen werden (Voß & Gebhardt, 2017b; Voß, Sikora & Hartke, 2017). Diese Hauptgütekriterien gewährleisten, dass Wahrnehmungs- und Beurteilungsfehler verringert werden (Voß et al., 2017). Des Weiteren ist die verlaufsdiagnostische Testgüte des Instrumentes zu überprüfen (Stufe 2) (Fuchs, 2004). Denn nur wenn Kriterien wie die Messinvarianz und Änderungssensibilität gegeben sind, können Entwicklungsverläufe abgebildet werden (Voß & Gebhardt, 2017b). Zuletzt sollte untersucht werden, ob das verlaufsdiagnostische Verfahren im entsprechenden pädagogischen Setting von Nutzen ist (Stufe 3) (Fuchs, 2004). Die Nebengütekriterien Nützlichkeit und Ökonomie sollten gewährleistet sein, da das verlaufsdiagnostische Verfahren nur dann regelmäßig im schulischen Alltag eingesetzt werden kann (Voß & Gebhardt, 2017a).

Wie die Verlaufsdiagnostik umgesetzt wird, ist abhängig vom entsprechenden Förderschwerpunkt (Casale et al., 2015b). Für den Bereich der Lernverlaufsdiagnostik liegen bereits einige evaluierte Verfahren vor. Im Bereich der Verhaltensverlaufsdiagnostik mangelt es hingegen noch an Verfahren, die Verhalten verlaufsdiagnostisch erfassen können (Casale, 2017; Voß & Gebhardt, 2017a). In den nachfolgenden Kapiteln werden beide Ansätze dargelegt.

3.1. Lernverlaufsdiagnostik

Der Begriff der Lernverlaufsdiagnostik beschreibt Testverfahren, welche entsprechend des formativen Assessments bzw. der formativen Diagnostik Lernverläufe erheben (Förster, Kuhn & Souvignier, 2017). Die Lernverlaufsdiagnostik stellt dabei eine spezifische Form der formativen Diagnostik dar (Gebhardt, Diehl & Mühling, 2016; Mühling, Gebhardt & Diehl, 2017). Die formative Diagnostik zielt darauf ab, die Leistungen der SchülerInnen zu erfassen, um unter Berücksichtigung dieser Informationen eine Modifikation der eingesetzten Fördermaßnahmen vorzunehmen (Voß et al., 2017; Jungjohann, Gebhardt, Diehl & Mühling, 2017). Bedeutsam ist dabei nicht das alleinige Erheben der Schülerdaten, sondern bedeutsam sind die Konsequenzen bzw. Schlüsse, die aus diesen Schülerdaten gezogen werden und zur Adaption von Unterricht und Förderung herangezogen werden (Jungjohann & Gebhardt, 2018). Im Sinne der formativen Diagnostik werden somit in regelmäßigen Abständen wiederholend Lernprozesse im Verlauf evaluiert (Klauer, 2014). Der formativen Diagnostik steht der Ansatz der summativen Diagnostik gegenüber (ebd.). Gemäß dieses Ansatzes wird am Ende eines Lernprozesses das erreichte Ergebnis bzw. der Leistungsstand gemessen (ebd.). Im englischsprachigen Raum wurde von Deno der Begriff „curriculum-based measurement“ (CBM) geprägt

(Deno, 2003). Ziel dieses Ansatzes war es festzustellen, inwieweit die SchülerInnen Inhalte, die aktuell im Unterricht behandelt wurden, beherrschten (Klauer, 2011). Klauer (2011) führt an, dass sich eine langfristig durchgeführte Lernverlaufsdiagnostik (z.B. im Zeitraum von einem Schuljahr) von diesem Ansatz lösen muss. Im Gegensatz zu dem, was aktuell im Unterricht behandelt wird, sollte vielmehr das erhoben werden, was die SchülerInnen am Ende des Schuljahres erreichen sollen. Im deutschen Sprachraum prägte Klauer zunächst den Begriff der Lernfortschrittsmessung und unter Berücksichtigung möglicher Lernstillstände und Lernverluste schließlich den Begriff der Lernverlaufsdiagnostik (Klauer, 2011). Weitere synonym verwendete Begriffe sind u.a. Progress Monitoring oder formative Unterrichtsevaluation (Voß et al., 2017). Lange Zeit fand diese Methode im deutschen Sprachraum keine Beachtung (Klauer, 2011). Angesichts der Weiterentwicklung des deutschen Schulsystems hin zu einem inklusiven Schulsystem ist sie zunehmend von Bedeutung, da sie eine besonders enge Verzahnung der Bereiche Diagnostik und Förderung ermöglicht und einen Beitrag zur Individualisierung des Unterrichts leisten kann (Mühling et al, 2017; Wilbert, 2014). Dabei ist es von zentraler Bedeutung, dass Schülerleistungen und -entwicklungen valide eingeschätzt, Leistungsrück- und -stillstände zeitnah erkannt und angemessene Fördermaßnahmen zur Unterstützung der Lernentwicklungen ergriffen werden (Gebhardt et al., 2015a). Dies kann durch die Lernverlaufsdiagnostik gewährleistet werden (ebd.). Im Fokus der Lernverlaufsdiagnostik steht die Veränderung (Wilbert & Linnemann, 2011). Ähnlich zur CBM können mithilfe der Lernverlaufsdiagnostik Lernverläufe abgebildet werden (Klauer, 2011). Dazu werden über einen längeren Zeitraum in regelmäßigen kurzen Abständen ökonomische Testinstrumente eingesetzt, die wiederholend dieselbe Kompetenz erheben (Gebhardt, Heine, Zeuch & Förster, 2015b; Klauer 2011). Diese wiederholenden Messungen und die anschließenden Darstellungen der Lernverläufe ermöglichen im Ganzen eine Abbildung des Lernverlaufs (Gebhardt, et al., 2015b; Klauer, 2011). Auf Basis der Darstellungen der Lernverläufe können Lehrkräfte die Wirksamkeit des Unterrichts evaluieren und Instruktionen für einzelne oder alle SchülerInnen einer Klasse modifizieren (Gebhardt, et al., 2015b). Dazu müssen die erhobenen Leistungen der SchülerInnen in Bezug zu einem anzustrebenden Verlauf, einem „Soll-Verlauf“ bzw. einer „goal-line“ gesetzt werden (Voß & Hartke, 2014, S. 96). Lernverlaufsdiagnostik ermöglicht eine systematische Evaluation eingesetzter Fördermaßnahmen und somit eine bessere Passung zwischen Lernbedürfnissen und Förderung (Grosche, 2014). Insbesondere in Klassen, in denen die SchülerInnen eine große Heterogenität aufweisen kann die Lernverlaufsdiagnostik eine bedeutende Informationsbasis liefern, auf der relevante pädagogische Entscheidungen getroffen werden können (Gebhardt, et al., 2015b). Auf Basis entsprechender Erhebungen kann für alle SchülerInnen unabhängig des Leistungsniveaus erhoben werden, ob eingesetzte Unterrichtsmaßnahmen zu einem Lernfortschritt führen (ebd.). Lernschwierigkeiten, die poten-

tiell den weiteren Lernerfolg gefährden, können mittels Lernverlaufsdiagnostik frühzeitig erfasst werden (Mühling et al., 2017). Zielsetzungen der Lernverlaufsmessungen orientieren sich am schulischen Curriculum, welches Lernziele für alle SchülerInnen formuliert (Voß & Gebhardt, 2017a). Auf Basis der wiederholten Messungen sind zudem Interpretationen unter Berücksichtigung der individuellen Bezugsnorm möglich (Förster et al., 2017). Generell sind im Zuge der Lernverlaufsdiagnostik klare Definitionen der Kompetenzen notwendig, die wiederholt erfasst werden sollen (Klauer, 2014). Eine Herausforderung in der Testkonstruktion stellt der Aspekt dar, dass für jede Messung neue Tests vorliegen müssen, da bei wiederholtem Einsatz des gleichen Instruments ein Lerneffekt die Ergebnisse verzerren kann (Klauer, 2011; Voß 2014). Die Instrumente müssen dabei jeweils dieselbe Kompetenz erfassen, also eine hohe Validität aufweisen und durchweg homogen schwierig sein (Klauer, 2011; Strathmann & Klauer, 2010). Entsprechende Testinstrumente können als Paralleltests bezeichnet werden (Voß, 2014). Sie erleichtern die Kommunikation über eingesetzte Fördermaßnahmen (Grosche, 2014). Des Weiteren müssen lernverlaufsdiagnostische Testverfahren Lernzuwächse, also Veränderungen der betreffenden Kompetenzen, sensibel erfassen können und das Gütekriterium der Ökonomie erfüllen, also schnell durchführbar, auswertbar und im schulischen Kontext praktikabel sein (Gebhardt, et al., 2015b; Klauer, 2011; Strathmann & Klauer, 2010). Eine ökonomische Auswertung und Ergebnisdokumentation können beispielsweise computerbasierte Testverfahren ermöglichen (Gebhardt, et al., 2015b). Computerbasierte Testverfahren werden beispielsweise kostenfrei von der Onlineplattform „LEVUMI“ bereitgestellt. Diese Plattform bietet umfangreich evaluierte Testverfahren für die Lernbereiche Deutsch und Mathematik (vgl. www.levumi.de; Mühling et al., 2017). Im Allgemeinen sind standardisierte, empirisch evaluierte Testinstrumente notwendig (ebd.). Diesbezüglich kommen klassisch konstruierte Testinstrumente nur bedingt infrage, da sie i.d.R. nur einmalig eingesetzt werden und aufgrund einer klassenstufenspezifischen Normierung Informationen zum Lernstand der SchülerInnen unter Berücksichtigung der sozialen Bezugsnorm liefern (Gebhardt, et al., 2015b; Klauer, 2011). Darüber hinaus sind sie zumeist sehr umfangreich und es liegen oftmals zu wenige Paralleltest vor (Gebhardt, et al., 2015b). Beispiele statusdiagnostischer Instrumente sind u.a. Klassenarbeiten und Schulleistungstests (Grosche, 2014). Sollen im Speziellen individuelle Lernentwicklungen von SchülerInnen einer leistungsheterogenen Klasse erfasst werden, sind aussagekräftige Testinstrumente notwendig, die entsprechend der Lernausgangslagen der SchülerInnen adaptiert werden können (Balt et al., 2017). Während es im englischen Sprachraum für zahlreiche schulische Förderbereiche eine große Anzahl evaluierter Testinstrumente gibt, liegen für den deutschen Sprachraum deutlich weniger Testinstrumente vor (Huber & Rietz, 2015). Testinstrumente zum Bereich Lesen lieferten beispielsweise Diehl und Hartke (2011) sowie Walter (2008). Diehl und Hartke (2011) konnten für das Inventar zur Erfassung der Lesekompetenz von Erstklässlern (IEL-1) eine gute Objektivität

und eine ausreichende Reliabilität und Validität nachweisen. Walter (2008) konnte für ein Set von Leseabschnitten auf der Basis der Technik des 1-Minute-Lauten-Lesens (CBM-LL) eine hohe Änderungssensibilität nachweisen. Ein Testinstrument zum Lernbereich Mathematik wurde beispielsweise von Gebhardt et al. (2015b) evaluiert. Gebhardt et al. (2015b) konnten für die Testreihe zur Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse sowohl auf der Dimension der Vorläuferkompetenzen als auch auf der Dimension der curricularen Kompetenz zufriedenstellende Reliabilitätswerte feststellen. In einer Überblicksarbeit konnten Stecker, Fuchs und Fuchs (2005) aufzeigen, dass der Einsatz von CBM zu einem signifikanten Anstieg der Schülerleistungen führte. Die Autoren konnten außerdem positive Effekte der Lernverlaufsdiagnostik für SchülerInnen mit sonderpädagogischen Förderbedarf feststellen. In einer deutschen Studie konnten Förster und Souvignier (2015) aufzeigen, dass SchülerInnen höhere Leseleistungen zeigten, wenn Lehrkräfte Rückmeldungen zum Verlauf der Leseleistungen erhielten und darin geschult wurden diese Informationen für Entscheidungen über weitere Instruktionen heranzuziehen. Aufgrund positiver Forschungsbefunde ist die formative Evaluation Gegenstand des pädagogischen Diskurses, insbesondere dann, wenn es um die Prävention schulischer Schwierigkeiten geht (Voß et al., 2017). Es zeigt sich, dass eine regelmäßig durchgeführte Verlaufsdiagnostik einen positiven Einfluss auf die Lernentwicklung von SchülerInnen hat (Balt et al., 2017). Dies konnte in einer Vielzahl von Studien nachgewiesen werden (Black & Wiliam, 1998; Hattie et al., 2013; Kingston & Nash, 2011; Stecker et al., 2005). Voß, Sikora und Hartke (2017) weisen in ihrem Beitrag zur Lernverlaufsdiagnostik auf die Hattie-Studie hin. Diese Studie sollte nicht unreflektiert herangezogen werden, jedoch kann sie Hinweise auf die Effektivität bestimmter Unterrichtsmerkmale und –konzepte liefern (Voß et al., 2017). Hattie führt in dieser Studie die „Formative Evaluation des Unterrichts“ als einen Einflussfaktor auf das Lernen der SchülerInnen auf (Hattie et al., 2013, S. 215f.). Die Effektivität des Einflussfaktors

formative Evaluation liegt unabhängig von dem Alter der Lernenden und der Art des sonderpädagogischen Förderbedarfs der SchülerInnen vor (ebd.). In einer qualitativen Überblicksarbeit konnten Black und William (1998) ebenfalls aufzeigen, dass SchülerInnen mit Lernschwierigkeiten vom Einsatz verlaufsdiagnostischer Verfahren, in diesem Fall im Sinne eines formativen Feedbacks, profitieren konnten. Der alleinige Einsatz verlaufsdiagnostischer Verfahren stellt dabei jedoch noch keine hinreichende Bedingung für Lernerfolge dar (Balt et al., 2017). Lernerfolge können dann auf den Einsatz verlaufsdiagnostischer Verfahren zurückgeführt werden, wenn unterrichtliches Handeln mit Bezug zu erfassten Lernentwicklungen angepasst wird (Stecker et al., 2005) oder die SchülerInnen ein Feedback erhalten (Hattie & Timperley, 2007; Shute, 2008). Bei letzterem Aspekt ist jedoch zu beachten, dass nicht jedes Feedback in gleicher Weise effektiv ist (Hattie & Timperley, 2007; Shute, 2008). Welche Art von Feedback eingesetzt und wie dieses kommuniziert wird, ist genauso von Relevanz, (Hattie & Timperley,

2007) wie das Motiv des Feedbacks, ein Einsatz zum richtigen Zeitpunkt und der Aspekt, dass das Feedback von SchülerInnen als Hilfsmittel aufgefasst wird (Shute, 2008). Voß et al. (2017) beschreiben, dass im Überblick ein positiver Trend der formativen Evaluation von Leistungen festzustellen ist. Dieser positive Trend bietet großes Potenzial für die schulische Praxis. Verlaufsdiagnostische Ansätze finden auch in der Verhaltensdiagnostik zunehmend Anklang.

3.2. Verhaltensverlaufsdiagnostik

Sozial-emotionale Entwicklungsrückstände und negative Verhaltenstendenzen sollten frühzeitig erfasst werden, um individuelle präventive Fördermaßnahmen ableiten und negativen Entwicklungsverläufen entgegenwirken zu können (Casale et al., 2015b; Voß & Gebhardt, 2017a; Wiedebusch & Petermann, 2011). Dazu ist es notwendig, status- und verlaufsdiagnostische Instrumente regelmäßig einzusetzen (Wiedebusch & Petermann, 2011). Grundsätzlich ist jedoch zu beachten, dass breit gefasste statusdiagnostische Instrumente keine präzisen Messungen der Verhaltensveränderungen von SchülerInnen ermöglichen (Conroy et al., 2008). Entscheidungen ausschließlich auf statusdiagnostisch erhobene Ergebnisse zu stützten, würde demnach außer Acht lassen, dass spezifische Ereignisse die emotionale Konstitution und das Verhalten von SchülerInnen rapide verändern können (Dever et al., 2015). Dementsprechend sind verlaufsdiagnostische Instrumente für den Förderschwerpunkt emotionale und soziale Entwicklung von besonderer Bedeutung (Casale et al., 2015a). Casale et al. (2015a) bezeichnen die Diagnostik von Verläufen im Förderschwerpunkt emotionale und soziale Entwicklung entsprechend als Verhaltensverlaufsdiagnostik.

Die Verlaufsdiagnostik ermöglicht die Identifikation problematischen Verhaltens und eine Evaluation eingesetzter pädagogischer Maßnahmen auf deren Grundlage Entscheidungen im Hinblick auf notwendige Adaptionen dieser Maßnahmen getroffen werden können (Voß & Gebhardt, 2017a). Die Verhaltensverlaufsdiagnostik ist somit von zentraler Bedeutung für mehrstufige Fördersysteme (wie den RTI-Ansatz), die auf den Bereich des Verhaltens abzielen und trägt zu einer evidenzbasierten und präventiven sonderpädagogischen Praxis bei (Casale, 2017). Instrumente der Verhaltensverlaufsdiagnostik können bei Bedarf bis zu mehrmals täglich eingesetzt werden und ermöglichen so zahlreiche Messungen (Casale et al., 2017). Verlaufsdiagnostische Instrumente werden in der Regel nicht für alle SchülerInnen einer Klasse durchgeführt, sondern lediglich für eine bestimmte Anzahl von SchülerInnen (ebd.). Entsprechende SchülerInnen können beispielsweise mittels universeller Screeningverfahren identifiziert werden (ebd.). Ihr Verhalten wird im Vergleich zum Verhalten der MitschülerInnen und somit unter Berücksichtigung der sozialen Bezugsnorm als problematisch eingeschätzt (ebd.). Zusätzliche Fördermaßnahmen erscheinen notwendig (ebd.). Die Inhalte der Verhaltensverlaufsdiagnostik werden mit Bezug zum situativen Kontexten ausgewählt und sind im Gegensatz zur Lernverlaufsdiagnostik nicht schulorganisatorisch festgelegt (Voß & Gebhardt, 2017a). Dabei ist zu beachten, das Schülerverhalten im Gegensatz zu objektiven Testwerten

der Lernverlaufsdiagnostik nicht im Hinblick auf Richtigkeit, sondern nur im Hinblick auf Angemessenheit beurteilt werden kann (Casale et al., 2015a). Die Entscheidung darüber, ob Verhaltensweisen angemessen sind, ist subjektiv geprägt und sollte immer vor dem Hintergrund spezifischer schulischer Settings sowie unter Berücksichtigung von Variablen der Klassensituation und der Lehrperson zu betrachten (Voß & Gebhardt, 2017a). Die Auswahl spezifischer Verhaltensweisen ist somit von großer Bedeutung für die Verhaltensverlaufsdiagnostik im schulischen Setting (Casale et al., 2015a). Die Forschung zur Verhaltensverlaufsdiagnostik und die Entwicklung und Evaluation von verlaufsdiagnostischen Instrumenten steht noch am Anfang (Huber & Rietz, 2015). Es mangelt an deutschsprachigen verlaufsdiagnostischen Instrumenten für die große Gruppe inklusiv zu beschulender SchülerInnen mit dem Förderschwerpunkt emotionale und soziale Entwicklung (Casale et al., 2015a). Eine Entwicklung valider und reliabler Instrumente zur Verhaltensverlaufsdiagnostik ist von Bedeutung, da sie eine angemessene Förderung von SchülerInnen mit Verhaltensproblemen und einen effizienten Einsatz von Ressourcen im schulischen Setting ermöglicht (Huber & Rietz, 2015). Zudem konnten Forschungsarbeiten aufzeigen, dass ein Feedback zu Lern- und Entwicklungsständen einen positiven Einfluss auf die Effizienz pädagogischer Interventionen hat (vgl. z.B. (Black & Wiliam, 1998; Hattie et al., 2013; Kingston & Nash, 2011; Stecker et al., 2005). Interventionen für diese Schülergruppen sind aufwändig und sollten daher hinsichtlich gewünschter Erfolge überprüft werden (Huber & Rietz, 2015). Es sind weitere Forschungstätigkeiten zur engmaschigen und ökonomischen Evaluationen von Interventionen bei Verhaltensproblemen notwendig. Dabei gilt es zu berücksichtigen, dass Ergebnisse und Testgüte der Verhaltensverlaufsdiagnostik von den Variablen Rater und Zeitpunkt beeinflusst werden (Casale, 2017).

Die geringe Anzahl verlaufsdiagnostischer Instrumente kann auf zwei Problematiken zurückgeführt werden. Erstens müssen verlaufsdiagnostische Instrumente zahlreiche Gütekriterien erfüllen (Casale et al., 2015a). Zweitens stellt sich im Zuge der Verlaufsdiagnostik die Frage, wie Verhalten als Indikator des emotionalen und sozialen Entwicklungsstandes der SchülerInnen unter Berücksichtigung der relevanten Gütekriterien erfasst werden kann (ebd.). Casale et al. (2015a) führen elf relevante Gütekriterien der Verhaltensverlaufsdiagnostik auf: Objektivität, Reliabilität, Validität, Skalierung, Ökonomie, Anwendung eines gültigen Messmodells, Eindimensionalität, Veränderungssensitivität, Inferenz, Direktheit und Orientierung an einer individuellen Bezugsnorm. Aufgrund des Umfangs dieser Arbeit werden die Gütekriterien nachfolgend nur kurz erläutert. Die im Rahmen der vorliegenden Pilotierungsstudie untersuchten und fokussierten Gütekriterien werden im Kapitel 6.3.1 (Studiendesign) tiefergehend erläutert. Das Gütekriterium der Objektivität erfasst, ob Testergebnisse unabhängig vom Testleiter und von den Zeitpunkten der Messungen vorliegen (Casale et al., 2015a). Reliabilität beschreibt, mit welcher Messgenauigkeit ein spezifisches Merkmal erfasst wird. Dazu kann beispielsweise

die interne Konsistenz der Items ermittelt werden (ebd.). Validität erhebt, „ob ein Test tatsächlich das misst, was er zu messen beansprucht“ (ebd., S. 39). Im schulischen Kontext ist vor allem die soziale Validität von Bedeutung, mit der untersucht wird, ob ausgewählte Verhaltensweisen unmittelbar von Relevanz sind für das schulische Setting (Casale et al., 2015a). Das Kriterium der Skalierung untersucht die zur Bildung eines Testwertes verwendete Verrechnungsvorschrift (ebd.). Ein Testverfahren kann dann als ökonomisch bezeichnet werden, wenn es schnell und leicht durchgeführt und ausgewertet werden kann (ebd.). Im Zuge der Untersuchung des Messmodells wird überprüft, ob ein gültiges Messmodell angewendet wird (ebd.). Es wird untersucht welche Wirkrichtung zwischen latenten, also nicht-beobachtbaren Variablen und manifesten bzw. beobachtbaren Variablen vorliegt (ebd.). Das Kriterium der Eindimensionalität liegt dann vor, wenn Korrelationen zwischen beantworteten Items einzig auf eine bestimmte latente Variable zurückgeführt werden und andere Einflussgrößen nicht von Relevanz sind (Bühner, 2011). Das Gütekriterium der Änderungssensitivität untersucht, ob über kurze Zeitspannen kleine Veränderungen latenter Merkmale mittels der konstruierten Items erfasst werden können (Casale et al., 2015a). Mit dem Kriterium der Inferenz wird „der Aufwand bzw. die Komplexität der schlussfolgernden Kognitionen“ erhoben, die zum Ausfüllen oder zur Beantwortung eines Tests bzw. Items notwendig sind (ebd., S. 41). Das Kriterium der Direktheit beschreibt den zeitlichen Abstand des Einsatzes eines Verfahren zur Situation, in dem das Verhalten aufgetreten ist (ebd.). Verlaufsdiagnostische Verfahren sollten zudem Vergleiche mit der individuellen Bezugsnorm ermöglichen. Dabei sind individuelle Verhaltensveränderungen über die Zeit von Relevanz (ebd.). Casale et al. (2015a) konnten in ihrer theoretischen Überblicksarbeit herausstellen, dass keine der gängigen verhaltensdiagnostischen Methoden alle verlaufsdiagnostisch relevanten Gütekriterien erfüllt. Zwei dieser Methoden, die Verhaltensbeobachtung und die Verhaltensbeurteilung mit Ratingskalen, entsprechen jedoch zahlreichen Gütekriterien (ebd.). Diese beiden Methoden werden nachfolgend erläutert und ihr Potential für die Verlaufsdiagnostik wird dargestellt.

3.2.1. direkte systematische Verhaltensbeobachtung

Verhaltensbeobachtungen ermöglichen die Erfassung von diagnostisch relevanten Informationen zur Häufigkeit, Dauer oder Intensität konkreter Verhaltensweisen (Casale et al., 2015a; Schmidt-Atzert, Amelang, Fydrich, Moosbrugger & Zielinski, 2012). Fremdbeobachtungen stellen dabei den Regelfall dar, aber auch Selbstbeobachtungen sind möglich (Casale et al., 2015a). Sie können unterschiedlichster Form sein: frei oder systematisch, direkt oder indirekt, verdeckt oder offen, teilnehmend oder nichtteilnehmend (ebd.).

Direkte Verhaltensbeobachtungen sind eine der am häufigsten verwendeten Methode zur Verhaltensdiagnostik (Steege, Davin & Hathaway, 2001). Steege, Davin und Hathaway (2001) bezeichnen direkte Verhaltensbeobachtungen als das Herzstück der Verhaltensdiagnostik. Bei direkten Verhaltensbeobachtungen ist es den Beobachtern möglich die Aufmerksamkeit

auf interessante Geschehnisse zu lenken (Schmidt-Atzert et al., 2012). Ein Nachteil von direkten Verhaltensbeobachtungen ist, dass Beobachter nicht gleichzeitig beobachten und registrieren können (ebd.).

Bei systematischen Verhaltensbeobachtungen wird i.d.R. ein spezifischer Verhaltensbereich fokussiert, wie z.B. Aggression (ebd.). Dieser wird über einen bestimmten Zeitraum (z.B. zwei Wochen) in kurzen zeitlichen Intervallen (z.B. minutiös) in relevanten Unterrichtssettings erfasst und dokumentiert (Casale et al., 2015b). Eine exakte Operationalisierung der zu beobachtenden Verhaltensweisen und eine Schulung der beobachtenden Personen ist notwendig (Huber & Rietz, 2015). Der Beobachter erhält Vorgaben dazu, welche Verhaltensweisen beobachtet und wie diese protokolliert werden sollen, z.B. mittels Strichlisten in Protokollbögen oder Time-Sampling Methoden (Schmidt-Atzert et al., 2012). Den Verfahren liegen entweder Zeichen- oder Kategoriensysteme zugrunde. Zeichensystemen gehen mit der Erfassung von einzelnen Verhaltensbereichen bzw. bestimmten Teilen eines Verhaltens einher. Kategoriensysteme hingegen zielen auf die vollständige Erfassung eines Verhaltensaspektes ab (Schmidt-Atzert et al., 2012). Dabei werden Segmentierungsprozesse durchgeführt: Relevante Verhaltensweisen werden von weniger relevanten Verhaltensweisen unterschieden und ihnen wird eine vermutete Bedeutung zugeschrieben (ebd.). Aussagen über Häufigkeit, Dauer und Intensität machen die beobachteten Verhaltensweisen quantifizierbar (ebd.).

Die methodische Kombination dieser beiden Formen, die direkte systematische Verhaltensbeobachtung (engl. Systematic Direct Observation (SDO)) ermöglicht exakteste und differenzierteste Verhaltensbeobachtungen (Huber & Rietz, 2015). Die direkte systematische Verhaltensbeobachtung erfüllt wichtige psychometrische Gütekriterien wie Objektivität, Reliabilität, Validität, Skalierung und Nebengütekriterien wie Eindimensionalität, Änderungssensitivität und Direktheit (Casale et al., 2015a). Die Inferenz der Verfahren kann niedrig gehalten werden, wenn spezifische Verhaltensweisen beobachtet werden und auch Vergleiche mit Bezug zur individuellen Bezugsnorm sind möglich (ebd.). Sie basieren zudem auf spezifizierten Messmodellen (ebd.). Folglich können systematische Verhaltensbeobachtungen Verhaltensveränderungen objektiv, reliabel und sensitiv abbilden (Casale et al., 2015b). Chafouleas et al. (2009a) führen die Flexibilität als einen Vorteil der Methode an. Das Instrument und die Vorgehensweise können kontextabhängig ausgewählt werden. Die Testgüte von Verfahren der systematischen Verhaltensbeobachtung steht in Abhängigkeit zu diversen Einflussfaktoren, wie der Beobachtungsdauer, dem Beobachter, dem Setting und den Zielverhaltensweisen (Casale et al., 2015a). Forschungsergebnisse zeigen, dass für ausreichende Beurteilerübereinstimmungen zwischen 7 und 20 Messzeitpunkte notwendig sind (Huber & Rietz, 2015). Da die Methode jedoch sehr aufwändig ist, ist ein engmaschiger effizienter Einsatz im schulischen Kontext nur unter Einsatz zusätzlicher personeller Ressourcen realisierbar (Casale et al., 2015b; Chafouleas et al., 2009a). Die engmaschige Anwendung im schulischen Kontext ist jedoch ein

zentrales Merkmal der Verlaufsdiagnostik (Casale et al., 2015b). Folglich erfüllt die Methode nicht das Kriterium der Ökonomie (Casale et al., 2015a).

3.2.2. Verhaltensbeurteilung mit Ratingskalen

Ein weiteres Verfahren, welches sich in Ansätzen für verlaufsdiagnostische Zwecke eignet, ist die Verhaltensbeurteilung mit Ratingskalen. Ratingverfahren im Allgemeinen erfahren großen Zuspruch und werden häufig zur Verhaltensbeurteilung eingesetzt (Voß & Gebhardt, 2017a). Methodisch erfolgt eine retrospektive Einschätzung von beobachteten Verhaltensweisen zumeist mittels standardisierter mehrstufiger Ratingskalen (Casale et al., 2015a; Huber & Rietz, 2015). Ein Beispiel eines solchen Verfahrens ist der Strengths and Difficulties Questionnaire (SDQ), welcher im Kapitel 4 tiefergehend beschrieben wird. Es steht nicht das Auszählen konkreter Verhaltensweisen im Vordergrund, wie bei systematischen Verhaltensbeobachtungen, sondern die Wahrnehmung der beurteilenden Person in Bezug auf die Verhaltensweisen (Huber & Rietz, 2015). Vorgegebene Verhaltensweisen werden durch Ankreuzen von Items auf einer Ratingskala beurteilt (Schmidt-Atzert et al., 2012). Die Items sollen ein spezifisches Konstrukt abbilden und sind somit reflektiver Art (Casale et al., 2015a). Auf diese Weise können Daten zur Häufigkeit und Dauer beobachteter Verhaltensweisen erhoben werden (ebd.). Verhaltensbeurteilungen mit Ratingskalen entsprechen einer abstrahierenden Methode (Huber & Rietz, 2015). Der Beobachtungszeitraum ist zumeist festgelegt und umfasst z.B. mehrere Unterrichtsstunden oder Wochen (ebd.). Verhaltensbeurteilungen können sich auf Beobachtungssequenzen beziehen, die bereits mehrere Tage oder Wochen zurückliegen (ebd.). Es können zudem verschiedene Schulfächer berücksichtigt werden (ebd.). Eine Ratingskala umfasst spezifische Verhaltensweisen (z.B. nimmt eigene Gefühle wahr), die verschiedenen Verhaltensdimensionen zugeordnet (z.B. Selbstwahrnehmung) und mithilfe einer Likert-Skala (z.B. von 0 bis 3) bewertet werden können (Casale et al., 2015b). Daten, die mit Ratingskalen erhobenen werden, können Anhaltspunkte für spezifische Verhaltensprobleme liefern (ebd.).

Verhaltensbeurteilungen mit Ratingskalen können im Gegensatz zur systematischen Verhaltensbeobachtung i.d.R. ökonomisch durchgeführt werden, da mittels Ratingskalen detaillierte Informationen über SchülerInnen schnell und unter Verwendung weniger Ressourcen erhoben werden können (Casale et al., 2015b; Schmidt-Atzert et al., 2012; Chafouleas et al., 2009a; Voß & Gebhardt, 2017a). Verhaltensbeurteilungen mit Ratingskalen können außerdem als valide und reliabel bezeichnet werden (Casale et al., 2015a). Es liegt ein spezifisches meistens reflexives Messmodell zugrunde (ebd.). Sie streben auf eindimensionale Darstellungen spezifischer Merkmalsausprägungen ab (ebd.). Da in der Regel spezifisch formulierte Items vorzufinden sind und Ratingskalen mit einer geringen Komplexität schlussfolgernder Kognitionen einhergehen, liegt eine niedrige Inferenz vor (ebd.). Ein Nachteil dieser Methode ist jedoch, dass die Beurteilungen der beobachtenden Personen auf Schlussfolgerungen zur möglichen

Bedeutung des Verhaltens beruhen und somit subjektiv geprägt sind (Schmidt-Atzert et al., 2012). Daher können Verhaltensbeurteilungen mit Urteilsfehlern einhergehen (Schmidt-Atzert et al., 2012). Ein möglicher Urteilsfehler ist der „Baseline-Error“ (Döring & Bortz, 2016, S. 255). Dieser kann insbesondere dann auftreten, wenn sich Beurteilungen auf größere Zeitspannen beziehen, wobei „die Auftretenswahrscheinlichkeit von Ereignissen [...] falsch eingeschätzt [wird], weil man sich nicht an der objektiven Häufigkeit, der sog. Baseline, orientiert, sondern irrtümlich besonders prägnante, im Gedächtnis gerade verfügbare oder typische Ereignisse irrtümlich für sehr wahrscheinlich hält“ (ebd., S. 255). Ein weiterer Urteilsfehler liegt dann vor, wenn Beurteilungen mit Ratingskalen mit einer mangelnden Differenzierung einhergehen (Döring & Bortz, 2016). Eine mangelnde Differenzierung meint, dass die beurteilenden Personen nicht die gesamte Breite der Skala zur Beurteilung heranziehen, sondern die Beurteilungen auf einen Teilbereich der Skala fokussieren (ebd.). Eine mangelnde Differenzierung kann Hinweise auf die Notwendigkeit einer Neukonstruktion der Ratingskala liefern (ebd.). Dabei sollte der Bereich, in dem die konzentrierten Beurteilungen vorzufinden sind, differenzierter gestaltet werden (ebd.). Aufgrund möglicher Urteilsfehler ist die Objektivität dieser Verfahren eingeschränkt (Casale et al., 2015a). Sie gehen zudem mit einer relativ langen Latenzzeit zwischen Beobachtungen und Beurteilungen einher (i.d.R. zwei Wochen) und erfolgen dementsprechend indirekt (Casale et al., 2015a). Des Weiteren sind sie mitunter nicht änderungssensitiv: Sie sind nicht dazu konzipiert häufig wiederholend eingesetzt zu werden und können somit kleine aber dennoch relevante Verhaltensänderungen nicht sensitiv erfassen (Casale et al., 2015b; Chafouleas et al., 2009a). Nur über längere Zeitspannen können Veränderungen sensitiv für Einzelfälle erfasst und Vergleiche im Hinblick auf die individuelle Bezugsnorm vorgenommen werden (Casale et al., 2015a). Darüber hinaus ist die Flexibilität der Verfahren begrenzt, da die Items festgelegt sind und signifikante Änderungen an den Skalen zu Verletzungen der Testgüte führen würden Chafouleas et al., 2009a). Dies schränkt die kontextuelle Relevanz der einzelnen Verfahren ein. Sie sind somit nicht für jeden Kontext von Relevanz (ebd.). Verhaltensbeurteilungen mit Ratingskalen können nur eingeschränkt für verlaufsdiagnostische Zwecke herangezogen werden.

Eine Gegenüberstellung beider Methoden zeigt, dass sie zusammen allen relevanten Gütekriterien der Verlaufsdiagnostik entsprechen (Casale et al., 2015a). In einer Kombination beider Methoden liegt somit das Potential für eine zur Verlaufsdiagnostik geeigneten Methode (ebd.). Mit dem Direct Behavior Rating (DBR) wurde im nordamerikansichen Raum eine entsprechende Methode entwickelt, die potentiell zur Evaluation von Verhaltensverläufen geeignet zu sein scheint (Casale et al., 2015b; Christ et al., 2009). Sie wird im Kapitel 5 vorgestellt. Im nachfolgenden Kapitel wird zunächst der SDQ als ein weit verbreitetes Screeninginstrument der Verhaltensbeurteilung mit Ratingskalen erläutert.

4. Strengths and Difficulties Questionnaire

SchülerInnen mit Verhaltensschwierigkeiten können mithilfe von Screeningverfahren identifiziert werden und so frühzeitig von notwendigen Förder- und Präventionsmaßnahmen profitieren (Dever et al., 2015). Nachfolgend wird ein solches Screeningverfahren, der Strengths and Difficulties Questionnaire (SDQ) vorgestellt. Entwickelt wurde der SDQ im englischsprachigen Raum von Goodman (1997) mit dem Ziel das prosoziale und problematische Verhalten von Kindern und Jugendlichen als Verhaltensscreening erfassen zu können (Goodman, 1997). International ist es eines der am weitesten verbreiteten Instrumente zur Erfassung von Verhaltensweisen bei Kindern und Jugendlichen (DeVries et al., 2017; Goodman et al., 2010). Auf Basis des Instrumentes werden Entscheidungen von hoher pädagogischer oder schulischer Relevanz getroffen (Voß & Gebhardt, 2017a). Es verfügt über ein kompaktes Format und besteht aus einem kurzen, einseitigen Fragebogen (Goodman, 1997; 2001). Es ist kostenfrei im Internet (www.sdqinfo.com) verfügbar und kann ökonomisch bearbeitet und ausgewertet werden (Klasen, Woerner, Rothenberger & Goodman, 2003). Goodman und Scott (1999) geben an, dass eine Bearbeitung lediglich 5 Minuten dauert. Es weist zudem eine hohe Flexibilität auf und stellt ein „sehr praxisfreundliches Verfahren“ mit einem umfangreichen Einsatzgebiet dar (Koglin, Barquero, Mayer, Scheithauer & Petermann, 2007, S. 175) . Die Einsatzgebiete erstrecken sich über Forschung und Praxis (Goodman, 2001; Voß & Gebhardt, 2017a). Es kann im schulischen Setting für Screeningzwecke (Goodman, Ford, Simmons, Gatward & Meltzer, 2000; Voß & Gebhardt, 2017a) und zur Erfassung der psychischen und psychosozialen Entwicklung eingesetzt werden (DeVries et al.., 2017). Die mithilfe des SDQs identifizierten Problemfelder und Ressourcen dienen als Anknüpfungspunkte für weitere differenzierte Diagnoseprozesse und als Grundlage für die Auswahl angemessener Unterstützungs- und Interventionsmaßnahmen (Voß & Gebhardt, 2017a; Woerner er al., 2004). Es kann erhoben werden, welche Kinder darin Unterstützung benötigen Verhaltensprobleme zu bewältigen (Kersten et al., 2016).

Inhaltlich erhebt der SDQ sowohl Stärken als auch Verhaltensprobleme von Kindern und Jugendlichen (Goodman, 1997). Er erfasst die häufigsten kindlichen Symptome von Verhaltensproblemen und stimmt oftmals sinngemäß und in seinen Funktionen mit anderen Verfahren wie z.B. der Child Behavior Checklist (CBCL) überein (DeVries et al., 2017; Klasen et al, 2003; Koglin et al., 2007). Insgesamt ist der SDQ in drei verschiedenen Versionen verfügbar: für Lehrer, Eltern und als Selbstbeurteilungsbogen. Die Versionen sind nahezu identisch und wurden nur geringfügig adaptiert (Koglin et al., 2007; Voß & Gebhardt, 2017a). Die Lehrkraft- und Elternversion ist für Kinder und Jugendliche im Alter von 4 bis 17 Jahren konzipiert worden, die Selbstbeurteilungsversion für Kinder und Jugendliche im Alter von 11 bis 17 Jahren (Strengths and Difficulties Questionnaire, 2015). Der SDQ umfasst 25 Items, welche mithilfe einer dreistufigen Likert-Skala („nicht zutreffend“, „teilweise zutreffend“, „eindeutig zutreffend“)

erhoben werden (Strengths and Difficulties Questionnaire, 2015). Die Beurteilung darüber, inwieweit das Kind die entsprechenden Verhaltensweisen gezeigt hat, beziehen sich auf die vergangenen sechs Monate (Haller et al., 2016). Zudem existiert eine Version für Kinder im Vorschulalter (2 - 4 Jahre), die von Eltern und Erzieherinnen und Erziehern ausgefüllt werden kann (Strengths and Difficulties Questionnaire, 2015). Des Weiteren gibt es doppelseitige Versionen, in denen zusätzlich danach gefragt wird, ob bei dem Kind Schwierigkeiten vorliegen und wenn ja, worauf diese Schwierigkeiten Auswirkungen haben (Strengths and Difficulties Questionnaire, 2015).

Unabhängig der verschiedenen Versionen gibt es drei unterschiedliche Ansätze zur Auswertung des Instrumentes: ein 5-Faktor-, ein 3-Faktor- und ein Bi-Faktormodell. Es herrscht Uneinigkeit darüber, welche dimensionale Struktur zur Interpretation vorzuziehen ist (DeVries et al., 2017; Goodman et al., 2010). Als Hauptmodelle sind das 5-Faktor- und das 3-Faktormodell aufzuführen (Goodman et al., 2010). Das 5-Faktormodell wurde basierend auf den Hauptkategorien aktueller Klassifikationssysteme zu psychischen Störungen von Kindern und Jugendlichen wie dem „Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)“ entwickelt (ebd.). Es besteht aus fünf Skalen („Hyperaktivität“, „Verhaltensprobleme“, „Verhaltensprobleme mit Gleichaltrigen“, „Emotionale Probleme“ und „Prosoziales Verhalten“) mit jeweils fünf Items (Casale et al., 2017). Die vier erstgenannten Skalen können zu einem Problemfaktor zusammengefasst werden und ergeben in Summe den Gesamtproblemwert (Lohbeck, Schultheiß, Petermann & Petermann, 2015b). Voß und Gebhardt (2017a) bezeichnen den Gesamtproblemwert als den „wichtigsten Wert“ des Instrumentes. International existieren Normen, die eine Einordnung des Gesamtproblemwertes und der fünf einzelnen Skalen in die Kategorien „normal“, „grenzwertig“ und „auffällig“ ermöglichen (Koglin et al., 2007; Voß & Gebhardt, 2017a). Dem Problemfaktor steht im Bi-Faktormodell die Skala „Prosoziales Verhalten“ gegenüber (Kóbor, Takács & Urbán, 2013). Wird der Problemfaktor in die zwei Subskalen „Externalisierende Verhaltensprobleme“ und „Internalisierende Verhaltensprobleme“ unterteilt, so ergibt sich unter Berücksichtigung der Skala „Prosoziales Verhalten“ das 3-Faktormodell (Goodman et al., 2010). Die Skala „Externalisierende Verhaltensprobleme“ wird aus den Skalen „Hyperaktivität“ und „Verhaltensprobleme“ und die Skala „Internalisierende Verhaltensprobleme“ aus den Skalen „Verhaltensprobleme mit Gleichaltrigen“ und „Emotionale Probleme“ gebildet (ebd.). Die 5-faktorielle Struktur konnte mittels Faktorenanalyse u.a. für die englischsprachige Version (Goodman, 1997; 2001; Goodman et al., 2010), die deutschsprachige Elternversion (Woerner et al., 2002), die deutschsprachige Lehrer- / Erzieherversion (Koglin et al., 2007; Rothenberger, Erhart, Wille, Ravens-Sieberer & die BELLA Arbeitsgruppe, 2008; Saile, 2007) und die deutsche Selbstbeurteilungsversion (Lohbeck, et al., 2015b) bestätigt werden. Es konnte außerdem aufgezeigt werden, dass die Skalen in ein 3-faktorielles Modell gegliedert werden können (DeVries et al., 2017; Di Riso et al., 2010; Goodman et al., 2010).

Beide Modelle konnten gute Passungen aufweisen (DeVries et al., 2017; Essau et al., 2012; Niclasen, Skovgaard, Andersen, Niclasen, Skovgaard, Andersen, Sømhovd & Obel, 2013). Goodman et al. (2010) kommen zu dem Schluss, dass die 3-faktorielle Struktur zwar nicht die 5-faktorielle Struktur im Sinne eines Modells erster Ordnung ersetzen sollten. Die 3-faktorielle Struktur könne jedoch als Modell zweiter Ordnung verwendet werden und im Zuge dessen als übergeordnete Faktorstruktur die Subskalen subsumieren (Goodman et al., 2010). DeVries et al. (2017) empfehlen die Verwendung des 5-Faktormodells zur Erfassung spezifischer Verhaltensweisen und -schwierigkeiten, welche übergeordnet externalisierenden oder internalisierenden Verhaltensweisen zugeordnet werden können. Goodman et al. (2010) geben an, dass die Verwendung des 5-Faktorenmodells nur für Kinder mit hohen Risikofaktoren, die psychische Störungen und/oder hohe Ergebnisse in den Subskalen aufweisen, gerechtfertigt sei. Demgegenüber empfehlen sie die Verwendung des breiteren 3-Faktorenmodells für Untersuchungen von Stichproben mit Kindern und Jugendlichen, die wenige Risikofaktoren aufweisen. Dies sei der konservativere Ansatz mit dem zum einen eine zutreffende Beschreibung der erhobenen Informationen sichergestellt werden könnte und zum anderen könnten auf diese Weise vergleichbare Erkenntnisse generiert werden (ebd.). Sie merken an, dass die optimale Wahl des Faktorenmodells von der jeweiligen Stichprobe und den Forschungszielen abhängt (ebd.). Neben dem 3-faktoriellen und dem 5-faktoriellen Modell konnte in einer Studie auch die dimensionale Struktur des Bi-Faktormodells bestätigt werden (Kóbor et al., 2013).

Die psychometrischen Eigenschaften des SDQ sind in einer Vielzahl nationaler und internationaler Studien untersucht worden. Die Forschungsergebnisse diesbezüglich sind sehr heterogen (Lohbeck, et al., 2015b). Saile (2007) und Koglin et al. (2007) konnten für die Lehrerversion hohe Reliabilitätswerte feststellen. Im Vergleich konnten Lohbeck et al. (2015) für die Selbstbeurteilungsversion zufriedenstellende Werte erheben (Lohbeck, et al., 2015b). Die Reliabilität der Elternverison wurde von Woerner et al. (2002) ermittelt lag für die Einzelskalen bei Werten zwischen α=.58 und α=.76, wohingegen der Wert der Gesamtproblemskala einen höheren Wert von .82 aufweist. Deutschsprachige Normierungen sind von Woerner et al. (2002) für die Elternversion, von Koglin et al. (2007) für die Lehrer-/Erzieherversion und von Altendorfer-Kling; Ardelt-Gattinger & Thun-Hohenstein (2007) für die Selbstbeurteilungsversion durchgeführt worden.

Voß und Gebhardt (2017a) geben an, dass die statusdiagnostische Eignung hinreichend erforscht worden ist. Da aktuell Verfahren fehlen, die zur Verlaufsdiagnostik eingesetzt werden können, ist es von Bedeutung zu erheben, inwieweit der SDQ für verlaufsdiagnostische bzw. Monitoring-Zwecke eingesetzt werden kann (Voß & Gebhardt, 2017a). Diesbezüglich gilt es zu erfassen, ob das Instrument die psychometrische Eigenschaft der Messinvarianz besitzt (DeVries et al., 2017; Voß & Gebhardt, 2017a). DeVries et al. (2017) konnten aufzeigen, dass sowohl das 3- als auch für das 5-Faktorenmodell eine starke Messinvarianz aufweisen. Auch

Voß und Gebhardt (2017a) konnten in ihrer Studie eine grundlegende Messinvarianz über vier Messzeitpunkte feststellen

Auf Basis der Forschungslage, der Reliabilitäts- und Validitätswerte, kann der SDQ als ein nützliches Instrument für Screeningzwecke und zur Erfassung problematischen und prosozialen Verhaltensweisen angesehen werden (Döpfner & Petermann, 2012; Goodman, 2001; Lohbeck et al., 2015b). Voß und Gebhardt (2017a) geben an, dass der SDQ grundsätzlich, unter Berücksichtigung einiger Überarbeitungshinweise, in Zeitspannen wie viertel- oder halbjährlich zur Verlaufsdiagnostik eingesetzt werden kann. Sie raten jedoch von einem täglichen oder wöchentlichen Einsatz ab, da der Bezugsrahmen zu breit und die Anzahl der Items zu gering sei. Für einen täglichen oder wöchentlichen Einsatz verweisen die Autoren auf den Ansatz des Direct Behavior Ratings (ebd.). Voß und Gebhardt (2017) empfehlen in schulischen Settings breite Verfahren wie das SDQ, mit Direct Behavior Ratings zu kombinieren, da diese spezifische Messungen erlauben. Sie sehen dabei eine Entwicklung des DBRs aus dem SDQ als vielversprechend an. Im Zuge dessen sollten Items hinzugefügt werden, die grenzwertige Ausprägungen von Verhaltensauffälligkeiten erfassen können und über eine geringe Inferenz verfügen (ebd.). Die Items sollten somit eine eindeutige Operationalisierung aufweisen und ihnen sollten eingängige Sachverhalte zugrunde liegen (ebd.). Voß und Gebhardt (2017) bezeichnen diese Items als „leichtere“ Items und schlagen als Beispiele Items aus dem Bereich des Arbeitsverhaltens vor, wie „Redet oft dazwischen“ oder „Meldet sich häufig im Unterricht“ (ebd., S. 31). Denn diese weisen insbesondere im schulischen Setting eine hohe Relevanz auf (ebd.). Des Weiteren sollten Items hinzugefügt werden, auf deren Basis solche Kinder differenziert beurteilt werden können, die charakteristische Verhaltensweisen für Verhaltensauffälligkeiten zeigen (ebd.). Es geht somit um Items, die im „oberen Messbereich“ liegen, wie „Beleidigt Mitschülerinnen und Mitschüler“ oder „Stört den Unterricht“ (ebd., S. 31). Der Ansatz des Direct Behavior Ratings wird im nachfolgenden Kapitel 5 vorgestellt wird.

5. Direct Behavior Rating

Eine evidenzbasierte pädagogische Praxis erfordert Verfahren, die Fördererfolge in Einzelfällen überprüfen können, wie beispielsweise Verfahren zur Verlaufsdiagnostik (Casale, 2017). Zur engmaschigen Erfassung von Schülerverhalten fehlen aktuell evaluierte verlaufsdiagnostische Verfahren (ebd.). Die bisher genutzten diagnostischen Instrumente im Förderschwerpunkt emotionale und soziale Entwicklung entsprechen den Gütekriterien der Verhaltensverlaufsdiagnostik nicht vollständig und können deswegen nur eingeschränkt zur Analyse von Verhaltensverläufen herangezogen werden (Casale et al., 2015a). Als alternativer und vielversprechender Ansatz zur Verhaltensverlaufsdiagnostik wird das aus dem nordamerikanischen Raum bekannte Direct Behavior Rating (DBR) in der Forschungsliteratur diskutiert (Christ et al., 2009; Huber & Rietz, 2015). Dieses Verfahren hat sich bereits in internationalen als auch in ersten deutschen Studien zur Verhaltensverlaufsdiagnostik bewähren können (Christ et al.,

2009; Huber & Rietz, 2015). Im deutschen Sprachraum findet sich neben dem Begriff des Direct Behavior Ratings (DBR) auch die deutsche Übersetzung der Direkten Verhaltensbeurteilung (DVB) (vgl. z.B. Huber & Rietz, 2015). DBR ähnliche Verfahren existieren bereits seit längerer Zeit (Chafouleas, Hagermoser Sanetti, Jaffery & Fallon, 2012b). Mit der systematischen Konzeptualisierung und Evaluation des Einsatzes der sog. „daily behavior report cards" (DBRC) wurde zunehmend der Begriff des Direct Behavior Ratings geprägt (Chafouleas, Riley-Tillman & McDougal, 2002; Christ et al., 2009). DBRC sind mittlerweile als ein spezifisches Verfahren der übergeordneten Methode DBR anzusehen (Christ et al., 2009). Den Begriffswechsel begründen Christ et al. (2009) durch die Direktheit der Methode (Direct), die Betonung des Aspektes Verhalten (Behavior) und durch die Erhebung der Informationen auf Basis von Ratingverfahren (Rating). Die Autoren bezeichnen diese drei Aspekte als die wesentlichen Eigenschaften des DBRs (Christ et al., 2009). Die Eigenschaft der Direktheit umfasst, dass Beurteilungen unmittelbar im Anschluss an die Beobachtung, also in direkter Nähe zum Auftreten des Verhaltens, durchgeführt werden (Casale et al., 2015a; Christ et al., 2009) (Christ et al., 2009). Dabei liegt ein gewisser Interpretationsspielraum vor, der sich durch die Beobachtungszeiträume begründet. Die Beobachtungszeiträume können wenige Sekunden bis hin zu einem ganzen Tag lang sein und während oder nach den Beobachtungen durchgeführt werden können (Christ et al., 2009). Je nachdem wie ökonomisch das Verfahren eingesetzt werden soll und wie hoch die Wahrscheinlichkeit des Auftretens des Verhaltens ist, können in Orientierung an den Kontext und den Zweck der Beurteilung entsprechende Vorgehensweisen ausgewählt werden (ebd.). Wenn Beurteilungen nicht unmittelbar nach der Beobachtung des Verhaltens durchgeführt werden können, so sollten sie dennoch vor der Beobachtung neuer relevanter Situationen durchgeführt werden, um den potentiellen Einfluss der dazwischenliegenden Beobachtungen zu begrenzen (Christ et al., 2009). Im Sinne der zweiten Eigenschaft, der Verhaltenskomponente, sollten alle am Förderprozess beteiligten Personen die interessierenden Verhaltensweisen verstehen und nachvollziehen können (Casale et al., 2015a; Christ et al., 2009). Dazu ist eine sorgfältige Operationalisierung notwendig. Anstelle von abstrakten hypothetischen Konstrukten sollten konkret beobachtbare Verhaltensweisen vorliegen (ebd.). Im Zuge der dritten Eigenschaft, der Ratingkomponente steht die Einschätzung des Verhaltens mithilfe von Ratingskalen im Vordergrund (ebd.). Da Verhaltenseinschätzungen von der Wahrnehmung und dem Eindruck der Rater beeinflusst werden, wird eine Quantifizierung der Wahrnehmung des Raters auf Basis einer Ratingskala notwendig (Christ et al., 2009).

Konzeptionell setzt sich der DBR-Ansatz aus den Methoden der Verhaltensbeurteilung mit Ratingskalen und der systematischen direkten Verhaltensbeobachtung zusammen (Casale et al., 2015b). Er kombiniert die Vorteile beider Methoden, indem ein konkret operationalisiertes

Verhalten in einer für dieses Verhalten relevanten Situation beobachtet und unmittelbar danach mittels Ratingskala beurteilt wird (Chafouleas, 2011). Huber und Rietz (2015) bezeichnen es als Ziel der direkten Verhaltensbeurteilung die Zeit zwischen der Beobachtung und der Beurteilung des Verhaltens zu verringern, sodass neben einer ökonomischen Verhaltenseinschätzung eine hohe „situative Validität" (S. 79) gewährleistet werden kann. Diese Kombination von Ökonomie und Direktheit macht DBR-Verfahren zu einer einzigartigen Methode der Verhaltensverlaufsdiagnostik (Chafouleas et al., 2009a).

Das Kriterium der Ökonomie greifen Christ et al. (2009) als eines von vier leitenden Prinzipien für DBR-Verfahren auf: „DBR should be *defensible, flexible, repeatable, and efficient"* (S. 207). Das erste Prinzip („defensible") fordert eine angemessene Testgüte (ebd.). Überprüfungen der Testgüte sollten fortlaufend erfolgen, um die DBR-Verfahren und Instrumente weiterentwickeln zu können (ebd.). Es sollte überprüft werden, ob mithilfe des Verfahrens Verhaltensverläufe von SchülerInnen reliabel, valide und möglichst genau erfasst werden können (Casale et al., 2017). Diesbezüglich ist eine systematische Forschung notwendig, da DBR-Verfahren auf Fremdeinschätzungen der Rater beruhen, welche mitunter von „teilweise abhängigen systematischen Fehlerquellen beeinflusst" werden (Casale et al., 2015c, S. 259).

Weiterhin sollten DBR-Verfahren im Sinne des zweiten Prinzips flexibel („flexible") sein und für eine Vielzahl von Verhaltensweisen, Zwecken und in zahlreichen Kontexten einsetzbar sein (Christ et al., 2009). Die Methode wird nicht definiert durch ein spezifisches Instrument oder eine spezifische Gruppe von Verhaltensweisen, sondern kann flexibel angepasst werden (ebd.). Dieses hohe Maß an Flexibilität bringt bei der Überprüfung der psychometrischen Testgüte große Herausforderungen mit sich, denn DBR-Verfahren können unter verschiedenen Messbedingungen von unterschiedlichen Ratern angewendet werden (Casale et al., 2017).

Des Weiteren sollten DBR-Verfahren ökonomisch („efficient") einsetzbar sein. Ein ökonomischer Einsatz ist dann gegeben, wenn Personen aus dem jeweiligen Setting mittels Ratingskalen Beurteilungen auf Basis von kurzen Beobachtungssituationen durchführen und diese Situationen zugleich mit wenigen Unterbrechungen und Störungen einhergehen (Casale et al., 2015c; Christ et al., 2009). Ein weiterer wichtiger Aspekt ist diesbezüglich die präzise Operationalisierung der Zielverhaltensweisen (Casale et al., 2017). Zudem sollten DBR-Verfahren wiederholbar („repeatable") sein, also in kurzen Abständen hochfrequente und regelmäßige Messungen ermöglichen (Casale et al., 2015a). Das Merkmal der Wiederholbarkeit ist entscheidend, da es eine fortlaufende Erfassung von Schülerdaten innerhalb und zwischen Kontexten ermöglicht, auf deren Basis wiederum das temporäre Verhalten und Verhaltensverläufe evaluiert und abgebildet werden können (Christ et al., 2009).

Chafouleas et al. (2009a) beschreiben zusätzlich das Prinzip „feasible" auf, welches als durchführbar oder praktikabel übersetzt werden kann. DBR-Verfahren sollten folglich praktikabel und nützlich sein.

Werden diese Prinzipien eingehalten, kann die Verwendung von Direct Behavior Ratings in einer Vielzahl von Situationen und für eine Vielzahl von Zwecken sichergestellt werden (Christ et al., 2009). Unterschiedlichen Personen, wie Lehrern oder Eltern, ist es dann möglich das DBR-Verfahren im Sinne einer evidenzbasierten Methode einzusetzen (ebd.). Durchgeführt wird die Methode wie folgt: Ein Rater beobachtet eine bestimmte Anzahl zuvor festgelegter und konkret operationalisierter Verhaltensweisen (wie z.B. konzentriertes Arbeiten) in einem spezifischen festgelegten Zeitraum und in einer Situation, in der diese Verhaltensweisen bedeutsam sind (hier z.B. Stillarbeitsphasen) und beurteilt diese Verhaltensweisen auf Basis einer kurzen Ratingskala unmittelbar nach dem Auftreten der Verhaltensweisen (Casale et al., 2017; Chafouleas et al, 2012a; Chafouleas et al., 2012b). DBR-Verfahren können in Kontexten und Settings, in denen datenbasierte Entscheidungen in multiprofessionellen Teams getroffen werden müssen, wie z.B. der Schule, zum Einsatz kommen (Christ et al., 2009). Damit die Ergebnisse zur datenbasierten Entscheidungsfindung herangezogen werden können, sollten sie eine hohe kontextuelle Relevanz aufweisen (Chafouleas et al., 2009a). Im schulischen Setting werden sie zur Erfassung von Verhaltensverläufen im Unterricht und zur Evaluation pädagogischer Maßnahmen hinsichtlich der Wirksamkeit eingesetzt (Casale et al., 2017; Chafouleas, 2011). Sie eignen sich somit für den Einsatz in präventiv gestuften Förderprogrammen wie beispielsweise dem RTI-Ansatz (Huber & Rietz, 2015). Werden die Evaluationen der Verhaltensverläufe als „Response“ in Abhängigkeit zu den eingesetzten Fördermaßnahmen („Intervention“) evaluiert, dann kann frühzeitig über die Wirksamkeit eingesetzter Fördermaßnahmen entschieden und es können neue Ansätze für die weitere pädagogische Arbeit abgeleitet werden (Casale et al., 2015b; Huber & Rietz, 2015). Die engmaschige und ressourcenorientierte Begleitung von Lern- und Förderprozessen mithilfe der DBR-Verfahren trägt auf diese Weise „zur Evidenzbasierung von konkreten Fördermaßnahmen im Einzelfall“ bei (Casale, 2017, S. 4). DBR-Verfahren können folglich einen sinnvollen Beitrag zur Diagnostik im sonderpädagogischen Bereich leisten (Casale et al., 2015c).

In erster Linie sind DBR-Verfahren als förderdiagnostische Instrumente anzusehen, welche eine Überprüfung der Wirkungen eingesetzter Fördermaßnahmen zum Ziel haben und nicht die Etikettierung von SchülerInnen (Huber & Rietz, 2015). Der primäre Einsatzbereich liegt dementsprechend in der Prozess- und Verlaufsdiagnostik (ebd.). Huber und Rietz (2015) bezeichnen dies als das besondere Merkmal des DBR. Dabei ist es weniger von Bedeutung, wie das Verhalten im Verhältnis zum Durchschnitt beurteilt wird, sondern vielmehr, ob eine Entwicklung über die Zeit zu verzeichnen ist (ebd.). Wie die Ergebnisse interpretiert werden, hängt davon ab, welche Bezugsnorm zur Orientierung dient. Die Bezugsnorm sollte in Abhängigkeit zum Ziel ausgewählt werden (Casale et al., 2015a). Eine soziale Bezugsnorm sollte dann gewählt werden, wenn problematische Verhaltensweisen identifiziert oder die Ausprägungen dieser Verhaltensweisen eingeschätzt werden (ebd.). Sollen hingegen Verhaltensveränderungen

untersucht und abgebildet werden, eignet sich die individuelle Bezugsnorm (ebd.). DBR-Verfahren, die für verschiedene Zwecke eingesetzt werden, z.B. Screening oder Verlaufsdiagnostik, können unterschiedliche zeitliche und personelle Ressourcen erfordern (Chafouleas et al., 2010).

Im nordamerikanischen Raum wird der DBR-Ansatz bereits seit einigen Jahren erforscht (Huber & Rietz, 2015). Dort konnte sich die Methode anfänglich bewähren (Christ et al., 2009). Forschungsergebnisse aus dem englischsprachigen Raum können nur bedingt auf den deutschsprachigen Raum übertragen werden. Es müssen neue DBR-Verfahren entwickelt und hinsichtlich der Testgüte überprüft werden (Casale et al., 2015a). Insgesamt liegen wenige Forschungsergebnisse vor. Zum einen werden diese erst seit einem relativ kurzen Zeitraum erforscht. Zum anderen kann die Methode unter verschiedensten Bedingungen eingesetzt werden, welche es zu berücksichtigen gilt (Casale, 2017; Huber & Rietz, 2015). Die vorliegenden Forschungsergebnisse sind tendenziell positiv und zeigen auf, dass es sich bei dem DBR um ein reliables und valides Instrument handelt (Casale, 2017) (Casale, 2017; Huber & Rietz, 2015). Die meisten Forschungsbefunde liegen zur psychometrischen Güte von Single-Item-Skalen vor und liefern überwiegend Hinweise zur statusdiagnostischen Testgüte der Verfahren (Casale et al., 2015a). Nachfolgend werden entsprechende Forschungsergebnisse erläutert.

Während Chafouleas et al. (2007) und Volpe & Briesch (2012) in ihren Studien einen bedeutenden Anteil der Varianzaufklärung auf die Facette der Rater zurückführen konnten, konntenChafouleas et al. (2010) sowie Casale et al. (2015c) lediglich einen geringen Varianzanteil feststellen. Casale et al. (2015c) setzen ihre Ergebnisse zur Interraterreliabilität in Bezug zu anderen Beurteilungsverfahren und stellen eine „mehr als akzeptable“ Interraterreliabilität fest (S. 265). Auch Christ, Riley-Tillman, Chafouleas & Jaffery (2011) konnten in ihrer Studie eine moderate bis hohe Interrater-Reliabilität erheben und geben an, dass sogar unerfahrene Rater für viele Verhaltensweisen grundlegend konsistente Ratings liefern. Huber und Rietz (2015) führen auf, dass insgesamt eine angemessene Interrater-Reliabilität vorliegt. Diesbezüglich ist es von Bedeutung, welche Personengruppe die Verhaltensbeurteilungen durchgeführt hat (Casale et al., 2017). Es wird davon ausgegangen, dass individuelle Eigenschaften der Lehrkräfte die Ergebnisse der DBR-Verfahren beeinflussen, wie z.B. Vertrautheit mit den betreffenden SchülerInnen, konkurrierende Anforderungen während der Beobachtungsphase und Vorerfahrungen in der Arbeit mit verlaufsdiagnostischen Verfahren (Chafouleas et al.., 2010) In ihrer Studie konnten Chafouleas et al., 2010 aufzeigen, dass sich Reliabilitätsschätzungen in Abhängigkeit zu den verschiedenen Ratern (Lehrer, geschulte Beobachter) unterschieden. Für die Klassenlehrer konnten nach zehn Beobachtungen akzeptable Reliabilitätswerte für den Zweck relativer und absoluter Entscheidungsfindungen ermittelt werden, wohingegen die Werte der geschulten Beobachter unterhalb des Grenzwertes lagen. Als Rater kommen somit

primär Personen infrage, die regelmäßig im unmittelbaren Kontakt zu den betreffenden SchülerInnen stehen und im natürlichen Setting der Beobachtungssituation agieren, also Personen, die hochfrequente Beurteilungen gewährleisten können (Chafouleas et al., 2010; Christ et al., 2009). Das können beispielsweise Lehrer, pädagogische Fachkräfte oder Eltern sein (Casale et al., 2015b). Grundsätzlich kann davon ausgegangen werden, dass Lehrkräfte dazu in der Lage sind relativ valide Auskünfte zum Schülerverhalten zu tätigen (Landscheidt, 2001). Aber auch die SchülerInnen selbst kommen als Rater infrage, sie können Selbstbeurteilungen ihres Verhaltens durchführen (Christ et al., 2009). Im Vergleich zu SDO-Verfahren ist es bei DBR-Verfahren von immenser Bedeutung, welcher Rater das Verhalten beurteilt (Huber & Rietz, 2015). Untersuchungen der konkurrenten Validität, die DBR-Verfahren den direkten systematischen Verhaltensbeobachtungen gegenüberstellen, zeigen, dass die Facette der Rater tendenziell einen größeren Einfluss auf die Beurteilungen mittels DBR-Verfahren hat als auf die Beobachtungen mittels direkter systematischer Verhaltensbeobachtung (Huber & Rietz, 2015). So konnten beispielsweise Briesch, Chafouleas und Riley-Tillman (2010) in ihrer Studie aufzeigen, dass der Großteil der Varianz in den Beobachtungen der direkten systematischen Verhaltensbeobachtung durch die Facette der Schüler aufgeklärt wurde, wohingegen der Großteil der DBR-Varianzaufklärung auf die Facette Rater zurückzuführen ist (Briesch, Chafouleas & Riley-Tillman, 2010). Folglich sind statusdiagnostische Entscheidungen, die auf Basis von DBR-Daten getroffen werden etwas unsicherer als Entscheidungen, die auf Daten beruhen, die mittels direkter systematischer Verhaltensbeobachtung erhoben wurden (Huber & Rietz, 2015). Zugleich konnte heraus gestellt werden, dass beide Methoden intraindividuelle Veränderungen im gleichen Maße sensitiv abbilden (Briesch et al., 2010). Diese Ergebnisse werden von Chafouleas et al. (2012a) gestützt. In weiteren Studien wurden die DBR-Daten den Daten von standardisierten Verfahren zu Verhaltensbeurteilung gegenüber gestellt (Chafouleas, Kilgus & Hernandez, 2009b; Kilgus, Chafouleas, Riley-Tillman & Welsh, 2012). Es konnten in beiden Studien moderate bis hohe Übereinstimmungen zwischen den beiden Verfahren und somit eine angemessene Kriteriums-Validität festgestellt werden (Chafouleas et al., 2009b; Kilgus et al., 2012).

Bei größeren Klassenteams bietet sich die Gelegenheit, dass mehrere Personen das Verhalten parallel beurteilen (Casale, 2017). Die Anzahl der Rater, die parallel raten, hat einen Einfluss auf die Ergebnisse (Chafouleas et al., 2007; Christ, Riley-Tillman, Chafouleas & Boice, 2010). Casale et al. (2015c) konnten belegen, dass mit der Erhöhung der Anzahl der Rater die Generalisierbarkeit und Zuverlässigkeit der erhobenen Schülerdaten zwar ansteigt, akzeptable Werte jedoch bereits für eine Anzahl von zwei Ratern festzustellen sind. Von grundlegender Bedeutung ist dabei, dass eine Beurteilung konsistent von derselben Person durchgeführt wird, da numerisch invariante Beurteilungen zu erwarten sind (Casale, 2017).

Casale et al. (2017) konnten weiterhin herausstellen, dass die Facette Rater nur bei global formulierten Single-Item-Skalen (SIS) einen substanziellen Varianzanteil aufklärt. Hingegen beeinflusst die Facette Rater die Zuverlässigkeit nicht substantiell, wenn konkret formulierte Multi-Item-Skalen (MIS) verwendet werden (ebd.).

Die Stabilität der Raterurteile über die Zeit sollte im Hinblick auf die prozessdiagnostische Eignung von DBR-Ratingskalen untersucht werden (Huber & Rietz, 2015). Huber und Rietz (2015) bezeichnen diesen Aspekt als eine bedeutende Voraussetzung für einen Einsatz des DBRs im Sinne der Prozessdiagnostik. Insgesamt können für die Test-Retest-Reliabilität des DBRs relativ hohe Werte angegeben werden, sofern die Verhaltensbeurteilungen auf mehreren Messungen basieren (Riley-Tillman, Christ, Chafouleas, Boice-Mallach & Briesch, 2011). Dabei fällt die Test-Retest-Reliabilität unterschiedlich in Abhängigkeit zur Situation, in der beobachtet wurde und zur Länge der Beobachtungssequenz aus. Im Allgemeinen kann der Beobachtungszeitraum wenige Sekunden bis einen ganzen Schultag umfassen und sollte in Orientierung an den Kontext und Zweck der Beurteilung ausgewählt werden (Christ et al., 2009). Für kürzere Sequenzen (10-minütig) wurden dabei etwas niedrigere Werte erhoben als für längere (20-minütige) Beobachtungssequenzen (Riley-Tillman et al., 2011). Riley-Tillman et al. (2011) konnten zudem für einen Beobachtungszeitraum (20 min), der in vier Sequenzen (je 5 min) unterteilt wurde, höhere Test-Retest-Reliabilitätswerte über einen Zeitraum von einer Woche feststellen als für einzelne Messungen. Generell führt die Bildung eines Mittelwertes über mehrere Messungen zu höheren Reliabilitätswerten (Riley-Tillman et al., 2011; Volpe & Briesch, 2015). Huber und Rietz (2015) schlussfolgern, dass diese relativ hohe Übereinstimmung dafür spricht, dass auch unerfahrene Rater Verhalten auf Basis von DBR-Verfahren zeitlich überdauernd beurteilen können. Sie geben an, dass die intrapersonelle Einschätzung von Schülerverhalten als relativ stabil eingeschätzt werden kann, wenn sich diese ersten positiven Ergebnisse in weiteren Studien nachweisen lassen (Huber & Rietz, 2015).

Ein weiterer bedeutender Varianzanteil unter Verwendung der DBR-Methode konnte auf die Interaktion zwischen Ratern und Schülern zurückgeführt werden (Briesch et al., 2010; Casale et al., 2015c; Chafouleas et al., 2010), wohingegen die Interaktion zwischen Ratern und Schülern bei direkten systematischen Verhaltensbeobachtungen nur mit einem geringen Varianzanteil auftrat (Briesch et al., 2010). Dieser Varianzanteil ist auch dann hoch, wenn die Rater zuvor umfangreich geschult wurden (Casale et al., 2015c). Als systematische Fehlerquelle muss die Subjektivität von Verhaltenseinschätzungen Beachtung finden (ebd.). Diese systematische Fehlerquelle kann auch trotz intensiver Schulungen nicht ausgeschlossen werden (ebd.). Eine mögliche Erklärung kann der Halo-Bias liefern (ebd.).

Vorliegende Forschungsarbeiten deuten darauf hin, dass sich bereits eine kurze Schulung positiv auf die Gütekriterien der Validität und Reliabilität auswirken kann (Huber & Rietz, 2015). Während Schlientz et al. (2009) in ihrer Studie aufzeigen konnten, dass genauere Ratings

möglich waren, wenn die Rater zuvor an einer Schulung teilgenommen haben, konnte in einer weiteren Studien festgestellt werden, dass bereits kurze Schulungen ausreichen, um genaue Ratings zu erhalten (LeBel, Kilgus, Briesch & Chafouleas, 2010). Intensive Schulungen sind dann notwendig, wenn wenig eindeutige Verhaltensweisen in qualitativ mittlerer Ausprägung beurteilt werden sollen (Huber & Rietz, 2015). Grundsätzlich können jedoch bereits Laienbeobachter nach kurzer Schulung zufriedenstellende Beurteilungen vornehmen (ebd.). Somit ist es von Relevanz, dass Schulungen im Alltag durchgeführt werden und dass die Beobachtungsgüte der Rater wiederholt überprüft wird (ebd.).

Auf die Facette der Person konnte eine moderate bis hohe Varianzaufklärung zurückgeführt werden (Briesch et al., 2010; Casale et al., 2015c; Chafouleas et al., 2010). Fokussierte und genaue Verhaltensbeurteilungen sind bis zu einer Schülerzahl von sieben Schülern möglich (Chafouleas et al., 2010). Im Allgemeinen ist eine Varianz im Schülerverhalten zu erwarten (ebd.). Unter Berücksichtigung der Generalisierbarkeitstheorie sollte diese Varianz in erster Linie in Interaktionen zwischen den Facetten der Person und unterschiedlicher Beurteilungssettings erwartet werden, wie zum Beispiel dem Tag und der Zeit (ebd.). In der Studie von Briesch et al. (2010) fiel der Varianzanteil, der auf die Interaktion zwischen Schüler und Tag bzw. Beobachtungssetting zurückgeführt werden konnte, bei den DBR-Ratings kleiner aus als bei der direkten systematischen Verhaltensbeobachtung. Die Autoren geben als mögliche Begründung an, dass ein „general impression halo effect“ vorliegen könnte (S. 418). Dabei wird die Einschätzung des Raters durch vorherige Ratings oder durch einen allgemeinen Eindruck von den SchülerInnen beeinflusst (ebd.). Ein Halo-Effekt kann vermieden werden, indem Ratings unmittelbar nach der Beobachtung des Verhaltens durchgeführt und die Anzahl der Ratings verringert werden, sodass die Rater nicht überfordert werden oder eine Erschöpfung erleben (ebd.). In der Studie von Chafouleas, Christ et al. (2007) zeigt sich diesbezüglich keine bedeutsame Varianzaufklärung durch die Interaktion der beiden Facetten Person und Tag für Vorschulkinder. Demgegenüber konnten Chafouleas et al. (2010) eine signifikante Varianz in der Interaktion zwischen Person und Tag (11 bis 13%) bei MittelstufenschülerInnen feststellen. Die Autoren führen dies darauf zurück, dass MittelstufenschülerInnen bereits seit mehreren Schuljahren mit schulischen Anforderungen und Erwartungen im Hinblick auf ihr Verhalten vertraut sind, während VorschülerInnen angemessene Verhaltensweisen im schulischen Setting erst erlenen. Die Ergebnisse werden der Studie von Casale et al. (2017) gestützt, die eine substantielle Varianzaufklärung der Interaktion zwischen Person und Tag sowohl für SI-Skalen (19,8%) als auch für MI-Skalen (22,1%) feststellten. Diese Ergebnisse liefern Hinweise auf die Änderungssensitivität von DBR-Verfahren bei der Beurteilung von Schülerverhalten über die Zeit (Chafouleas et al., 2010). Es kann geschlussfolgert werden, dass die Zuverlässigkeit der DBR-Daten von den Beobachtungssituationen abhängt und dass die Rater die gleichen SchülerInnen zu unterschiedlichen Beobachtungszeitpunkten unterschiedlich beurteilten (Casale et

al., 2017). Casale et al. (2017) regen an, dass Verhaltensbeurteilungen zu standardisierten Messzeitpunkten durchgeführt werden sollten. Würde die Varianzaufklärung in diesem Zuge gering ausfallen, so würden die Daten eine höhere Stabilität aufweisen und mit einer besseren Interpretierbarkeit einhergehen (ebd.). Chafouleas et al., 2010 konnten in ihrer Studie herausstellen, dass Rater dann besonders konsistent Verhalten beurteilen konnten, wenn sie das Verhalten von Schülern mit „extremen" Werten beurteilten. Größere Varianzen wurden für die Beurteilungen von SchülerInnen festgestellt, deren Verhalten variabler auftrat (ebd.).
Die Anzahl der Messzeitpunkte, die für zuverlässige Entscheidungen notwendig sind, unterscheiden sich in Abhängigkeit zum angestrebtem Zweck und zum interessierenden Verhalten (ebd.). So sollten Empfehlungen zu Grenzwerten für Screeningzwecke, konservativer ausfallen, als wenn DBR-Daten als Grundlage für absolute oder kriteriumsorientierte Entscheidungen herangezogen werden (ebd.). Casale et al. (2017) unterscheiden in ihrer Studie zwischen normorientierten und intraindividuellen Entscheidungen. Für erstere sind mittels SIS zuverlässige Ergebnisse nach vier Messungen zu erwarten, für MIS nach fünf Messungen (ebd.). Für intraindividuelle Entscheidungen konnten die Autoren lediglich für MI-Skalen eine Überschreitung des kritischen Werts (nach 13 Messungen) unter Berücksichtigung eines ökonomischen Zeitraums ermitteln (ebd.). Casale et al. (2017) schlussfolgern, dass mittels DBR-Verfahren Verhaltensveränderungen in einem im Vergleich zu traditionellen diagnostischen Verfahren relativ kurzen Zeitraum abgebildet werden können. MI-Skalen sollten dabei SI-Skalen vorgezogen werden (ebd.). Weitere Ergebnisse zur verlaufsdiagnostischen Eignung von MI-Skalen liefern Daniels, Volpe, Briesch & Gadow (2017). Sie konnten herausstellen, dass zuverlässige Entscheidungen unter Berücksichtigung der individuellen Bezugsnorm dann getroffen werden können, wenn eine 3-Item-Skala an acht Messzeitpunkten eingesetzt wird, eine 4-Item-Skala oder 5-Item-Skala an vier Messzeitpunkten und eine 6-Item-Skala an drei Messzeitpunkten (ebd.). In zwei weiteren Studien waren 7 bis 10 Messungen (Chafouleas et al., 2007) bzw. 10 bis 20 Messungen (Chafouleas et al., 2010) notwendig, um reliable Einschätzungen des Zielverhaltens zu erhalten. Christ et al. (2010) konnten in ihrer Studie erheben, dass ca. fünf Messungen von einem Rater über mehrere Beobachtungen oder von mehreren Ratern im Zuge eines Beobachtungssettings notwendig sind, um Entscheidungen von geringer Tragweite treffen zu können. Wohingegen Entscheidungen von großer Tragweite dann getätigt werden können, wenn 15 bis 20 Messungen zusammengefasst werden (ebd.). Huber und Rietz (2015) kommen auf Basis ihrer Überblicksarbeit zu dem Schluss, dass mindestens fünf Messungen durchgeführt werden sollten, um starke Messwertstreuungen zwischen den beurteilenden Personen zu vermeiden, da ansonsten im Verhältnis ein größerer Varianzanteil durch die Rater aufgeklärt wird als durch Unterschiede im Verhalten der beobachteten Personen. Bis zu 20 Datenpunkte sollten zusammengefasst werden, wenn Entscheidungen von hoher Relevanz

mithilfe von DBR-Daten beantwortet werden sollten (ebd.). Christ et al. (2009) empfehlen zusammenfassend für grundlegende Aussagen zum Schülerverhalten eine Anzahl von 10 Ratings, wenn die Ratings von demselben Rater durchgeführt werden. Insgesamt ist dabei die Anzahl der Ratings wichtiger ist als die Anzahl der Beobachtungstage (Chafouleas et al., 2007). Diese Forschungsergebnisse zeigen, dass DBR-Verfahren für verlaufsdiagnostische Zwecke geeignet sind (Casale et al., 2015c). Trotz der hohen Anzahl notweniger Messungen stellt das DBR-Verfahren ein weniger aufwendiges Verfahren dar als die direkte systematische Verhaltensbeobachtung (Casale et al., 2017). Ein kombinierter Einsatz verschiedener Methoden der Verhaltensverlaufsdiagnostik, die jeweils unterschiedliche Zwecke abdecken (z.B. Screening, Verlaufsdiagnostik) und auf unterschiedliche Ressourcen zurückgreifen (z.B. personell, zeitlich), kann als effektiv bezeichnet werden (Daniels et al., 2017).

Bei den Ratingskalen von DBR-Verfahren handelt es sich ähnlich zur Verhaltensbeurteilung um mehrstufige Likert-Skalen, die z.B. numerisch (0 – 10) kodiert werden (Huber & Rietz, 2015). Die eingesetzten Ratingskalen können nominalskalierte, ordinalskalierte oder intervallskalierte Merkmale aufweisen (ebd.). Die Anzahl der Skalengradienten hat nur einen geringen Einfluss auf die Ergebnisse von DBR-Verfahren (Briesch et al., 2010). Die Genauigkeit der Beurteilungen wird dabei nicht durch die Skalenbreite oder die Beschriftung der Skala (z.B. prozentual) beeinflusst (Christ et al., 2010; Riley-Tillman et al., 2011; Chafouleas, Christ & Riley-Tillman, 2009b). DBR-Ergebnisse können folglich über Skalengradienten hinweg verallgemeinert werden. Dies ist für den Aspekt der Kommunikation relevant. Die Anzahl der Skalengradienten kann für jüngere SchülerInnen zur besseren Verständlichkeit angepasst werden ohne die psychometrische Qualität des Verfahrens negativ zu beeinflussen (Briesch, Kilgus, Chafouleas, Riley-Tillman & Christ, 2013). Es werden Skalen mit mindestens sechs Stufen empfohlen (Chafouleas, 2011; Christ et al., 2009). Insbesondere siebenstufige Skalen weisen gute Reliabilitäts- und Validitätswerte auf und werden in der Forschungsliteratur empfohlen (Bühner, 2011; Döring & Bortz, 2016; Preston & Colman, 2000). So konnten Preston & Colman (2000) in ihrer Studie feststellen, dass die Werte der internen Konsistenz für Skalen mit sieben Stufen oder mehr am höchsten ausfielen. Gleichzeitig sinkt die Test-Retest-Reliabilität bei Skalen, die mehr als 10 Stufen umfassen (ebd.).

Es werden Single-Item-Skalen (SIS) und Multi-Item-Skalen (MIS) unterschieden. Verfahren mit lediglich einem einzigen Item werden als SIS bezeichnet, wohingegen Verfahren mit mehreren Items als MIS bezeichnet werden (Casale et al., 2015b). SIS sind dann ökonomisch, wenn sie globale Verhaltensbereiche erfassen (Casale et al., 2017). Huber und Rietz (2015) empfehlen bei einem Einsatz von SIS nach Möglichkeit positiv und global formulierte Items, die die Anwesenheit eines Verhaltensbereichs erheben. Konkrete Verhaltensweisen können mit SIS nur begrenzt erfasst werden (Casale et al., 2017). Konkrete Verhaltensweisen sind dann von Bedeutung, wenn individuelle Entwicklungsverläufe erfasst werden sollen (ebd.). Für

die Abbildung individueller Entwicklungsverläufe eignen sich besonders MIS, da mehrere Items (i.d.R. drei bis fünf) als Indikator für eine übergeordnete Verhaltensdimension herangezogen werden (ebd.). Die einzelnen Items können abhängig vom Ziel des Einsatzes der Ratingskala einzeln oder aufsummiert ausgewertet werden (ebd.).

Casale et al. (2017) kommen auf Basis ihrer Studie zu dem Schluss, dass für intraindividuelle Vergleiche, wie Entwicklungsverläufe, eine höhere Zuverlässigkeit der Beurteilungen mittels (konkreten und spezifischen) MIS erreicht werden kann. Wohingegen für relative Vergleiche zwischen verschiedenen SchülerInnen eine hohe Zuverlässigkeit der Beurteilungen sowohl bei SIS als auch MIS nachgewiesen werden konnte (ebd.). Ähnliche Werte konnten auch in der Studie von Volpe und Briesch (2012) ermittelt werden. Die Autoren stellten für DBR-MIS höhere Reliabilitätswerte fest als für DBR-SIS, sowohl für den Verhaltensbereich der Beteiligung am Unterricht als für störende Verhaltensweisen. Gleichzeitig sind MIS stabiler und stimmen im höheren Maße mit systematischen Verhaltensbeobachtungen und dem „wahren Verhalten“ überein (Huber & Rietz, 2015). SIS können leichter eingesetzt werden als MIS, denn mit einer höheren Anzahl der Items steigt zugleich die „Komplexität bzw. Schwierigkeit der Bewertungsaufgabe“ (ebd., S. 92). Die MIS weisen jedoch eine (scheinbar) höhere testdiagnostische Genauigkeit auf. Sie können spezifische und globale Messungen miteinander verbinden, denn die Rater müssen spezifische Indikatoren eines allgemeineren Verhaltenskonstrukts erfassen (Volpe & Briesch, 2012). Die vorliegenden Forschungsergebnisse weisen darauf hin, dass MI-Skalen besser als SI-Skalen dazu geeignet sind, intraindividuelle Verhaltensveränderungen über die Zeit sensitiv abzubilden (ebd.).

Des Weiteren zeigt sich, dass die Anzahl der Items die Testgüte nur gering beeinflusst (Casale et al., 2015c). Casale et al. (2015c) konnten herausstellen, dass die Generalisierbarkeit und Zuverlässigkeit einer Ratingskala mit geringerer Itemzahl lediglich minimal sinkt. Beurteilungen, die auf einem Item beruhen, weisen lediglich etwas geringere Werte auf als Beurteilungen, die auf fünf Items beruhen (ebd.). Grundsätzlich werden für zeitnahe Entscheidungen DBR-MIS empfohlen, da die Verwendung einer MIS die Anzahl der Ratings verringert, die benötigt werden, um ein angemessenes Maß an Zuverlässigkeit zu erreichen (Volpe & Briesch, 2012). MI-Skalen liefern dabei viele einzelne Messungen (Huber & Rietz, 2015). Im Allgemeinen führen viele Messungen und die Bildung eines Mittelwertes zu stabileren Werten (ebd.). Soll Schülerverhalten verlaufsdiagnostisch und unter Berücksichtigung der individuellen Bezugsnorm erhoben werden, ist eine Anzahl von bis zu fünf Items empfehlenswert (Volpe & Briesch, 2012).Soll Schülerverhalten hingegen statusdiagnostisch erhoben werden erscheinen mitunter bis zu 50 Items sinnvoll (Volpe & Fabiano, 2013).

Zielverhaltensweisen werden mithilfe von Items erfasst, die als Indikatoren übergeordnete Konstrukte abbilden (Casale et al., 2015a). Dementsprechend wird ein reflexives Messmodell

verwendet (ebd.). Die Items können hinsichtlich zwei unterschiedlicher Formen der Formulierung unterschieden werden. Zum einen können sie global und zum anderen spezifisch formuliert sein (ebd.). Beispiele für global formulierte Items sind: „störendes Verhalten“ oder „Lern- und Arbeitsverhalten“ (ebd., S. 47). Spezifisch formulierte Items sind beispielsweise: „Ärgert andere“, „Arbeitet konzentriert“ oder „Meldet sich bei Fragen“ (ebd., S. 48). Sowohl mittels globaler als auch mittels spezifischer Items werden konkrete und beobachtbare Verhaltensweisen und keine abstrakten Konstrukte beurteilt (ebd.). Globale und spezifische Formulierungen stellen Pole entlang eines Kontinuums dar (Christ et al., 2011). Dabei liegt das Dilemma vor, dass im Zuge der Erfassung globaler Verhaltensbereiche wichtige Veränderungen in spezifischen Verhaltensweisen nicht erhoben werden können. Werden demgegenüber allerdings nur wenige spezifische Zielverhaltensweisen erfasst, dann können potenziell umfassendere Effekte der Interventionsmaßnahmen unentdeckt bleiben (Volpe & Briesch, 2012).

Im Rahmen von DBR-Verfahren werden ausschließlich Zielverhaltensweisen beobachtet, die zuvor definiert wurden. Andere Verhaltensweisen werden im Zuge der Beurteilungen nicht berücksichtigt (Huber & Rietz, 2015). Die Konzeption von DBR-Verfahren ermöglicht die Einschätzung unterschiedlichster Verhaltensweisen (Casale et al., 2015a). Das ausgewählte Verhalten kann ein Verhalten sein, welches als problematisch empfunden wird, welches gefördert werden soll oder welches generelle Aussagen ermöglichen soll (Casale, 2017; Casale et al., 2015b; Chafouleas et al, 2012a). Neben Verhaltensweisen, die im Unterricht gezeigt werden, z.B. Verhalten in Gruppenarbeiten, kann auch Verhalten relevant sein, welches in Pausen- oder Übergangszeiten (z.B. auf dem Weg zu anderen Räumlichkeiten) gezeigt wird (Casale, 2017).

Insgesamt kann aufgezeigt werden, dass die Ergebnisse zur psychometrischen Güte in Abhängigkeit zu spezifischen Verhaltensweisen variieren (Chafouleas et al., 2007; Christ et al., 2011; Huber & Rietz, 2015). Chafouleas (2011) hebt im besonderen drei Verhaltensbereiche hervor, die sogenannten „Big 3“ des DBRs: Beteiligung am Unterricht, störendes Verhalten und respektvolles Verhalten. Die Verhaltensbereiche wurden auf Basis von Literaturrecherchen herausgearbeitet und danach ausgewählt, ob sie für viele SchülerInnen von Relevanz sind und ob sie einen Einfluss auf die Lernerfolge der SchülerInnen haben (ebd.). Huber und Rietz (2015) stellen in ihrer Überblicksarbeit heraus, dass eine besonders gute Beobachtungsgüte für den Bereich der Beteiligung am Unterricht und störendes Verhalten nachgewiesen werden konnte. Die Beobachtungsgenauigkeit für respektvolles Verhalten lag in einem niedrigeren aber immer noch akzeptablen Bereich (ebd.). Ferner führen die Autoren auf, dass Forschungsergebnisse zum Verhaltensbereich der internalisierenden Verhaltensweisen fehlen (ebd.). Die Forschungsarbeiten zeigen, dass die Formulierungen bzw. die Valenz der Items die Genauigkeit der Beurteilungen beeinflussen (Casale et al., 2017; Chafouleas, Jaffery, Riley-Tillman, Christ & Sen, 2013; Christ et al., 2011; Riley-Tillman, Chafouleas, Christ,

Briesch & LeBel, 2009). Die in diesen Studien erhobenen Ergebnisse sind nicht einheitlich. Sie deuten darauf hin, dass die Items für den Verhaltensbereich der Beteiligung am Unterricht positiv formuliert und die Items für die Bereiche des störenden und respektvollen Verhaltens negativ formuliert sein sollten (Chafouleas et al., 2013; Christ et al., 2011; Riley-Tillman et al.,2009). Unabhängig von diesen spezifischen Forschungsergebnissen ist es wichtig, dass das Verhalten beobachtbar, also sorgfältig operationalisiert und situativ relevant ist (Casale, 2017). Eine entsprechende Operationalisierung kann zu einem Anstieg der Beobachtungsgüte führen (Huber & Rietz, 2015). Sehr auffällige Verhaltensweisen können leichter erfasst werden als weniger auffällige Verhaltensweisen (Huber & Rietz, 2015; Landscheidt, 2001). Die Verhaltensweisen sollten unabhängig voneinander betrachtet werden, da die Art, die Formulierungen, die Valenz und die Spezifität der Zielverhaltensweisen zu unterschiedlichen Ergebnissen in der Verhaltensbeurteilung führen können (Chafouleas, 2011; Chafouleas et al., 2010; Riley-Tillman et al., 2009). Vergleiche über verschiedene Verhaltensbereichen erscheinen somit nicht angemessen (Chafouleas et al.., 2010).

Zudem gilt es zu beachten, dass die Inferenz der Items eine Auswirkung auf die Zuverlässigkeit der DBR-Messungen hat (Casale et al., 2017). Items mit einer hohen Interferenz, wie sie beispielsweise eher bei globalen Items gegeben ist, ermöglichen im Allgemeinen weniger zuverlässige Beurteilungen als Items mit einer niedrigeren bzw. mittleren Interferenz (ebd.). Damit Items dem Gütekriterium der Inferenz im Sinne der Verlaufsdiagnostik entsprechen, sollten Items tendenziell eher niedrig bzw. mittel inferent und konkret operationalisiert sein (Casale et al., 2015a)(Christ et al., 2010; Conley, Marchant & Caldarella, 2014).

Die Items können zum einen über nomothetische und zum anderen über idiografische Ansätze generiert werden. Nomothetische Ansätze nutzen große Stichproben, um Skalen auszuwählen, die sich psychometrisch als am besten geeignet erwiesen haben (Volpe & Gadow, 2010). Vorteile des nomothetischen Ansatzes sind, dass die Skalen eine hohe Testgüte aufweisen und die Testgüte bereits entwickelter Items leicht ausgewertet werden kann (ebd.). Mitunter können sie jedoch für den Einzelfall nicht geeignet sein (Casale et al., 2015a). Idiografische Ansätze erfordern demgegenüber spezifische Kenntnisse über die SchülerInnen, z.B. aus vorherigen Verhaltensbeurteilungen (Volpe & Gadow, 2010). Diese Ansätze gehen von den einzelnen SchülerInnen aus (Casale et al., 2015a). Für jeden Schüler und jede Schülerin werden spezifische Items und jeweils ein eigenes DBR-Instrument entwickelt (ebd.). Dementsprechend liegt bei diesem Ansatz eine sehr gute Passung zwischen Item und Kind vor (ebd.). Jedoch kann aufgrund der geringen Stichprobengröße die Testgüte nur erschwert evaluiert werden (ebd.). Chafouleas et al. (2010) empfehlen eine verlaufsdiagnostische Methode, die idiographische und nomothetische Ansätze verknüpft. Insgesamt können Direct Behavior Ratings flexibel in Abhängigkeit zum jeweiligen Zweck gestaltet werden (Christ et al., 2009).

Zur Darstellung der erhobenen Daten kann ein anschauliches Liniendiagramm gewählt werden (Casale, 2017). Dieses kann die Entwicklung des Kindes grafisch bzw. als „Kurve“ darstellen, sodass diese für Rater, SchülerInnen, Eltern und weitere am Förderprozess beteiligte Personen erkennbar ist (Huber & Rietz, 2015, S. 93; Grosche, 2014). Auf Basis dieser Darstellung ist eine Evaluation der Wirksamkeit eingesetzter pädagogischer Maßnahmen möglich (Huber & Rietz, 2015).

Casale et al. (2015a) weisen auf Einschränkungen des Instrumentes als auch auf Einschränkungen im Hinblick auf die Erfassung von Verhalten hin. Zum einen bedarf es weiterhin einer Klärung, was unter Schülerverhalten verstanden wird (ebd.). Obwohl bestimmte Verhaltensdimensionen, wie Sozialverhalten oder Lern- und Arbeitsverhalten, wiederholt in der Literatur aufgegriffen werden, können sich die zugrundeliegenden Konzepte unterscheiden. Die Begrifflichkeiten sollten unter Berücksichtigung des Forschungs- und Anwendungsbereiches betrachtet werden (Casale et al., 2015a; Jurkowski & Hänze 2014). Zum anderen fehlen sowohl theoretische als auch empirische Forschungsarbeiten zur „Schwierigkeit der diagnostischen Situation“ (Huber & Rietz, 2015, S. 94). Dieser Aspekt ist von Relevanz, da die Auftretenswahrscheinlichkeit eines Verhaltens in Abhängigkeit zu verschiedenen inhaltlichen, methodischen und didaktischen Settings steht (ebd.). Dabei ist unklar, wie Verhaltensverlaufsentwicklungen in Anbetracht unterschiedlicher Situationsschwierigkeiten interpretiert werden können (ebd.). Die Verhaltensentwicklungen können ohne systematische Kontrollen nur unzureichend interpretiert werden, da Verhaltensentwicklungen ebenfalls auf die Situation oder ein ungenügendes Testverfahren zurückgeführt werden könnten (Casale et al., 2015a).

Im Allgemeinen sollten Befunde zur Testgüte nicht unüberlegt auf andere Kontexte, wie z.B. andere Rater, andere Verhaltensbereiche oder andere Schülergruppen übertragen werden (Chafouleas et al., 2010). Bereits vorliegende Forschungsarbeiten deuten darauf hin, dass DBR-Verfahren Zielverhaltensweisen zuverlässig erfassen und als Grundlage für intraindividuelle und normorientierte Interpretationen herangezogen werden können. Dazu müssen bestimmte Voraussetzungen getroffen werden, wie z.B. klar definierte Beobachtungszeiträume und eindeutige Operationalisierungen von Items (Casale et al., 2015c). Huber und Rietz (2015) empfehlen weitere Studien unter systematischer Variation der situativen Bedingungen und auch der Itemformulierungen. Des Weiteren fehlen Studien, in deren Fokus das Verhalten von SchülerInnen mit problematischen Verhaltensweisen steht (Casale et al., 2017). Dabei stellt diese Gruppe eine wesentliche Zielgruppe der Verlaufsdiagnostik im RTI-Modell dar (ebd.). Im deutschen Sprachraum sind weitere Studien zur Testgüte des DBRs notwendig (Huber & Rietz, 2015).

6. Pilotierungsstudie

Zahlreiche Forschungsergebnisse zur Methode des DBR stammen aus dem englischen Sprachraum. Diese Forschungsergebnisse können nur bedingt auf den deutschen Sprachraum übertragen werden (Huber & Rietz, 2015). Folglich sollten auch im deutschen Sprachraum Ratingskalen zur direkten Verhaltensbeurteilung entwickelt und hinsichtlich ihrer psychometrischen Testgüte evaluiert werden (Casale et al., 2015c). Dazu will die vorliegende Pilotierungsstudie einen Beitrag leisten.

6.1. Zielsetzung und Fragestellung

Im Rahmen dieser wird der Einsatz einer Direct Behavior Ratingskala zu vier Verhaltensbereichen (schulbezogenes, externalisierendes, internalisierendes und prosoziales Verhalten) erprobt. Für die Onlineplattform LEVUMI wurde auf Basis einer adaptierten Version des SDQ eine Multi-Item-Skala entwickelt, welche zur Verhaltensverlaufsdiagnostik eingesetzt werden soll. Die Ratingskala wurde im Rahmen einer Masterarbeit von Sauerland (i.D.) qualitativ weiterentwickelt. Da sich der aktuelle Forschungsstand überwiegend auf Single-Item-Skalen bezieht (Casale et al., 2017), soll diese Studie einen Beitrag zur Erforschung von MIS liefern. Grundsätzlich sollte jede neu konstruierte Ratingskala hinsichtlich der für die Verlaufsdiagnostik relevanten Gütekriterien überprüft werden, da Ratingskalen sehr flexibel eingesetzt und gestaltet werden können (Chafouleas et al., 2010). Die neu konstruierte und weiterentwickelte Ratingskala wird im Zuge der vorliegenden Pilotierungsstudie im inklusiven Setting der Grundschule und Gesamtschule erprobt. Die Pilotierungsstudie verfolgt zwei Forschungsziele. Zum einen soll sie Hinweise zur Notwendigkeit einer Weiterentwicklung der vorliegenden Ratingskala liefern. Zum anderen sollen anhand Studie die Verhaltensveränderungen der Grund- und GesamtschülerInnen untersucht werden. Vorab werden zunächst die Messbedingungen erhoben, unter denen die Ratingskala eingesetzt wurde. Zudem wird erhoben, ob die Skala die Gütekriterien der Ökonomie und Nützlichkeit erfüllt. Im Anschluss daran wird die psychometrische Güte der Ratingskala überprüft, indem unter Berücksichtigung einer 4-faktoriellen Struktur eine Schätzung der Reliabilitätswerte (interne Konsistenz) und eine Überprüfung der Skalierung sowie der Änderungssensibilität erfolgt. Darauf aufbauend wird untersucht, welche Mittelwerte die SchülerInnen der beiden Schulformen aufweisen, welche Unterschiede sowohl zwischen den Schülergruppen als auch über die Zeit vorliegen. Zum Abschluss wird die Variationsbreite der Verhaltensbeurteilungen durch die Lehrkräfte untersucht.

Zur Beantwortung dieser Forschungsinteressen wurden folgende Forschungsfragen formuliert:

Forschungsfrage 1: Wie setzen Lehrkräfte an Grund- und Gesamtschulen die Ratingskala ein? Handelt es sich bei der Ratingskala um ein ökonomisches und nützliches Testinstrument?

Forschungsfrage 2: Wie ist die interne Konsistenz der Skalen des 4-Faktormodells pro Messzeitpunkt und über die Messzeitpunkte?

Forschungsfrage 3: Wie korrelieren die Werte der fünf Messzeitpunkte der vier Skalen unter Berücksichtigung der Schulform miteinander?

Forschungsfrage 4: Welche Mittelwerte weisen die Grund- und GesamtschülerInnen pro Messzeitpunkt und über die Messzeitpunkte auf?

Forschungsfrage 5: Welche Variationsbreite weisen die Beurteilungen der Lehrkräfte der Grund- und Gesamtschulen auf?

6.2. Onlineplattform LEVUMI

Die Onlineplattform LEVUMI wurde im Rahmen des gleichnamigen Forschungsprojektes der Technischen Universität Dortmund (Prof. Dr. Markus Gebhardt), der Europa-Universität Flensburg (Prof. Dr. Kirsten Diehl) und der Universität Kiel (Prof. Dr. Andreas Mühling) entwickelt. Im Fokus dieser Onlineplattform stehen vertiefende Forschungen zur Lernverlaufsdiagnostik und die Entwicklung eines praktikablen und schulrelevanten Onlineinstrumentes (Gebhardt, Diehl & Mühling, 2016; Jungjohann, DeVries, Gebhardt & Mühling, 2018). Derzeit (Stand Juli 2018) sind auf der Plattform (https://www.levumi.de/tests) Tests für die Lernbereiche Deutsch (Leseflüssigkeit, Rechtschreibung, Sinnentnehmendes Lesen, Wortschatz) und Mathematik (Zahlen lesen, Zahlenreihen, Zahlenstrahl) verfügbar (Gebhardt & Mühling, n.d.). Die Tests sind in bis zu sieben verschiedenen Niveaustufen verfügbar (Gebhardt & Mühling, n.d.). Für den Bereich der Leseflüssigkeit sind zudem konkrete Fördermaßnahmen (Vorläuferfähigkeiten, Buchstaben-Lautbeziehung, Richtiges Lesen, Wörter kennen und erkennen, Inhalte verstehen, Miteinander lesen) entwickelt worden, die ebenfalls auf der Internetseite der Lernplattform (https://www.levumi.de/materials) kostenfrei verfügbar sind (Jungjohann, Gebhardt, Diehl & Mühling, 2017). Die Onlineplattform bietet eine computergestützte Auswertung und Darstellung der Testverfahren an (Gebhardt et al., 2016). Ein Vorteil dieses Auswertungsverfahrens ist, dass alle notwendigen Aufgaben und Analysen automatisch computergestützt erfolgen (Gebhardt et al., 2016). Lehrkräfte können dabei sowohl ein quantitatives als auch ein qualitatives Feedback erhalten (Mühling et al., 2017; Jungjohann & Gebhardt, 2018). Neben den aktuell verfügbaren und umfangreich evaluierten Tests zur Lernverlaufsdiagnostik sollen langfristig weitere Tests entwickelt werden (Gebhardt et al., 2016). Weitere Tests können der Onlineplattform problemlos hinzugefügt werden (Gebhardt et al., 2015a). Zukünftig soll die Onlineplattform um den Themenbereich des Verhaltens erweitert werden. Dazu wird in der vorliegenden Pilotierungsstudie eine Direct Behavior Ratingskala erprobt.

6.3. Methodisches Vorgehen

Nachfolgend soll erläutert werden, welches methodische Vorgehen zur Beantwortung der vorgestellten Forschungsfragen gewählt wurde. Zunächst wird zur Einordnung der Pilotierungs-

studie das übergeordnete Forschungsprojekt LEVUMI beschrieben. In dem Kapitel 6.3.1 werden die Methoden, die zur Erhebung und Auswertung der Daten herangezogen wurden, vorgestellt. Darauf folgt eine Beschreibung des Aufbaus und der Operationalisierung der Ratingskala. In dem Kapitel 6.3.4 wird erläutert, wie die Pilotierungsstudie durchgeführt wurde. Das letzte Kapitel (6.3.5) dient der Stichprobendarstellung.

6.3.1. Studiendesign

Erhebungsmethoden

Die Daten werden zum einen quantitativ mit der zu evaluierenden Ratingskala und zum anderen qualitativ durch vier halbstrukturierte Interviews erhoben. Es liegt somit ein Mix-Method Ansatz vor (Döring & Bortz, 2016). Ratingskalen sind in der Forschungspraxis von großer Bedeutung (ebd.). Sie können zur Fremd- und Selbstbeurteilung psychologischer und sozialer Variablen eingesetzt werden (ebd.). Dabei beurteilen die Rater nach subjektiven Einschätzungen die quantitative Ausprägung eines Merkmals oder einer Eigenschaft einer Person, in diesem Fall das Verhalten der SchülerInnen (Bühner, 2011; Döring & Bortz, 2016). Die Stufen einer Ratingskala entsprechen „markierten Abschnitten eines Merkmalskontinuums“ (Döring & Bortz, 2016, S. 245), die i.d.R. intervallskaliert beurteilt werden. Somit wird der Abstand zwischen den Stufen als gleich groß interpretiert (ebd.). Im wissenschaftlichen Diskurs besteht Uneinigkeit darüber, ob Ratingskalen als intervallskaliert aufgefasst und inferenzstatistische Analysen basierend auf Daten durchgeführt werden können, die durch Ratingskalen gewonnen wurden (ebd.). Die vorliegende Studie orientiert sich an der Argumentationslinie, dass „die Verletzungen der Intervallskaleneigenschaften [...] bei Ratingskalen nicht so gravierend [sind], als dass man auf die Verwendung parametrischer interferenzstatistischer Verfahren gänzlich verzichten müsste“ (ebd., S. 250). Grundsätzlich ist jedoch zu beachten, dass intervallskalierte Messungen mit Ratingskalen „ein auf Hypothesen gegründetes Unterfangen“ darstellen (ebd., S. 250). Vorteile von Ratingskalen sind, dass sie differenzierte Aussagen über Merkmalsausprägungen ermöglichen und ökonomisch eingesetzt und ausgewertet werden können (Bühner, 2011; Casale et al., 2017). Diese Vorteile ermöglichen es eine umfassende Datengrundlage für die vorliegende Studie zu generieren. Zur Auswertung dieser Daten wurden die Programme IBM SPSS 25 und Microsoft Office Excel 2016 verwendet.

Interviews bieten den Vorteil, dass „Aspekte des subjektiven Erlebens“, wie z.B. Gefühle oder Einstellungen erfasst und zurückliegende Ereignisse kommuniziert werden können (Döring & Bortz, 2016, S. 365) Die Interviews liefern umfängliche und komplexe Schilderungen des Einsatzes der Methode (ebd.). Bei den durchgeführten Interviews handelt es sich um halbstrukturierte Interviews, da zur Strukturierung des Interviews ein Interview-Leitfaden (vgl. Anhang 3.5) herangezogen wurde (Döring & Bortz, 2016; Walter-Klose, 2015). Dem Interview-Leitfaden liegen offene („Wie war der Einsatz der Ratingskala?“), sowie geschlossene („Kann die

Ratingskala im Schulalltag häufig und schnell eingesetzt werden?“) Fragen zugrunde. Die Reihenfolge der Fragen und die Formulierungen wurden je nach Gesprächsverlauf flexibel verändert und angepasst, um einen Gesprächsfluss aufrechtzuerhalten und um den Lehrkräften die Möglichkeit zu geben, eigene Anliegen und Anregungen vorzubringen (Döring & Bortz, 2016; Meuser & Nagel, 2010; Walter-Klose, 2015) . Bei den Interviews handelt es sich um Experteninterviews, da die interviewten Lehrkräfte als Fachpersonen zur Beurteilung von Schülerverhalten im schulischen Setting angesehen werden (Walter-Klose, 2015). Sie wurden in drei von vier Fällen telefonisch durchgeführt. Vorteile des telefonischen Leitfaden-Interviews sind einerseits die ökonomische Durchführung und andererseits die größere Anonymität, welche es den interviewten Personen erleichtert spezifische Themen anzusprechen (Döring & Bortz, 2016a). Da zur Beantwortung der Forschungsfragen die inhaltlichen Aussagen der befragten Personen im Vordergrund standen, wurden die Interviews basierend auf dem einfachen Transkriptionssystem nach Dresing und Pehl (2017) ausgewertet (vgl. Anhang 3.6). Im Sinne einer Teiltranskription wurden nur solche Interviewsequenzen transkribiert, die für die Beantwortung der Forschungsfragen von Relevanz waren (Döring & Bortz, 2016a). Ein Interview wurde auf Wunsch nicht auditiv dokumentiert. Die inhaltlich relevanten Aussagen wurden teilweise wörtlich, überwiegend stichpunktartig im Verlauf des Gesprächs notiert. Eine qualitative Auswertung der Transkripte findet sich im Anhang (vgl. Anhang 3.7 – 3.10). Diese orientiert sich an der qualitativen Inhaltsanalyse nach Mayring (2010). Die Interviews wurden mithilfe der App „Sprachmemos“ dokumentiert und mithilfe des Programms „f5“ transkribiert.

Auswertungsmethoden

Eine Evaluation der Testgüte von DBR Ratingskalen ist von besonderer Bedeutung, da Beurteilungen des Schülerverhaltens auf Fremdeinschätzungen sowie Schlussfolgerungen der Rater basieren, welche mit Urteilsfehlern einhergehen (Casale et al., 2015c; Schmidt-Atzert et al., 2012). Eine Evaluation möglicher Fehlerquellen ist nur bedingt auf Basis gängiger forschungstheoretischer Ansätze möglich (Casale et al., 2015c). Chafouleas et al. (2012) und Voß und Gebhardt (2017a) verweisen zur Evaluation der Testgüte verlaufsdiagnostischer Instrumente auf den Ansatz von Fuchs (2004), da dieser Ansatz auch für Forschungen zur Methode des Direct Behavior Ratings einen Orientierungsrahmen bieten kann (Chafouleas et al, 2012a). Fuchs (2004) gibt an, dass zum Nachweis der Vertretbarkeit eines Instrumentes zur Verlaufsdiagnostik Untersuchungen auf drei Forschungsstufen notwendig sind (Fuchs, 2004). Diese Stufen werden von Voß und Gebhardt (2017a) aufgegriffen. Sie formulieren drei mit den Stufen übereinstimmende „Forderungen an verlaufsdiagnostische Verfahren“ (S. 21). Erstens sollten verlaufsdiagnostische Instrumente die psychometrischen Gütekriterien der Statusdiagnostik erfüllen (ebd.). Dazu sind Schätzungen der Hauptgütekriterien Objektivität, Reliabilität und Validität durchzuführen (ebd.). Zweitens sollten die Instrumente hinsichtlich psychometri-

scher Eigenschaften der Verlaufsdiagnostik untersucht werden (ebd.). Es sollte überprüft werden, ob das Instrument Entwicklungsverläufe änderungssensibel abbilden kann (ebd.). Als dritte Forderung führen sie auf, dass ein ökonomischer Einsatz des Instrumentes möglich sein und das Instrument einen positiven Einfluss auf schulische Unterrichtsprozesse haben sollte (ebd.). Die ersten beiden Stufen können Hinweise zur Angemessenheit und die dritte Stufe Hinweise zur Nützlichkeit des Instrumentes für den Unterricht liefern (Fuchs, 2004). Die vorliegende Studie orientiert sich zur Evaluation der Testgüte der neu entwickelten Ratingskala an diesen Stufen bzw. Forderungen.

Bevor Untersuchungen zur Testgüte vorgenommen werden ist zu erheben, unter welchen Messbedingungen die Ratingskala eingesetzt wurde. Es gilt zu überprüfen, unter welchen Messbedingungen die Methode des DBRs Verhaltensweisen von SchülerInnen zuverlässig erfassen kann (Casale et al., 2017). Dies ist für einen verlaufsdiagnostischen Einsatz der Methode im praktischen Handlungsfeld Schule von großer Bedeutung (ebd.). Da die Studie im Sinne einer Feldstudie im schulischen Setting durchgeführt wurde, stimmen die Untersuchungsbedingungen weitestgehend mit den Bedingungen der Situationen überein, in denen die Ratingskala zukünftig eingesetzt werden soll (Döring & Bortz, 2016). Zur Beantwortung der ersten Forschungsfrage wurde dementsprechend untersucht, wie und unter welchen Bedingungen die Lehrkräfte die Methode im inklusiven Handlungsfeld der Grund- und Gesamtschulen eingesetzt haben. Die Lehrkräfte wurden nach der Beurteilung des Verhaltens gebeten, Angaben zum Beobachtungssetting zu machen. Auf Grundlage dieser Daten wurden Häufigkeitsanalysen vorgenommen. Die Lehrkräfte wurden zudem interviewt, um die Erfahrungen und Einschätzungen der Lehrkräfte, die als Experten im schulischen Setting angesehen werden, zu erheben. Ein zentrales Anliegen dieser Studie war es, das „Spezialwissen (strukturelles Fachwissen und/oder Praxis-/Handlungswissen)“ (ebd., S. 376) der Lehrkräfte bei der Weiterentwicklung der Methode zu berücksichtigen. Diese „Fachexpertise“ (ebd., S. 360) kann als Eigenschaft der Lehrer Einfluss auf die Genauigkeit der Urteile im schulischen Setting haben (Casale et al., 2015a; Südkamp, Kaiser & Möller, 2012). Südkamp et al. (2012) führen dies in ihrem Modell der teacher judgement accuracy auf. Da Lehrerurteile im Allgemeinen nicht fehlerfrei bleiben, sollten bekannte Quellen dieser Fehler bei der Entwicklung von Testinstrumenten berücksichtigt werden (Casale et al., 2015a; Südkamp et al., 2012). Dementsprechend findet in dieser Studie die Variable der Schulform besondere Berücksichtigung und damit einhergehend auch das Alter der SchülerInnen. Dieses führen Südkamp, Kaiser und Möller (2012) ebenfalls als einen Einflussfaktor auf die Genauigkeit von Lehrerurteilen an.

Gleichzeitig kann mittels der Interviews auch die dritte Forderung nach Voß und Gebhardt (2017a) überprüft werden, indem erhoben wird, ob die Ratingskala aus Sicht der Lehrkräfte die Gütekriterien der Ökonomie und Nützlichkeit erfüllt. Testverfahren können dann als ökonomisch bezeichnet werden, wenn sie bei geringem Materialaufwand, einfacher Handhabung

in einer kurzen und angemessenen Zeitspanne (für Gruppen) durchgeführt sowie schnell und mit wenig Aufwand ausgewertet werden können (Bundschuh & Winkler, 2014; Bühner, 2011). Für verlaufsdiagnostische Instrumente ist das Kriterium der Ökonomie von immenser Bedeutung, da dieses Kriterium eine schnelle Durchführung und einen häufigen sowie kontinuierlichen Einsatz gewährleistet kann (Voß & Gebhardt, 2017a). Diesbezüglich wurde den Lehrkräften die Frage gestellt, ob die Ratingskala im Unterricht schnell und häufig eingesetzt werden kann. Im schulischen Setting sind kurze und schnelle Instrumente notwendig, die „wenig personelle Ressourcen erfordern“ (Casale et al., 2015a, S. 40). Das Kriterium der Nützlichkeit ist insbesondere im sonderpädagogischen Kontext von großer Bedeutung. Es erfasst zum einen, ob die gemessenen Verhaltensbereiche in der Praxis von Relevanz sind und zum anderen, ob das Testverfahren einen Beitrag zur Förderung der SchülerInnen leisten kann (Bundschuh & Winkler, 2014) (Bühner, 2011). Casale et al. (2015a) greifen den ersten Aspekt unter dem Kriterium der sozialen Validität auf. Es sollten Verhaltensweisen erfasst werden, die für das Verhalten in der Schule von praktischer Relevanz sind, da Entscheidungen über Fördermaßnahmen anderenfalls auf pädagogisch fragwürdigen Begründungen beruhen würden (ebd.).

Zur Beantwortung der zweiten Forschungsfrage wird entsprechend der ersten Forderung nach Voß und Gebhardt (2017a) die Reliabilität des Verfahrens untersucht. Das Gütekriterium der Reliabilität gibt Auskunft über die „Zuverlässigkeit bzw. Genauigkeit einer Messung“ (Sikora, 2015a, S. 78). Die Messgenauigkeit eines Tests zu einem bestimmten Messzeitpunkt kann beispielsweise anhand der Korrelationen der Items eines Testverfahrens untereinander bei Berücksichtigung der Testlänge bestimmt werden (Bühner, 2011). Dieser Wert wird als innere bzw. interne Konsistenz bezeichnet und ermöglicht Aussagen darüber, ob die 19 Items der Ratingskala Indikatoren der vier übergeordneten Merkmale bzw. Skalen darstellen (Schmidt-Atzert et al., 2012). Eine Operationalisierung der faktoriellen Struktur der Ratingskala findet sich im nachfolgenden Kapitel 6.3.2 (Operationalisierung). Die Bestimmung der internen Konsistenz der vier Skalen erfolgt nach Cronbach’s Alpha. Dabei geben Werte von $\alpha \geq .70$ eine zufriedenstellende Reliabilität, Werte von $\alpha \geq .80$ eine gute Reliabilität und Werte von $\alpha < .50$ eine nicht akzeptable Reliabilität an (Dorsch, Wirtz & Strohmer, 2017). Nachteilig ist bei dieser Analyse, die im Sinne der Klassischen Testtheorie (KTT) erfolgt, dass nur der Faktor Item fokussiert wird und weitere Faktoren, die den Messfehler zusätzlich beeinflussen können, keine Berücksichtigung finden (Casale et al., 2017). Demgegenüber kann mittels Generalisierbarkeitstheorie (GT) der Einfluss mehrerer Faktoren, wie zum Beispiel der Items und Rater auf den Messfehler zugleich erhoben werden (ebd.). Die GT stellt somit eine Erweiterung der KTT dar (ebd.). Vor dem Hintergrund des Umfangs bzw. Anspruchs dieser Arbeit können Forschungstätigkeiten im Sinne der GT in dieser Arbeit nicht gewährleistet werden. Forschungstä-

tigkeiten im Sinne der KT erscheinen angemessen. Die Schätzung von Cronbach's Alpha ermöglicht eine Gesamtschätzung der Zuverlässigkeit der Daten (Hintze, 2005). Des Weiteren kann in diesem Zuge die Trennschärfe der Items festgestellt werden. Zur Bestimmung der Trennschärfe wird die Korrelation zwischen den Messwerten eines Items und den aufsummierten Skalenwerten der übrigen Items berechnet (Bühner, 2011). Mittels des Kennwertes der Trennschärfe, können Aussagen darüber getätigt werden, in welchem Maße ein Item Messungen des angestrebten Konstrukts ermöglicht (ebd.). Ein Item wird dann als trennscharf bezeichnet, wenn für Personen mit einem hohen bzw. niedrigen Gesamtergebnis der Skala auf dem Item ebenfalls ein hoher bzw. niedriger Wert festgestellt werden kann (Döring & Bortz, 2016). Anhand eines trennscharfen Items können Personen mit entsprechenden Ausprägungen des betreffenden Konstruktes identifiziert werden (ebd.). Dabei werden Werte von ri >.50 als hoch, Werte von ri = .30 - .50 als mittel und Werte von ri < .30 als niedrig angegeben (Bühner, 2011). Wenn für Items einer Skala gleiche Trennschärfen ermittelt werden können, schafft dies günstige Voraussetzungen dafür, dass das Gütekriterium der Skalierung erfüllt ist (ebd.). Das Gütekriterium der Skalierung überprüft, ob die Verrechnungsvorschrift, also die Bildung eines Testwertes (z.B. das Aufsummieren der Rohwerte) und die daraus folgenden numerischen Unterschiede zwischen den Testpersonen die Verschiedenheiten im Verhalten abbilden und somit zur Interpretation der Merkmalsausprägungen herangezogen werden können (Bühner, 2011; Casale et al., 2015a). Kann dieses Gütekriterium nicht nachgewiesen werden, so wären die mittels Verrechnungsvorschrift gebildeten Testwerte inhaltlich ohne Wert und nicht interpretierbar (Casale et al., 2015a). Das Gütekriterium der Skalierung ist im Kontext der Verhaltensverlaufsmessung von Bedeutung, da auf dessen Basis subjektiv über- oder unterschätzte Beurteilungen begrenzt werden können (ebd.).

Auf Basis von Reliabilitätsschätzungen können Aussagen darüber getroffen werden, ob sich das Verfahren für statusdiagnostische Zwecke eignet und Verhalten konsistent erfasst werden kann (Hintze, 2005) (Voß, 2014). Unklar ist dann weiterhin, ob es sich für verlaufsdiagnostische Zwecke eignet (Voß, 2014). Zur Überprüfung der verlaufsdiagnostischen Eignung des Verfahrens (Forderung 2) wird die Änderungssensibilität der Ratingskala untersucht. Mittels änderungssensitiver Testverfahren können Veränderungen von Merkmalen abgebildet werden (Casale et al., 2015a). Diese sollten lediglich mäßige bis mittelhohe Retest-Reliabilitätswerte aufweisen und die Korrelationen zwischen den Messungen sollten mit zunehmender Zeit geringer werden (Klauer, 2011). In Anlehnung an Klauer (2011) werden im Zuge der zweiten Forschungsfrage die Korrelationen der Skalenwerte zwischen dem ersten Messzeitpunkt und den weiteren vier Messzeitpunkten berechnet und in einem Liniendiagramm mit Trendlinie visualisiert. Grafische Darstellungen eignen sich im Besonderen zur Visualisierung von Verläu-

fen in Daten (Kazdin, 2005). Es wird die Produkt-Moment-Korrelation bestimmt, die zur Untersuchung von intervallskalierten Daten geeignet ist (Rasch, Friese, Hofmann & Naumann, 2014a; Stockheim, 2015). Dabei wird angenommen, dass Ratingskalen intervallskaliert sind.
Zur Beantwortung der vierten Forschungsfrage wurde untersucht, inwieweit die Ausprägungen der vier Verhaltensbereiche bzw. deren Veränderungen über die Zeit unter Berücksichtigung der zuvor erhobenen Messbedingungen variieren. Dazu werden die Kennwerte der Verhaltensbeurteilungen, die mittels Ratingskala erhoben wurden, systematisch dargestellt. Es wird zum einen das arithmetische Mittel und die Standardabweichung pro Skala und Messzeitpunkt ermittelt. Das Arithmetische Mittel bzw. der Mittelwert beschreibt den Durchschnitt aller gemessenen Werte (Rasch et al., 2014a). Es wird berechnet, indem die Summe aller Werte durch deren Anzahl n dividiert wird (ebd.). Die Standardabweichung gibt das Ausmaß der Abweichungen bzw. Streuungen der Werte einer Verteilung vom arithmetischen Mittel wieder (Rasch et al., 2014a; Wilbert, 2015). Zum anderen werden zur Visualisierung der Ergebnisse pro Skala und Schulform grafische Darstellungen herangezogen (Kazdin, 2005). Es werden Liniendiagramme mit Trendlinien erstellt. Sie dienen ausschließlich der Visualisierung der Verläufe. Liniendiagramme eignen sich ins besondere zur Visualisierung von Verläufen über die Zeit, da sie die Darstellung umfangreicher quantitativer Informationen auf einfache visuelle Weise ermöglichen (Johnson, Riley-Tillman & Chafouleas, 2016). Eine Trendlinie entspricht dabei der Linien, die am besten der Verlaufskurve der erhobenen Daten entspricht (ebd.). Basierend auf dem arithmetischen Mittel und der Standardabweichung wird untersucht, ob Unterschiede zwischen den Schulformen und über die Messzeitpunkte vorliegen. Um zu überprüfen, ob die Unterschiede nicht nur zufällig vorliegen, sondern mit hoher Wahrscheinlichkeit bestehen (Orthmann Bless, 2015) wird ein inferenzstatistisches Verfahren herangezogen: die Varianzanalyse (ANOVA) mit Messwiederholung. Die ANOVA mit Messwiederholung stellt ein bedeutsames Verfahren für die sonderpädagogische Forschung dar und wird bei abhängigen Stichproben herangezogen (Rasch, Friese, Naumann & Hofmann, 2014b; Sinner & Kuhl, 2015b). Sie ermöglicht Aussagen darüber, ob es sich bei den Mittelwertunterschieden der beiden Schülergruppen über die Messezeitpunkte um systematische Unterschiede oder zufällige Abweichungen handelt (Sinner & Kuhl, 2015a). Mittelwertunterschiede werden bei einem Signifikanzniveau kleiner als 5 % ($p < .005$) als systematisch bzw. signifikant aufgefasst (ebd.). Der Test der Innersubjekteffekte ermöglicht Aussagen darüber, ob sich die beiden Schülergruppen über die Messzeitpunkte unterschiedlich entwickelt haben (Sinner & Kuhl, 2015b).
Für den Umgang mit fehlenden Daten wurde das „traditionelle“ Verfahren des Fallweisen Löschens gewählt (Sedlmeier & Renkewitz, 2013, S. 783). Bei diesem Verfahren werden ganze Fälle gelöscht, wenn für eine oder mehr Variablen Werte fehlen (ebd.). Dabei ist zu beachten, dass Ausschlussverfahren mitunter zu unbemerkten Verzerrungen in den Ergebnissen führen können (Döring & Bortz, 2016). Dieses Verfahren wird herangezogen, da für die betrachteten

Werte die MCAR-Bedingung nicht zutrifft. Im Sinne der MCAR-Bedingung fehlen Daten „komplett zufällig“ und die Daten können „als Zufallsstichprobe aus der ursprünglichen vollständigen Stichprobe betrachtet werden“ (Sedlmeier & Renkewitz, 2013, S. 781). Diese Bedingung trifft bei Vergleichen abhängiger Stichproben über mehrere Messzeitpunkte nicht zu. Ferner führen Huber und Rietz (2015) an, dass fünf Messwerte notwendig sind, um die Unterschiede im Verhalten auf die beobachtete Person zurückzuführen. Bei einer geringeren Anzahl an Messwerten würde die Varianz im größeren Umfang durch die Person des Raters aufgeklärt werden als durch die Person des Schülers.

Die Qualität eines Testverfahrens wird zudem von designspezifischen Aspekten bedingt, wie z.B. dem Skalendesign (Casale et al., 2015a). Abschließend wird diesbezüglich erhoben, in welchem Bereich die Verhaltensbeurteilungen der Grundschullehrkräfte im Vergleich zu den Verhaltensbeurteilungen der Gesamtschullehrkräfte liegen. Dazu wird die Variationsbreite bzw. Spannweite (R) berechnet (Rasch et al., 2014a; Wilbert, 2015). Diese entspricht der Differenz zwischen dem maximalen und minimalen Wert der Verhaltensbeurteilung pro Schüler (ebd.) und wird zum einen pro Skala und zum anderen pro Item gebildet. Zur Interpretation der Werte werden erneut das arithmetische Mittel und die Standardabweichung herangezogen. Zudem wird das 25., 50. und 75. Perzentil berechnet (Döring & Bortz, 2016). Werden geringe Variationsbreiten erfasst bzw. erhoben, dass nicht die gesamte Skalenbreite zur Beurteilung des Verhaltens über die fünf Messzeitpunkte herangezogen wurde, wird dies als Indikator für den Urteilsfehler einer mangelnden Differenzierung aufgefasst (ebd.). Entsprechende Ergebnisse würden für die Neukonstruktion der Ratingskala sprechen (ebd.).

Weiterhin ist anzumerken, dass die Items festgelegte und vorgegebene Antwortkategorien darstellen und keine Möglichkeit bieten eigene Antworten einzubringen (Bühner, 2011). Einige der Lehrkräfte haben die Verhaltensbeurteilungen jedoch durch Kommentare bzw. Anmerkungen ergänzt. Diese Kommentare und Anmerkungen wurden dahingehend überprüft, ob es sich um Erläuterungen oder Beispiele für die Items handelt oder, ob grundlegende Sinnveränderungen vorgenommen wurden. Eine entsprechende Analyse findet sich im Anhang (vgl. Anhang 3.12). Erläuterungen und Beispiele führten nicht zum Ausschluss der Daten, grundlegende Sinnveränderungen hingegen schon.

6.3.2. Operationalisierung

Der vorliegenden Studie liegt eine Direct Behavior Ratingskala zugrunde, die im Zuge des Forschungsprojektes LEVUMI entwickelt wurde. Anfänglich lag eine adaptierte Ratingskala vor, die von Gebhardt, Casale, Jungjohann und DeVries (i.D.) in Anlehnung an die deutschsprachige Fremdbeurteilungsversion des SDQ entwickelt worden war. Diese umfasste die Verhaltensbereiche „Verhaltensprobleme“, „Hyperaktivität“ und „Emotionale Probleme“ sowie den neu hinzugefügten Verhaltensbereich „Schulbezogenes Verhalten“. Der SDQ ist ein Instrument, welches sich bereits in zahlreichen Studien zur Verhaltensbeurteilung bewähren konnte

(vgl. z.B. Goodman et al., 2010; Voß & Gebhardt, 2017a). Es umfasst relevante Verhaltensweisen des schulischen Settings und ermöglicht auf Basis der erhobenen Daten Entscheidungen, die für das entsprechende schulische Setting pädagogisch und praktisch von Bedeutung sind (Voß & Gebhardt, 2017a). Eine Überarbeitung und Erprobung der Ratingskala wurde von Sauerland (i.D.) im Rahmen ihrer Masterarbeit durchgeführt. Zunächst erweiterte sie die adaptierte Version der Ratingskala von Gebhardt et al. (i.D.) in Anlehnung an den SDQ um zwei Verhaltensbereiche („Prosoziales Verhalten", „Verhaltensprobleme mit Gleichaltrigen"). Im Anschluss daran wurde der Einsatz der Ratingskala im schulischen Setting erprobt. Stichprobenartig ausgewählte Lehrkräfte setzten die Ratingskala im schulischen Setting ein. Die Lehrkräfte wurden anschließend in Anbetracht ihrer Fachexpertise interviewt. Ihre Anmerkungen fanden in einer weiteren Überarbeitung der Ratingskala Berücksichtigung (Sauerland, i.D.). Im Anschluss daran wurde eine Untersuchung der Interrater-Reliabilität durchgeführt. Dazu beurteilten zwei geschulte Beobachterinnen und jeweils eine Lehrkraft parallel das Verhalten verschiedener SchülerInnen in der gleichen Beobachtungssituation. Die Interrater-Reliabilität lieferte zufriedenstellende Werte (zwischen ρ = .724 und ρ = .853) (ebd., i.D.). Insbesondere zwischen den beiden geschulten Beobachterinnen konnte eine starke positive Korrelation festgestellt werden (ρ = .837) (ebd., i.D.). Unter Berücksichtigung weiterer qualitativer Rückmeldungen sowohl der geschulten Beobachterinnen und der Lehrkräfte als auch eines externen Experten wurde die Ratingskala erneut überarbeitet. Die in diesem Zuge entwickelte Ratingskala liegt dieser Studie vor. Sie umfasst sechs Verhaltensbereiche: „Schulbezogenes Verhalten", „Verhaltensprobleme", „Hyperaktivität", „Emotionale Probleme", „Prosoziales Verhalten" und „Verhaltensprobleme mit Gleichaltrigen".

Die Items der Ratingskala wurden basierend auf einem nomothetischen Ansatz generiert. Die Autoren Gebhardt, Casale, Jungjohann und DeVries (i.D.) sowie Sauerland (i.D.) orientierten sich bei der Konstruktion der Ratingskala an den bereits umfänglich evaluierten Items der Ratingskala des SDQ (vgl. Kapitel 4). Insgesamt besteht die Ratingskala aus 19 Items. Die Items entsprechen sogenannten Indikatoren und stellen beobachtbare Variablen dar, mittels derer die Ausprägungen der zu untersuchenden theoretischen Konzepte bzw. Konstrukte erhoben werden können (Döring & Bortz, 2016a; Sikora, 2015b). Bei der vorliegenden Ratingskala handelt es sich um eine Multi-Item-Skala. Die der Ratingskala zugrundeliegenden Konstrukte werden jeweils über mehrere Indikatoren operationalisiert (Döring & Bortz, 2016). Die jeweilige Gruppierung der Indikatoren entspricht einer psychometrischen Skala (ebd.). Die Ratingskala umfasst die sechs psychometrischen Skalen SV (Schulbezogenes Verhalten), VP (Verhaltensprobleme), HY (Hyperaktivität), EP (Emotionale Probleme), PS (Prosoziales Verhalten) und VPG (Verhaltensprobleme mit Gleichaltrigen). Nachfolgend wird kurz erläutert werden, welche Verhaltensdimensionen mit den jeweiligen Skalen erhoben werden sollen. Die Skala SV erhebt das für den schulischen Kontext relevante Arbeits- und Sozialverhalten der SchülerInnen,

welches erfolgreiche schulisches Lernprozesse ermöglicht, wie z.B. die aktive Beteiligung am Unterricht, die konzentrierte Arbeit an Aufgaben und die Einhaltung von Regeln (Casale et al., 2015c; Henning et al., 2017; Lohbeck et al., 2015a). Die Skala VP erfasst Verhaltensprobleme, welche den externalisierenden Verhaltensproblemen zugeordnet werden können und somit nach außen, gegen die Lehrkräfte und MitschülerInnen gerichtet sind (Goodman et al., 2010). Es werden Verhaltensweisen aufgegriffen, die als oppositionell und aufsässig beschrieben werden können (Myschker & Stein, 2014). Die Skala HY erfasst ebenfalls externalisierende Verhaltensstörungen und im Spezifischen den Bereich ADHS (Goodman et al., 2010). In den drei Items der Ratingskala finden die drei Leitsymptome Unaufmerksamkeit, Hyperaktivität und Impulsivität Berücksichtigung (Steinhausen, 2016). Die Items der Skala EP operationalisieren die zwei internalisierenden Verhaltensstörungen der Angststörung und der depressiven Störung, welche sich auf der objektiven, körperlichen Ebene z.B. in Form von Anspannung oder Energielosigkeit zeigen (Goodman et al., 2010; Nevermann, 2008; Reicher & Rossmann, 2008). Die konkreten Items der Skala PS operationalisieren die sozialen Kompetenzen der SchülerInnen und erheben ihre Fertigkeiten zur Bildung positiver Sozialbeziehungen (Beelmann & Raabe, 2007). Dazu gehören u.a. Verhaltensweisen wie das Anbieten von Hilfeleistungen. Die Skala VPG erfasst ebenfalls internalisierende Verhaltensstörungen (Goodman et al., 2010). Verhaltensprobleme mit Gleichaltrigen werden in diesem Sinne durch Störungen des Sozialkontaktes operationalisiert. Dazu zählt z.B. die Ablehnung durch MitschülerInnen oder ein unreifes und erwachsenenabhängiges Sozialverhalten (Döpfner & Petermann, 2012; Lohbeck et al., 2014a). Die Skalen VP, HY, EP, PS und VPG entsprechen den fünf psychometrischen Skalen des SDQ. Diese 5-faktorielle Struktur konnte für den SDQ in zahlreichen nationalen und internationalen Studien nachgewiesen werden (vgl. z.B. Goodman et al., 2010; Koglin et al., 2007; Lohbeck et al., 2015b). Voß und Gebhardt (2017a) empfehlen, dass im Zuge einer Weiterentwicklung des SDQs zum Direct Behavior Rating Items hinzugefügt werden sollten, die es ermöglichen SchülerInnen zu identifizieren, die Verhaltensweisen in einer grenzwertigen Ausprägung zur Verhaltensauffälligkeit zeigen (Voß & Gebhardt, 2017a). Zum anderen liegt die Auswahl des Verhaltensbereichs „Schulbezogenes Verhalten“ darin begründet, dass die Ratingskala durch diese Erweiterung eine Erfassung des Arbeitsverhaltens ermöglicht, welches insbesondere im schulischen Setting von Relevanz ist (ebd.). Dementsprechend wurde die Ratingskala um eine sechste Skala, die Skala „Schulbezogenes Verhalten“ (SV) erweitert. Neben diesem 5-faktoriellen Modell konnte in der Forschungsliteratur ebenfalls ein 3-faktorielles Modell bestätigt werden (vgl., z.B. Di Riso et al., 2010; DeVries et al., 2017; Goodman et al., 2010), welches sich in die drei Skalen EXT (Externalisierende Verhaltensweisen), INT (Internalisierende Verhaltensweisen) und PS (Prosoziale Verhaltensweisen) gliedert. Die beiden Skalen EXT und INT setzen sich dabei im Sinne eines Modells zweiter Ordnung aus je zwei Subskalen des 5-faktoriellen Modells zusammen (Goodman et al., 2010). Die

erste Skala besteht aus den Subskalen VP und HY. Die zweite Skala umfasst die Subskalen EP und VPG. Durch die Skala SV erweitert sich die dimensionale Struktur des 3-Faktormodells um einen Faktor auf ein 4-Faktormodell. Dieses 4-faktorielle Modell wird zur Untersuchung der oben genannten Forschungsfragen herangezogen, da sich das zugrundeliegende 3-faktorielle Modell im Besonderen zur Untersuchung von Stichproben mit Kindern und Jugendlichen eignet, die wenige Risikofaktoren aufweisen und eine Verwendung des 5-faktoriellen Modells nicht gerechtfertigt scheint (ebd.). Goodman et al. (2010) geben an, dass das 5-faktorielle Modell nur dann herangezogen werden sollte, wenn Kinder mit hohen Risikofaktoren untersucht werden würden, also Kinder mit psychischen Störungen und/oder hohen Ergebnissen in den Subskalen. Dies trifft für die vorliegende Stichprobe nicht zu. Die Lehrkräfte wurden zwar dazu aufgefordert das Verhalten von SchülerInnen zu beobachten, die in mindestens einem der sechs aufgeführten Verhaltensbereiche auffällige Verhaltensweisen zeigen, jedoch wurden die Lehrkräfte zudem dazu aufgefordert alle Verhaltensbereiche zu beurteilen, unabhängig davon, ob der Schüler oder die Schülerin in allen oder nur einem Verhaltensbereich ein auffälliges Verhalten zeigte. Somit ist über alle Skalen mit durchschnittlichen und nicht extremen bzw. besonders hohen Mittelwerten zu rechnen. Im Zuge der Operationalisierung ist es von Bedeutung das Skalenniveau zu bestimmen (Sikora, 2015b). Wie bereits im Kapitel 6.3.1 erläutert, erlauben Ratingskalen intervallskalierte Messungen von psychologischen Variablen (Döring & Bortz, 2016). Die vorliegende Ratingskala stellt eine Intervallskala dar, da das Antwortformat als „annähernd gleichabständig“ (ebd., S. 245) bezeichnet werden kann. Die Abstände der Ratingskala werden durch numerische Marken (die Ziffern 1 bis 7) repräsentiert. Zur Charakterisierung der numerischen Abstufungen sind die Skalenendpunkten mit verbalen Marken (Nie und Immer) versehen worden. Die Begriffe „Nie“ und „Immer“ geben an, dass mittels der Ratingskala die Häufigkeit gezeigter Verhaltensweisen beurteilt wird (ebd.). Zur Interpretation der empirischen Unterschiede zwischen den SchülerInnen erscheint es sinnvoll die numerischen Marken von 2 bis 6 ebenfalls verbal zu charakterisieren. Die verbalen Marken sollten dann die Abstufungen der Ratingskala ebenfalls annähernd gleichabständig charakterisieren (ebd.). Döring und Bortz (2016) liefern zur Erhebung von Häufigkeiten die zwei folgenden 5-stufigen Beispiele:

- nie – selten – gelegentlich – oft – immer
- sehr selten – selten – gelegentlich – oft – sehr oft (ebd., S. 246).

Zur Interpretation der Ergebnisse der vorliegenden Studie werden die zwei 5-stufigen Skalierungen zu einer 7-stufigen Skalierung zusammengefasst und wie folgt angeordnet:

nie – sehr selten – selten – gelegentlich – oft – sehr oft – immer.

Da es sich bei dem Instrument um eine Ratingskala handelt, liegt ein gebundenes Antwortformat vor (Bühner, 2011). Die Antwortkategorien sind festgelegt und vorgegeben und bieten keine Möglichkeit eigene Antworten einzubringen (ebd.).

Neben der Ratingskala werden Interviews zur Beantwortung der dritten Forschungsfrage durchgeführt. Die Aussagen der Lehrkräfte in Bezug auf den Einsatz der Ratingskala im schulischen Setting stellen als „Expertenurteile“ (Sikora, 2015b, S. 71) Indikatoren für die Nebengütekriterien der Ökonomie und Nützlichkeit der Ratingskala dar. Sie werden mittels qualitativer Inhaltsanalyse nach Mayring (2015) ausgewertet. Im Sinne der Interpretationsform Zusammenfassung erfolgt eine Kategorienbildung (Mayring, 2015).

6.3.3. Aufbau der Erhebungsinstrumente

Zur Erprobung der Ratingskala wurde jeder Lehrkraft ein 32-seitiger Ordner überreicht. Dieser Ordner enthielt zur vereinfachten Durchführung und Übersichtlichkeit in chronologischer Reihenfolge ein Deckblatt zur Erhebung der Lehrerdaten, eine schriftliche Instruktion sowie für jeweils fünf SchülerInnen ein Deckblatt zur Erhebung der Schülerdaten und fünf Ratingskalen enthielt (vgl. Anhang 3.4). Auf der ersten Seite des Ordners, dem Deckblatt, wurden Angaben zur Schule und zur durchführenden Lehrkraft erfragt. Die Lehrkräfte sollten zum einen angeben, an welcher Schulform sie tätig sind und ob sie die Klasse der betreffenden SchülerInnen leiten. Zum anderen sollte sie Angaben zu ihrer Stundenanzahl pro Woche in der Klasse, zu ihrer Anstellung und zu ihren Erfahrungen in der Arbeit mit Ratingskalen machen. Sie hatten die Möglichkeit anzugeben, ob sie eine Auswertung der Verhaltensverlaufsmessung ihrer SchülerInnen erhalten möchten und konnten für Rückfragen freiwillig ihre Kontaktdaten angeben. Sie hatten die Möglichkeit ihre Namen anzugeben oder diesen mittels Code zu anonymisieren. Unabhängig davon, welche Angaben sie machten, wurden alle Daten anonymisiert erfasst und ausgewertet, um die Privatsphäre und Persönlichkeitsrechte der Lehrkräfte und auch der betreffenden SchülerInnen zu wahren. Darüber wurden die Lehrkräfte auf diesem Deckblatt und auch den Deckblättern zur Erhebung der Schülerdaten informiert. Auf der zweiten Seite des Ordners folgt die schriftliche Instruktion zur Durchführung des DBRs. Zunächst erhalten die Lehrkräfte allgemeine Informationen zur Methode des Direct Behavior Ratings. Darauf folgen Hinweise zum Beobachtungssetting und zur Durchführung und Dokumentation der Verhaltensverlaufsmessung. Eine ausführliche Erläuterung der Instruktion findet sich im nachfolgenden Kapitel 6.3.4. Ab der dritten Seite finden sich für fünf SchülerInnen je ein Deckblatt zur Erhebung der Schülerdaten sowie je fünf Ratingskalen. Mittels der Deckblätter werden Angaben zum Alter, zur Klassenstufe sowie zum Geschlecht der SchülerInnen erfragt. Zudem wird erhoben, ob das Kind einen diagnostizierten sonderpädagogischen Förderbedarf und einen Migrationshintergrund hat, also, ob das Kind im Ausland geboren wurde. Liegt ein diagnostizierter sonderpädagogischer Förderbedarf vor, wird die Art des sonderpädagogischen Förderbedarfs erfragt. Für jedes Kind lagen den Lehrkräften jeweils fünf Ratingskalen vor, da die Lehrkräfte das Verhalten der SchülerInnen an fünf aufeinanderfolgenden Tagen beurteilen sollten. Insgesamt setzt sich die Ratingskala aus 19 Items zusammen, welche entsprechend der sechs übergeordneten Verhaltensbereiche angeordnet und gruppiert sind. Als

erster Verhaltensbereich findet sich das schulbezogene Verhalten mit vier Items, gefolgt von den Bereichen Verhaltensprobleme, Hyperaktivität, emotionale Probleme, Prosoziales Verhalten und Verhaltensprobleme mit Gleichaltrigen mit je drei Items. Die Verhaltensbereiche entsprechen somit jeweils einer Multi-Item-Skala.

Bei der Skala handelt es sich um eine 7-stufige Ratingskala. Für jedes Item sind die numerischen Marken 1 bis 7 angegeben. Die Skalenendpunkte sind in der Spaltenbezeichnung mit den verbalen Marken „Nie“ (1) und „Immer“ (7) versehen. Die Ratingskala entspricht folglich einer unipolaren Häufigkeitsratingskala. Sie bildet „graduell abgestuft“ die Häufigkeit der Verhaltensweisen ab (Döring & Bortz, 2016, S. 245). Dabei ist zu beachten, dass hohe Werte in den Verhaltensbereichen Schulbezogenes Verhalten und Prosoziales Verhalten die Anwesenheit angemessener Verhaltensweisen repräsentierten. Wohingegen hohe Werte in den Verhaltensbereichen Verhaltensprobleme, Hyperaktivität, Emotionale Probleme und Verhaltensprobleme mit Gleichaltrigen die Anwesenheit unangemessener Verhaltensweisen abbilden.

Die Lehrkräfte wurden zudem dazu aufgefordert nach jeder Verhaltensbeurteilung Angaben zur Beobachtungssituation zu machen. In einem Feld unterhalb der Ratingskala sollten sie zum einen das Datum der Verhaltensbeurteilung eintragen und zum anderen angeben, ob der Beobachtungszeitraum einem Schultag, einer Schulstunde oder einem anderen Zeitraum entsprach. Bei der Länge einer Schulstunde, wurden sie darum gebeten auch das Schulfach anzugeben, welches in der entsprechenden Schulstunde unterrichtet wurde. Entsprach der Beobachtungszeitraum einem anderen Zeitraum, wie z.B. einer Doppelstunde, konnte sie diesen aufführen. Zudem wurde erfragt, ob sich die Beobachtungen überwiegend auf Situationen im Klassenverband oder auf andere Situationen bezogen. Auch an dieser Stelle konnte im Falle einer anderen Situation die jeweilige Situation aufgeführt werden.

Die Daten zum Einsatz der Ratingskala wurden zudem mithilfe von Interviews erhoben. Bei den Interviews handelt es sich um halbstrukturierte Interviews für die vorab ein Leitfaden entwickelt wurde (vgl. Anhang 3.5). Der Leitfaden umfasst einen Fragenkatalog, welcher sich in vier Themenblöcke gliedert. Der erste Themenblock „Einsatz der Ratingskala“ erfragt mit drei offenen Fragen, wie der Einsatz der Ratingskala war, wann ein Einsatz problemlos möglich war und wo Schwierigkeiten auftraten. Der zweite Themenblock „Instruktion“ erhebt zunächst mit einer geschlossenen Frage, ob die Instruktion verständlich war. Diese Frage bezieht sich sowohl auf die mündliche Schulung als auch auf die schriftliche Instruktion. Im Falle einer unklaren Instruktion, wurde mit offenen Fragen erfasst, was nicht verständlich war und welche Instruktion bzw. welcher Hinweis fehlte. Ebenso wurden im dritten Themenblock „Items“ erhoben, ob die Items verständlich waren. Wurde dies verneint, wurde erfragt, welche Items nicht verständlich waren. Im vierten Themenbereich wird das Kriterium der Ökonomie erhoben und mit einer geschlossenen Frage erfasst, ob die Ratingskala im Schulalltag häufig und schnell

eingesetzt werden kann. Wenn diese Frage verneint wurde, werden die Lehrkräfte darum gebeten zu erläutern, warum kein häufiger und schneller Einsatz der Ratingskala möglich ist. Diese Fragen dienten lediglich der Orientierung und konnten je nach Gesprächsverlauf angepasst werden. Zum Ende eines jeden Gespräches wurde den Lehrkräften die Möglichkeit gegeben eigene Anmerkungen machen und Erfahrungen zu erläutern, die nicht explizit erfragt wurden.

6.3.4. Durchführung

Zur Durchführung der Pilotierungsstudie wurden ab Januar 2018 Grundschulen und Gesamtschulen im Raum Dortmund und Münster kontaktiert. Interessierte SchulleiterInnen und Lehrkräfte erhielten zur ersten Information das offizielle Anschreiben zur Studie (vgl. Anhang 3.1) sowie eine Kurzinformation zum Forschungsprojekt LEVUMI (vgl. Anhang S. 3.2). Auf diesem Wege konnten 39 Lehrkräfte an neun verschiedenen Schulen für eine Teilnahme an der Studie gewonnen werden. Die Ordner wurden den Lehrkräften persönlich ausgehändigt. Die Übergabe wurde genutzt, um mindestens eine Lehrkraft pro Schule zu schulen. In diesem Zuge wurde das Forschungsprojekt LEVUMI, die Pilotierungsstudie, die Thematik der Verhaltensverlaufsdiagnostik und der Ansatz des Direct Behavior Ratings vorgestellt sowie der geplante Einsatz der Ratingskala erläutert. Für die Präsentation dieser Informationen wurde ein Poster bzw. Handout erstellt (vgl. Anhang S. 3.3). An einer Grundschule wurde die Pilotierungsstudie im Rahmen einer Lehrerkonferenz vorgestellt. Da es aufgrund der Größe der Stichprobe und der räumlichen Verteilung der Schulen zeitlich nicht möglich war jede teilnehmende Lehrkraft persönlich zu schulen, enthielt jeder Ordner eine schriftliche Instruktion zur Durchführung der Verhaltensbeurteilung. Zudem wurden die geschulten Lehrkräfte gebeten, für die weiteren Lehrkräfte der Schule als Ansprechpartner zur Verfügung zu stehen und bei Bedarf den Einsatz der Ratingskala zu erläutern. Die schriftliche Instruktion informiert über die Methode des Direct Behavior Ratings und gibt Hinweise zum Beobachtungssetting, zur Durchführung und Dokumentation der Verhaltensbeurteilungen. Die Lehrkräfte sollten darauf achten, dass der Beobachtungszeitraum nicht über einen Unterrichtstag hinausging und dass während der Beobachtungen Sichtkontakt zu den SchülerInnen bestand. Sie sollten den Fokus auf Situationen im Klassenverband legen und nach Möglichkeit Beobachtungssituationen wählen, die hinsichtlich Länge (z.B. Unterrichtstag) und Setting (z.B. Klassenverband) vergleichbar waren. Sie wurden dazu aufgefordert, die Items vor der Beobachtung zu lesen und im Anschluss daran fünf SchülerInnen auszuwählen, die nach ihrer Einschätzung in mindestens einem der sechs Verhaltensbereiche ein auffälliges Verhalten zeigten. (Anmerkung: Es wurde darauf verzichtet, vorab ein Screeningverfahren, wie den SDQ, zur Bestimmung problematischer Verhaltensbereiche einzusetzen, da Sauerland (i.D.) in ihrer Studie feststellte, dass die Lehrkräfte das Screeningverfahren nicht zur Auswahl problematischer Verhaltensweisen nutzten. Stattdes-

sen füllten sie die Ratingskala für alle Verhaltensbereiche aus, da sie sich auf dieser Grundlage eine umfassendere Einschätzung erhofften.) Die Lehrkräfte wurden gebeten, das Verhalten der fünf SchülerInnen an fünf aufeinanderfolgenden Tagen zu beurteilen. Als „aufeinanderfolgende“ Schultage werden die Schultage verstanden, an denen die Lehrkraft die SchülerInnen unterrichtet. Wenn die Lehrkraft die SchülerInnen an allen Wochentagen unterrichtet, sind das zum Beispiel alle Wochentage von Montag bis Freitag. Zwischen zwei aufeinanderfolgenden Tagen kann mitunter aber auch ein weiterer Schultag liegen, an dem die Lehrkraft die SchülerInnen nicht unterrichtet. Sie wurden außerdem dazu aufgefordert, ihre Beurteilungen auf die unmittelbar vorhergehenden Beobachtungen zu beziehen. Sofern sie eine Verhaltensweise auf Basis ihrer Beobachtungen nicht beurteilen konnten, sollten sie diese auslassen (Beispiel: keine Beurteilung des Verhaltens in Gruppensituationen, wenn keine Gruppenarbeit durchgeführt wurde). Ferner war es von Bedeutung, dass immer die gleiche Lehrkraft die Beurteilungen durchführte (vgl. Casale et al., 2015c). Für Rückfragen waren die Kontaktdaten der Studierenden aufgeführt. Bei Rückfragen wendeten sich die Lehrkräfte in den meisten Fällen an die geschulten KollegInnen, welche wiederum die Studierenden kontaktierten. Lediglich eine Lehrkraft kontaktierte die Studierenden auf direktem Wege. Die Lehrkräfte führten die Beurteilungen im Zeitraum vom 24.01.2018 bis 25.04.2018 eigenverantwortlich durch. Die Studierenden holten die Ordner nach Beendigung der Verhaltensbeurteilungen ab. Nach Beendigung der Verhaltensbeurteilungen sind vier Interviews durchgeführt worden. Pro Schulform wurden zwei Lehrkräfte interviewt, jeweils eine Regelschullehrkraft und eine sonderpädagogische Lehrkraft. Drei der Interviews wurden telefonisch durchgeführt und auditiv dokumentiert. Ein Interview wurde im Lehrerzimmer der Schule durchgeführt und auf Wunsch schriftlich dokumentiert. Dazu wurde ein Interviewleitfaden entworfen. Es wurde erfragt, wie der Einsatz der Ratingskala war, ob die Instruktion und die Items verständlich waren und ob die Ratingskala im Schulalltag häufig und schnell eingesetzt werden kann. Die Lehrkräfte hatten die Möglichkeit eigene Anmerkungen bzw. Anliegen vorzubringen.

6.3.5. Stichprobe

Die Pilotierungsstudie ist in einem Zeitraum von drei Monaten (24.01. bis 25.04.2018) im Raum Dortmund und Münster an fünf Grund- und vier Gesamtschulen durchgeführt worden. Insgesamt beurteilten N = 39 Lehrkräfte von neun verschiedenen Schulen das Verhalten von N = 205 SchülerInnen. Von den N = 39 Lehrkräften setzten N = 23 Lehrkräfte die Ratingskala an Grundschulen (59%) und N = 16 Lehrkräfte die Ratingskala an Gesamtschulen (41%) ein. Diese Verteilung ist darauf zurückzuführen, dass anteilig mehr Lehrkräfte der Primarstufe auf direktem Wege kontaktiert werden konnten. Wurde ein erster Kontakt zu den Lehrkräften auf indirektem Wege, z.B. über das Sekretariat hergestellt, wie es überwiegend bei den Gesamtschulen der Fall war, fielen die positiven Rückmeldungen und Zusagen der Lehrkräfte anteilig

deutlich niedriger aus. Die Grundschullehrkräfte haben das Verhalten von *N* = 108 SchülerInnen (52,7%) und die Gesamtschullehrkräfte das Verhalten von *N* = 97 SchülerInnen (47,3%) beurteilt. Insgesamt *N* = 7 Lehrkräfte (17,9%) sind als sonderpädagogische Lehrkräfte angestellt. Davon arbeiten *N* = 4 Lehrkräfte an Grundschulen und *N* = 3 Lehrkräfte an Gesamtschulen. Die übrigen *N* = 32 Lehrkräfte (82,1%) arbeiten als Regelschullehrkräfte an den jeweiligen Schulen. Die Verteilung der Regelschullehrkräfte auf die Schulformen kann der Tabelle 1 entnommen werden. *N* = 31 Lehrkräfte (79,5%). gaben an, als KlassenlehrerIn der betreffenden SchülerInnen zu arbeiten. Der Anteil der Klassenlehrer ist mit 82,6% an den Grundschullehrkräften (*N* = 23, davon Klassenlehrer *N* = 19) etwas höher als der Anteil an den Gesamtschullehrkräften mit 75%. Jeweils vier Lehrkräfte pro Schulform gaben an, nicht als Klassenleitung zu arbeiten. Ferner wurden die Lehrkräfte befragt, wie viele Schulstunden sie durchschnittlich pro Woche in den Klassen der beobachteten SchülerInnen tätig waren. Bei einem Vergleich der Mittelwerte wird deutlich, dass die Grundschullehrkräfte (*M* = 18,40) durchschnittlich fast dreimal länger mit den SchülerInnen arbeiteten als die Gesamtschullehrkräfte (*M* = 6,94). Diese Werte waren zu erwarten, da Lehrkräfte an Grundschulen vermehrt nach dem Klassenlehrerprinzip unterrichten, während Lehrkräfte an Gesamtschulen vorwiegend nach dem Fachlehrerprinzip arbeiten (KMK, 2017). Ein Drittel aller Lehrkräfte (*N* = 13) konnte Vorerfahrungen in der Arbeit mit Ratingskalen vorweisen. Unter Berücksichtigung der Schulformen haben acht Grundschullehrkräfte (34,8%) und fünf Gesamtschullehrkräfte (31,3%) bereits mit Ratingskalen gearbeitet.

Die Tabelle 1 stellt die Verteilungen der Lehrkräfte im Hinblick auf Anstellung, Klassenleitung und Erfahrungen in der Arbeit mit Ratingskalen pro Schulform dar.

Tabelle 1: Anzahl der LehrerInnen pro Schulform unter Berücksichtigung der Variablen Anstellung, Klassenleitung und Erfahrungen in der Arbeit mit Ratingskalen

		Grundschule (*n* = 23)	Gesamtschule (*n* = 16)	Gesamt (*N* = 39)
		N (%)	*N (%)*	*N*
Anstellung	Regelschullehrkraft	19 (82,6)	13 (81,3)	32
	sonderpädagogische Lehrkraft	4 (17,4)	3 (18,7)	7
Klassenleitung	Klassenleitung	19 (82,6)	12 (75)	31
	keine Klassenleitung	4 (17,4)	4 (25)	8
Erfahrungen	Erfahrungen	8 (34,8)	5 (31,3)	13
	keine Erfahrungen	15 (65,2)	11 (68,7)	26

Bei der Betrachtung der Geschlechterverteilung der SchülerInnen wird deutlich, dass,von den *N* = 205 SchülerInnen *N* = 60 (29,3%) weiblich und *N* = 145 (70,7%) männlich. In den zwei Schulformen ist eine ähnliche Geschlechterverteilung zu finden mit 28,7% weiblichen und 71,3% männlichen GrundschülerInnen sowie 29,9% weiblichen und 70,1% männlichen GesamtschülerInnen.

Die SchülerInnen, deren Verhaltensweisen von den Lehrkräften beobachtet wurden, waren zwischen 6;4 und 18;5 Jahre alt. Es liegt somit eine Altersspanne von 12;2 Jahren vor. Die Schüleranzahlen verteilen sich wie folgt auf die Jahrgangsstufen 1 bis 10.

Tabelle 2: Anzahl und durchschnittliches Alter der SchülerInnen pro Jahrgangsstufe

Jahrgangsstufen	Häufigkeit *N (%)*	Alter *M (SD)*
1 – 4	108 (52,7)	8,56 (1,29)
5 – 7	41 (20,0)	12,13 (0,86)
8 – 10	55 (26,8)	14,49 (1,12)

Für eine Schülerin der Gesamtschule wurde keine Angabe zur Klassenstufe gemacht.

Insgesamt weisen 17,1% der SchülerInnen (*N* = 35) einen sonderpädagogischen Förderbedarf auf. Auch im Hinblick auf diesen Aspekt ist eine ähnliche Verteilung in beiden Schulformen vorzufinden. So haben 16,7% der GrundschülerInnen und 17,5% der GesamtschülerInnen einen diagnostizierten sonderpädagogischen Förderbedarf. Die Verteilung über die verschiedenen Förderschwerpunkte kann der nachfolgenden Tabelle 3 entnommen werden.

Tabelle 3: Anzahl und prozentualer Anteil diagnostizierter Förderschwerpunkte

	Grundschule (*N* = 108)	Gesamtschule (*N* = 97)	Gesamt (*N* = 205)
	N (%)	N (%)	N (%)
FSP ESE	4 (3,7)	9 (9,3)	13 (6,3)
FSP LE	8 (7,4)	9 (9,3)	17 (8,3)
FSP SB	7 (6,5)	1 (1,0)	8 (3,9)
FP GG	0 (0,0)	1 (1,0)	1 (0,5)

Anmerkungen. FSP ESE = Förderschwerpunkt emotionale und soziale Entwicklung; FSP LE = Förderschwerpunkt Lernen; FSP SB = Förderschwerpunkt Sprache; FSP GG = Förderschwerpunkt geistige Entwicklung

In der Tabelle 3 finden sich *N* = 39 förderschwerpunktspezifische Angaben bei *N* = 35 SchülerInnen mit sonderpädagogischem Förderbedarf, da für einen Grundschüler und drei GesamtschülerInnen zwei diagnostizierte Förderschwerpunkte angegeben wurden.

20,5% aller SchülerInnen sind im Ausland geboren. Ein Vergleich zwischen den Schulformen zeigt einen etwas höheren Anteil von GrundschülerInnen mit Migrationshintergrund (25%) als GesamtschülerInnen mit Migrationshintergrund (15,6%). Für eine Gesamtschülerin wurden keine Angaben zum Migrationshintergrund gemacht. Da bis zu sieben Lehrkräfte einer Schule an der Studie teilnahmen, liegen N = 8 Fälle vor, in denen N = 4 GrundschülerInnen jeweils von zwei verschiedenen Lehrkräften beobachtet wurden. Es wurde überprüft, ob die Lehrkräfte zur selben Zeit in derselben Situation dasselbe Verhalten beurteilt haben. Dies war nicht der Fall. Somit werden die N = 8 unterschiedlichen Beurteilungen getrennt voneinander betrachtet und es liegen bei N = 205 unterschiedlichen Schülerinnen und Schülern N = 209 unabhängige Fälle vor. Unter Berücksichtigung der Schulformen liegen N = 112 Fälle (53,6%) aus Grundschulen und N = 97 Fälle (46,4%) aus Gesamtschulen vor. Die nachfolgenden Auswertungen beziehen sich auf die Fälle.

Insgesamt wurden vier Lehrkräfte interviewt, zwei Grundschul- und zwei Gesamtschullehrkräfte. Pro Schulform wurden jeweils eine Regelschullehrkraft und eine sonderpädagogische Lehrkraft befragt. Beide Regelschullehrkräfte sind Klassenlehrer der SchülerInnen, die sonderpädagogischen Lehrkräfte haben keine Klassenleitung inne. Die beiden sonderpädagogischen Lehrkräfte gaben an, Erfahrungen in der Arbeit mit Ratingskalen zu haben, wohingegen die Regelschullehrkräfte äußerten, über keine Erfahrungen zu verfügen. Die Regelschullehrerin der Grundschule unterrichtet 14 Stunden pro Woche in der Klasse der beobachteten SchülerInnen, die Regelschullehrerin der Gesamtschule sechs Stunden und die sonderpädagogische Lehrkraft der Gesamtschule ca. drei Stunden. Die sonderpädagogische Lehrkraft der Grundschule hat diesbezüglich keine Angaben gemacht.

6.4. Ergebnisse

In diesem Kapitel werden die Forschungsergebnisse der Pilotierungsstudie strukturiert dargestellt. Die Reihenfolge der Darstellung orientiert sich an den zuvor aufgestellten Forschungsfragen (vgl. Kapitel 6.1) sowie an dem chronologischen Auftreten der Skalen in der Ratingskala.

Forschungsfrage 1: Wie setzen Lehrkräfte an Grund- und Gesamtschulen die Ratingskala ein? Handelt es sich bei der Ratingskala um ein ökonomisches und nützliches Testinstrument?

Nachfolgend werden die Angaben der Lehrkräfte zu Beobachtungssituationen, Beobachtungszeiträumen, Beobachtungssettings und Erhebungszeiträumen dargestellt. Zur Untersuchung dieser Angaben wurden unter Berücksichtigung der Schulform Häufigkeitsanalysen durchgeführt. Des Weiteren werden Erfahrungen und Anmerkungen der interviewten Lehrkräfte erläutert.

Erhebungszeiträume

Zur Berechnung der Länge des Erhebungszeitraums wurde die Anzahl der Schultage zwischen dem ersten und dem letzten Messzeitpunkt bestimmt. Ausgeschlossen wurden Wochenenden, Ferientage und Feiertage. Tabelle 4 veranschaulicht über welchen Erhebungszeitraum Grundschul- und Gesamtschullehrkräfte fünf Beurteilungen des Schülerverhaltens vornahmen.

Tabelle 4: Häufigkeitsverteilung des Erhebungszeitraums (in Tagen) pro Schulform

Erhebungszeitraum (in Tagen)		Grundschullehrkräfte (*N* = 107)	Gesamtschullehrkräfte (*N* = 89)
Median		5,00	9,00
Minimum		5	5
Maximum		21	30
Perzentile	25	5,00	6,00
	50	5,00	9,00
	75	5,00	15,00
M (SD)		6,32 (3,53)	11,62 (7,01)

Anmerkung: keine Berücksichtigung in der Berechnung fanden Wochenenden, Ferientage und Feiertage.

Die Lehrkräfte der Grundschulen setzten die Ratingskalen in Zeiträumen von bis zu 21 Tagen ein. Gesamtschullehrkräfte führten fünf Beurteilungen an bis zu 30 Schultagen durch. Im Durchschnitt setzten die Grundschullehrkräfte die Ratingskalen in Zeiträumen von *M* = 6,32 Schultagen und die Gesamtschullehrkräfte in Zeiträumen von *M* = 11,62 Schultagen ein. 76,6% aller Grundschullehrer war es möglich die Ratingskala in einem Zeitraum von fünf unmittelbar aufeinanderfolgenden Schultagen einzusetzen. Hingegen war es nur 22,5% der Gesamtschullehrer möglich die Ratingskala im selbigen Zeitraum anzuwenden.

Beobachtungszeiträume

Außerdem wurden die Lehrkräfte befragt, wie lang die Beobachtungszeiträume waren, auf die sich ihre Beurteilungen bezogen. Die Tabelle 5 zeigt die Ergebnisse.

Tabelle 5: Häufigkeitsverteilung des Beobachtungszeitraum pro Schulform

Beobachtungszeitraum	Primarstufe (*N* = 560)		Sekundarstufe (*N* = 479)	
	Häufigkeit (N)	Prozent (%)	Häufigkeit (N)	Prozent (%)
eine Schulstunde	170	30,4	402	82,9
ein Schultag	365	65,2	0	0
anderer Zeitraum	25	4,5	77	15,9

Tabelle 5 verdeutlicht, dass mit 65,2% die Mehrheit, der in der Primarstufe tätigen Lehrkräfte, einen Zeitraum von einem Schultag beobachtet hat. Die Mehrheit der Gesamtschullehrkräfte (82,9%) wählte hingegen als Grundlage für die Verhaltensbeurteilungen einen Beobachtungszeitraum von einer Schulstunde. Diesen Zeitraum wählten die Grundschullehrkräfte in 30,4% der Fälle. Den Zeitraum von einem Schultag wählten Gesamtschullehrkräfte hingegen gar nicht. In 4,5% der Fälle beobachteten Lehrkräfte der Primarstufe und in 15,9% der Fällen Lehrkräfte der Sekundarstufe andere Zeiträume wie beispielsweise zwei Schulstunden bzw. eine Doppelstunde.

Beobachtungssituation

Von den *N* = 560 Grundschullehrkräften bezogen 89,5% und von den *N* = 479 Gesamtschullehrkräften bezogen 77,1% der Lehrkräfte ihre Beurteilungen auf Situationen im Klassenverband. In 10,5% der Fälle beurteilten Grundschullehrkräfte und in 21,6% der Fälle Gesamtschullehrkräfte das Verhalten der SchülerInnen vor dem Hintergrund anderer Situationen. Unter diese Kategorie fallen beispielsweise Verhaltensbeurteilungen im Setting der Kleingruppe oder des Einzelunterrichts, wie sie oftmals von sonderpädagogischen Lehrkräften angegeben wurden.

Tabelle 6: Häufigkeitsverteilung der Beobachtungssituationen pro Schulform

Beobachtungssituation	Primarstufe (*N* = 560)		Sekundarstufe (*N* = 479)	
	Häufigkeit (N)	Prozent (%)	Häufigkeit (N)	Prozent (%)
Klassenverband	501	89,5	374	77,1
anderer Zeitraum	59	10,5	105	21,6

Befragung

Es schließt sich eine strukturierte Darstellung der Ergebnisse der qualitativ ausgewerteten Interviews an, welche basierend auf der Fachexpertise der Lehrkräfte umfangreiche Informationen zum Einsatz der Ratingskala im schulischen Kontext liefern. Den Lehrkräften wurden Fragen zum allgemeinen Einsatz der Ratingskala, zur Verständlichkeit der Items und zur Verständlichkeit der Instruktion gestellt. Des Weiteren wurde erfragt, ob ein ökonomischer Einsatz der Ratingskala möglich war. Dazu wurden vier Lehrkräfte befragt interviewt. Die qualitative Auswertung der Interviews orientierte sich an der qualitativen Inhaltsanalyse nach Mayring (2015), genauer der Interpretationsform der Zusammenfassung (Mayring, 2015). Die zwei Reduktionsdurchgänge der Zusammenfassung zur qualitativen Kategorienbildung finden sich in tabellarischer Form im Anhang (vgl. Anhang 3.11).

Die Auswertung der Interviews zeigt, dass bei angemessener Schülerzahl ein problemloser Einsatz möglich ist. Die Anwendung der Ratingskala ist vergleichbar zu anderen Ratingskalen. Unter einer angemessenen Schülerzahl wird bei paralleler Beobachtung eine Anzahl von einem bis drei SchülerInnen verstanden. Im Rahmen dieser Studie wurden die Lehrkräfte darum gebeten fünf SchülerInnen zu beobachten. Beobachtungen von fünf SchülerInnen waren für den Sonderpädagogen der Gesamtschule umsetzbar. Die Regelschullehrkraft der Grundschule äußerte hingegen, dass sie diese Anzahl für zu umfangreich hielt. Eine Auswertung der Anzahl durchgeführter Verhaltensbeurteilungen zeigt, dass die teilnehmenden Lehrkräfte zwischen 2 und 20 verschiedenen Schülerinnen und Schülern beobachtet haben. Nicht alle SchülerInnen wurden dabei parallel in derselben Schulstunde beobachtet. Die Regelschullehrkräfte gaben an, dass der Einsatz der Ratingskala einen zusätzlichen Zeitaufwand darstelle. Insbesondere die Einarbeitungen, erste Durchführungen und die Reflexion der Beurteilungen würden zeitaufwändige Phasen im Einsatz der Ratingskala darstellen. Ferner gaben die Lehrkräfte an, dass dieser Zeitaufwand bei regelmäßigen, routinierten und unmittelbaren Einsatz der Ratingskala überschaubar bliebe und in einem angemessenen Zeitrahmen läge. Die Regelschullehrkraft der Gesamtschule merkte an, dass Beurteilungen bei entsprechenden Unterrichtsphasen, z.B. Stillarbeit, auch unterrichtsbegleitend möglich wären. Hingegen gab die Regelschullehrkraft der Grundschule an, dass das gleichzeitige Unterrichten, genaue Beobachtungen und Beurteilungen im Zuge von langen Beobachtungszeiträumen erschwerten. Lange Beobachtungszeiträume, wie ein Schultag, würden außerdem mit zahlreichen weiteren Eindrücken einhergehen, die die genauen Beobachtungen und Beurteilungen ebenfalls erschweren würden. Im Allgemeinen empfand sie es leichter extreme Ausprägungen sowohl angemessener als auch unangemessener Verhaltensweisen zu beurteilen. Die Lehrkräfte gaben an, dass die SchülerInnen über längere Beobachtungszeiträume große Schwankungen im Verhalten

zeigten, wobei unter längeren Zeiträumen sowohl ein Schultag als auch ein Schuljahr verstanden wird. Sie äußerten, dass der Einsatz der Ratingskala in diesem Zuge nützlich sei. Die Ratingskala könne dabei über lange Zeiträume und zahlreiche Messzeitpunkte schnell und häufig eingesetzt werden. Nützlich sei die Skala zudem unter dem Aspekt, dass sie reguläre Beobachtungen, die ohnehin durchgeführt werden, ergänze. Die Auswertung der Ratingskala könne Informationen liefern, die für persönliche Gespräche mit Kollegen und Eltern herangezogen werden könnten. Die Sonderpädagogin der Grundschule wies daraufhin, dass die Skala für Unterrichtssettings im Klassenverband geeignet ist, eine Übertragung auf das sonderpädagogische Unterrichtssetting jedoch nur bedingt möglich sei. Die Settings weisen große Unterschiede auf. Überdies hinaus treten in beiden Settings unterschiedliche Problematiken auf. Im sonderpädagogischen Setting seien beispielsweise aufgrund häufiger Einzel- oder Kleingruppensettings prosoziale Verhaltensweisen (Skala PS) und im Besonderen Gruppenarbeitsweisen (Item PS16) seltener zu beobachten als im Klassenverband. Sonderpädagogen würden des Weiteren häufiger mit SchülerInnen mit externalisierenden Verhaltensweisen arbeiten als mit SchülerInnen mit internalisierenden Verhaltensweisen. Es sei schwierig, emotionale Probleme im Allgemeinen auf Basis von Beobachtungen und im Speziellen auf Basis der vorliegenden Items dieser Skala zu beurteilen. Erschwert werden diese Beurteilungen durch die Anwesenheit vieler SchülerInnen. Der Sonderpädagoge der Gesamtschule gab an, dass die Skalen bzw. Verhaltensbereiche verständlich seien. Die Regelschullehrerin der Grundschule gab an, dass der Wechsel der Valenz zwischen den Skalen (Bsp.: hohe Werten der Skala SV repräsentieren angemessene Verhaltensweisen, hohe Werte der Skala VP repräsentieren unangemessene Verhaltensweisen) unpraktisch sei und auch bei erprobtem Einsatz Konzentration erfordere. Dies ist am Ende von längeren Beobachtungszeiträumen ungünstig. Aufgrund zahlreicher zusätzlicher Anforderungen im Schulalltag sei eine schnelle Bearbeitung der Skala wünschenswert. Ein Einsatz der Skala ist sowohl auf Basis der mündlichen Instruktion als auch auf Basis der schriftlichen Instruktion möglich. Die Instruktionen wurden von allen Lehrkräften als verständlich bezeichnet (drei der vier Lehrkräfte erhielten eine kurze Schulung vor Einsatz der Ratingskala, einer Lehrkraft lag lediglich die schriftliche Instruktion vor). Die Items wurden ebenfalls als verständlich bezeichnet. Insbesondere die Items SV01, SV02, SV03, VP05, HY08, HY09, PS16, VPG17, VPG19 konnten nach Angabe der Lehrkräfte aufgrund konkreter Operationalisierungen schnell beurteilt werden und ermöglichten eindeutige Aussagen. Eindeutige Items (z.B. SV01, SV02) würden die Beurteilungen erleichtern, mehrdeutige Items (SV04, VP05, VP06, VP07, VPG18) hingegen erschweren die Beurteilungen. Die beiden Lehrkräfte der Sekundarstufe sehen bei der Ratingskala aufgrund der geringen Anzahl der Items eine geringere Aussagekraft gegeben im Vergleich zu anderen testdiagnostischen Instrumenten. Die Skalenbreite der 7-Likert-Skala wurde zur Einschätzung der Verhaltensweisen als angemessen bezeichnet. Der Sonderpädagoge der Gesamtschule äußerte den Wunsch nach

einem computergestützten Auswertungstool. Die Sonderpädagogin der Grundschule regte eine Weiterentwicklung der Ratingskala für sonderpädagogische Settings und eine konkrete Operationalisierung der Items der Skala EP an.

Forschungsfrage 2: Wie ist die interne Konsistenz der Skalen des 4-Faktormodells pro Messzeitpunkt und über die Messzeitpunkte?

Nachfolgend wird zur Schätzung der Reliabilität die interne Konsistenz nach Cronbach's Alpha für die vier Skalen der Ratingskala bestimmt. Die Korrelationsanalysen wurden mit dem Programm SPSS durchgeführt. Fehlende Werte über die Messzeitpunkte führen zu einer fallweisen Löschung der Werte.
Wie der Tabelle 7 zu entnehmen ist, konnten für alle Skalen ausnahmslos über alle Messzeitpunkte zufriedenstellende Werte der internen Konsistenz ($\alpha \geq .70$) festgestellt werden. Gute Reliabilitätswerte ($\alpha \geq .80$) konnten im Besonderen für die Skalen EXT und PS erhoben werden.

Tabelle 7: Interne Konsistenz der Skalen nach Cronbach's Alpha (α) zu fünf Messzeitpunkten

Skala	MZ1		MZ2		MZ3		MZ4		MZ5	
	N	α	*N*	α	*N*	α	*N*	α	*N*	α
SV	173	.76	183	.77	182	.77	182	.79	179	.79
EXT	184	.88	181	.88	180	.87	182	.86	178	.89
INT	158	.75	158	.78	160	.79	162	.79	158	.79
PS	180	.92	170	.94	178	.93	174	.92	178	.93

Zur differenzierten Analyse der Items wurde die Trennschärfe der Items bestimmt. Hohe Trennschärfen (r_i > .50) zu den fünf Messzeitpunkten konnten für drei Items der Skala SV (SV02, SV03, SV04), alle Items der Skala EXT (VP05, VP06, VP07; HY08, HY09, HY10), drei Items der Skala INT (EP11, EP12, EP13) und alle Items der Skala PS (PS14, PS15, PS16) festgestellt werden. Wie die Tabelle 8 verdeutlicht, konnte für ein Item der Skala SV (SV01) und drei weitere Items der Skala INT (VPG17, VPG18, VPG19) in einigen Fällen nur mittelmäßige (.30 < r_i > .50) Werte festgestellt werden. Zu zwei Messzeitpunkten ließ sich sogar nur eine geringe (r_i < .30) Trennschärfe für das erste Item der Skala SV (SV01) feststellen.

Tabelle 8: Items mit geringen bis mittelmäßigen Itemtrennschärfen (r_i) nach Cronbach's Alpha zu den fünf Messzeitpunkten

Item	Skala	MZ1		MZ2		MZ3		MZ4		MZ5	
		N	r_i	*N*	r_i	*N*	r_i	*N*	r_i	*N*	r_i
SV01	SV	173	.274	183	.294	182	.314	182	.369	179	.323
VPG17	INT	158	.451	158	.548	160	.564	162	.543	158	.542
VPG18	INT	158	.395	158	.417	160	.395	162	.432	158	.447
VPG19	INT	158	.317	158	.397	160	.462	162	.370	158	.408

Forschungsfrage 3: Wie korrelieren die Werte der fünf Messzeitpunkte der vier Skalen unter Berücksichtigung der Schulform miteinander?

Zur Beantwortung der dritten Forschungsfrage wird pro Skala die Korrelation zwischen dem Mittelwert des ersten Messzeitpunktes und den Mittelwerten der weiteren Messzeitpunkte nach Pearson berechnet. Tabelle 9 zeigt die Korrelationswerte in der Übersicht.

Tabelle 9: Korrelationen des ersten Messzeitpunktes mit den weiteren Messzeitpunkten pro Skala und Schulform bzw. schulformübergreifend

Skala	Schulform		MZ1	MZ2	MZ3	MZ4	MZ5
SV	schulform-übergreifend	Korrelation nach Pearson	1	,812**	,723**	,715**	,679**
		N	173	162	162	161	155
	Grundschule	Korrelation nach Pearson	1	,867**	,790**	,772**	,751**
		N	97	91	95	92	87
	Gesamtschule	Korrelation nach Pearson	1	,719**	,590**	,620**	,564**
		N	76	71	67	69	68
EXT	schulform-übergreifend	Korrelation nach Pearson	1	,866**	,840**	,771**	,766**
		N	184	176	175	175	172
	Grundschule	Korrelation nach Pearson	1	,877**	,864**	,813**	,801**
		N	102	97	97	97	97
	Gesamtschule	Korrelation nach Pearson	1	,853**	,801**	,704**	,719**
		N	82	79	78	78	75
INT	schulform-übergreifend	Korrelation nach Pearson	1	,893**	,839**	,828**	,773**
		N	158	154	151	153	148
	Grundschule	Korrelation nach Pearson	1	,889**	,827**	,798**	,759**
		N	92	89	89	89	88
	Gesamtschule	Korrelation nach Pearson	1	,908**	,860**	,868**	,791**
		N	66	65	62	64	60
PS	schulform-übergreifend	Korrelation nach Pearson	1	,815**	,830**	,803**	,777**
		N	180	165	168	164	168
	Grundschule	Korrelation nach Pearson	1	,862**	,865**	,827**	,833**
		N	101	88	92	89	96
	Gesamtschule	Korrelation nach Pearson	1	,756**	,786**	,760**	,666**
		N	79	77	76	75	72

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Die Abbildungen 1 bis 12 visualisieren den tendenziellen Verlauf der berechneten Korrelationen über die Zeit. Die Abbildungen veranschaulichen für alle Skalen und pro Schulform bzw. schulformübergreifend einen Reliabilitätsabfall der Verhaltensbeurteilungen über die fünf Messzeitpunkte.

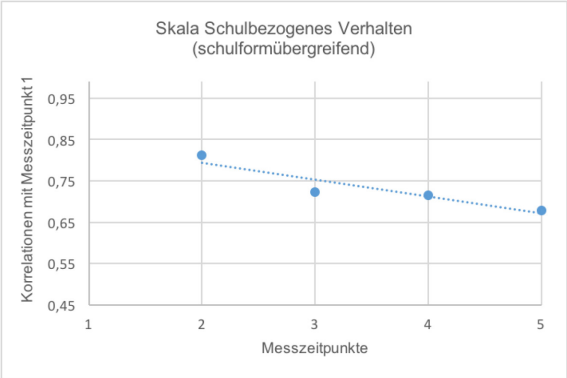

Abbildung 1: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala SV (schulformübergreifend)

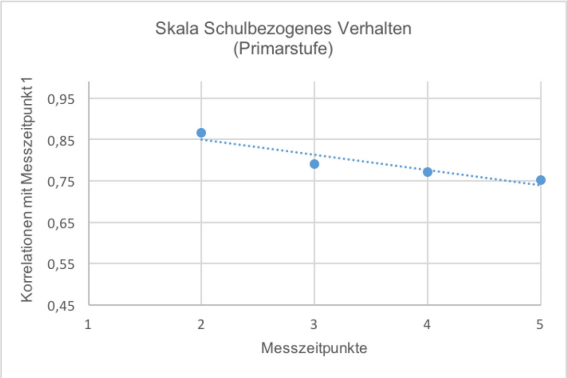

Abbildung 2: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala SV (Primarstufe)

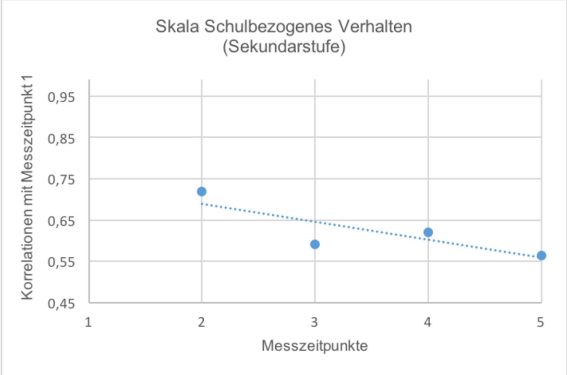

Abbildung 3: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala SV (Sekundarstufe)

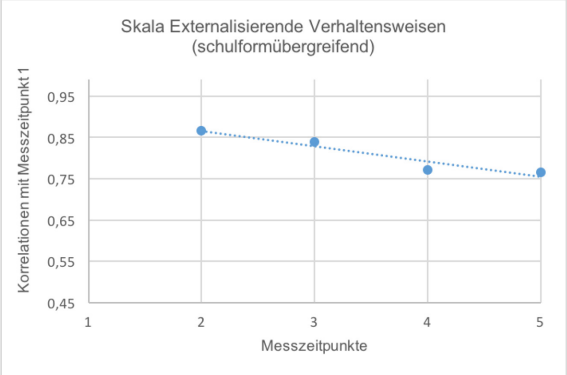

Abbildung 4: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala EXT (schulformübergreifend)

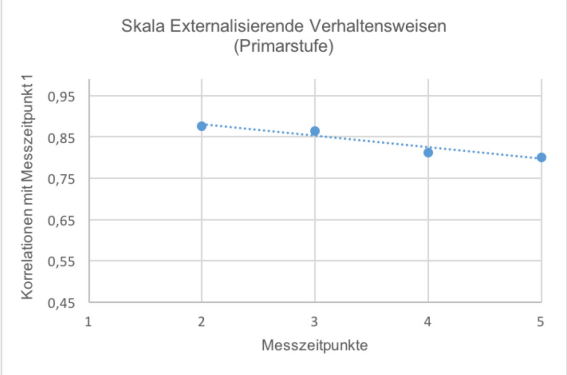

Abbildung 5: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala EXT (Primarstufe)

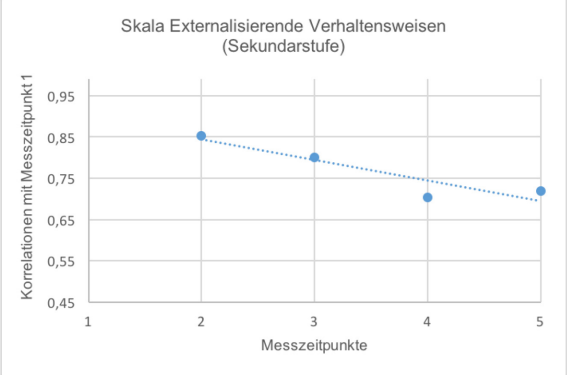

Abbildung 6: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala EXT (Sekundarstufe)

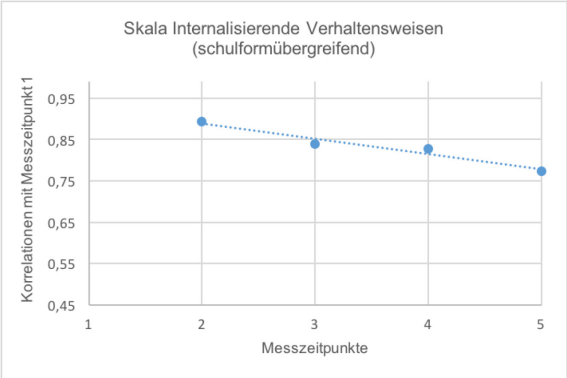

Abbildung 7: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala INT (schulformübergreifend)

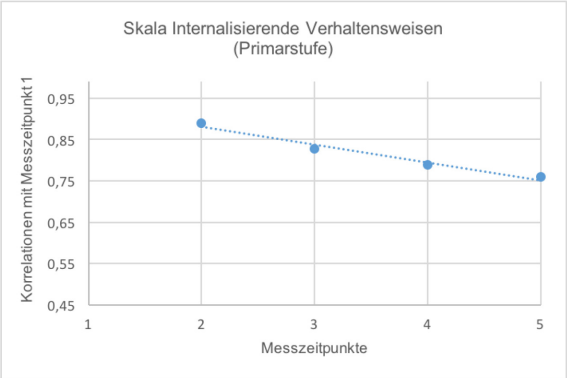

Abbildung 8: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala INT (Primarstufe)

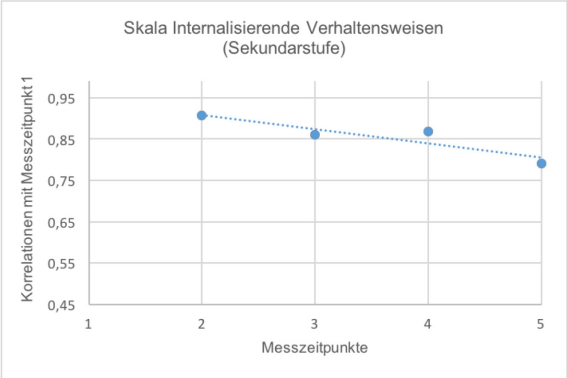

Abbildung 9: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala INT (Sekundarstufe)

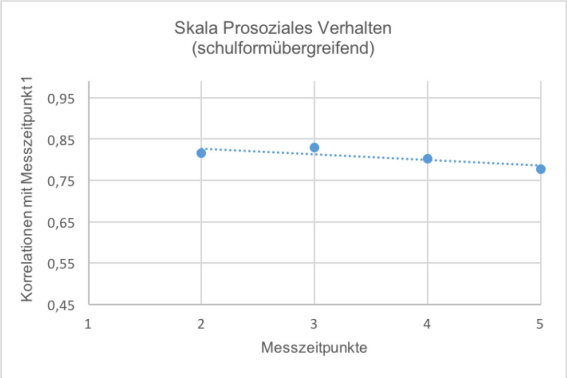

Abbildung 10: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala PS (schulformübergreifend)

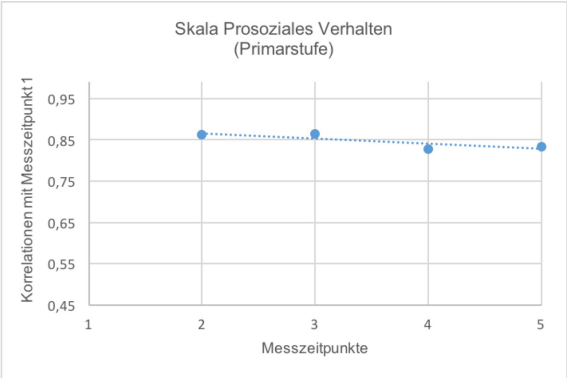

Abbildung 11: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala PS (Primarstufe)

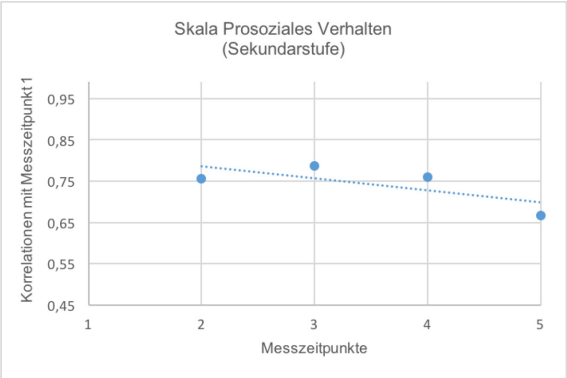

Abbildung 12: Korrelation des ersten Ratings mit den vier weiteren Ratings über die Zeit – Skala PS (Sekundarstufe)

Forschungsfrage 4: Welche Mittelwerte weisen die Grund- und GesamtschülerInnen pro Messzeitpunkt und über die Messzeitpunkte auf?

Zur Beantwortung der vierten Forschungsfrage werden nachfolgend die erhobenen Mittelwerte und Standardabweichungen der Grund- und GesamtschülerInnen pro Skala und Messzeitpunkt dargestellt. Zudem werden zur Darstellung der Verhaltensverläufe Liniendiagramme mit Trendlinien konfiguriert. Diese dienen ausschließlich der Visualisierung der Verläufe. Zur Überprüfung systematischer Unterschiede zwischen den beiden Schulformen in der Verhaltensveränderung über die Messzeitpunkte wurde pro Skala und Schulform eine Varianzanalyse durchgeführt.
Tabelle 10 führt die Mittelwerte und Standardabweichungen der beiden Schülergruppen für die vier Skalen und fünf Messzeitpunkte auf.

Tabelle 10: Skalenmittelwerte der SchülerInnen zu den fünf Messzeitpunkten pro Schulform und schulformübergreifend

Skalen	Schulform	MZ1 *M (SD)*	MZ2 *M (SD)*	MZ3 *M (SD)*	MZ4 *M (SD)*	MZ5 *M (SD)*
SV	Grundschule *N* = 80	4,15 (1,40)	4,18 (1,39)	4,28 (1,45)	4,25 (1,45)	4,27 (1,40)
	Gesamtschule *N* =61	4 (1,28)	4,01 (1,2)	4,1 (1,2)	4,37 (1,25)	4,14 (1,24)
	schulform-übergreifend *N* = 141	4,09 (1,35)	4,11 (1,31)	4,2 (1,35)	4,3 (1,36)	4,21 (1,33)
EXT	Grundschule *N* = 89	3,26 (1,58)	3,16 (1,56)	3,1 (1,58)	2,91 (1,4)	2,95 (1,45)
	Gesamtschule *N* = 72	3,15 (1,51)	3,14 (1,41)	2,97 (1,42)	2,7 (1,29)	2,77 (1,48)
	schulform-übergreifend *N* = 161	3,21 (1,55)	3,15 (1,49)	3,04 (1,5)	2,81 (1,35)	2,87 (1,46)
INT	Grundschule *N* = 81	2,61 (1,18)	2,34 (1,14)	2,34 (1,16)	2,22 (1,14)	2,21 (1,2)
	Gesamtschule *N* = 57	2,92 (1,3)	2,96 (1,26)	2,99 (1,27)	2,92 (1,3)	2,89 (1,22)
	schulform-übergreifend *N* = 138	2,74 (1,23)	2,6 (1,22)	2,61 (1,25)	2,51 (1,25)	2,49 (1,25)
PS	Grundschule *N* = 80	4,45 (1,92)	4,53 (1,79)	4,68 (1,82)	4,75 (1,79)	4,58 (1,85)
	Gesamtschule *N* = 70	4,62 (1,53)	4,25 (1,53)	4,42 (1,47)	4,81 (1,18)	4,53 (1,39)
	schulform-übergreifend *N* = 150	4,53 (1,79)	4,4 (1,68)	4,56 (1,67)	4,78 (1,53)	4,55 (1,65)

Nachfolgend wird zunächst erläutert, wie hoch die Mittelwerte der SchülerInnen schulformübergreifend ausgefallen sind. Der Tabelle 10 kann entnommen werden, dass die Lehrkräfte das schulbezogene Verhalten schulformübergreifend durchschnittlich mit dem Wert 4 beurteilten. Die externalisierenden Verhaltensweisen wurden von den Lehrkräften beider Schulformen durchschnittlich mit einem Skalenwert von 3 beurteilt. Für die Skala internalisierende Verhaltensweisen wurden im Vergleich mit den anderen Skalen schulformübergreifend die niedrigsten Mittelwerte berechnet. Die Skala Prosoziales Verhalten weist wiederum die höchsten Mittelwerte auf. Für diesen Verhaltensbereich zeigen sich die höchsten Standardabweichungen im Vergleich zu den anderen Skalen. Etwas niedrigere, aber dennoch hohe Standardabweichungen sind für die Skala externalisierende Verhaltensweisen festgestellt worden. Wiederum niedrigere Standardabweichungen pro Messzeitpunkt finden sich bei der Skala schulbezogenes Verhalten. Die Werte der Skala internalisierende Verhaltensweisen streuen am geringsten, jedoch weiterhin in einem Bereich von ungefähr einem Skalenwert.

Die Abbildungen 13 bis 24 dienen der Visualisierung der Verhaltensveränderungen der SchülerInnen pro Schulform und Skala.

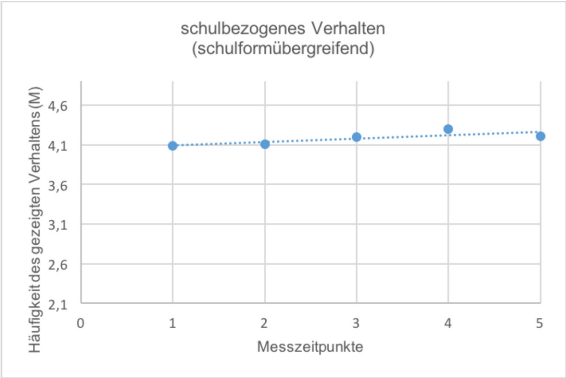

Abbildung 13: Verhaltensveränderungen – Skala SV (schulformübergreifend)

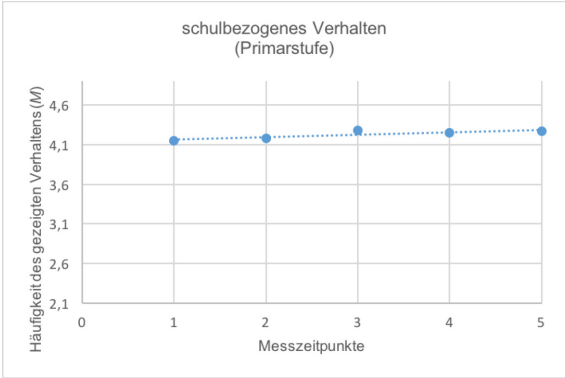

Abbildung 14: Verhaltensveränderungen – Skala SV (Primarstufe)

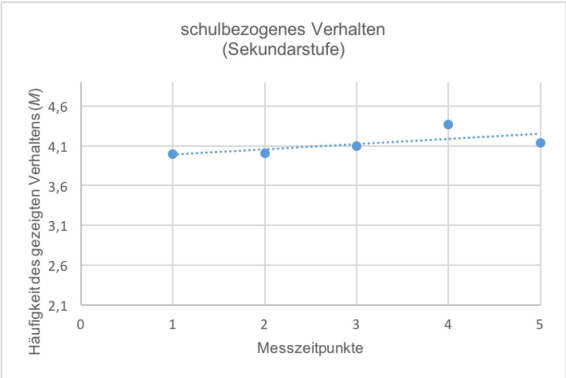

Abbildung 15: Verhaltensveränderungen – Skala SV (Sekundarstufe)

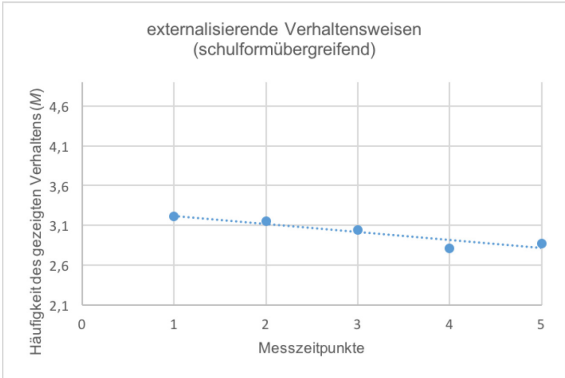

Abbildung 16: Verhaltensveränderungen – Skala EXT (schulformübergreifend)

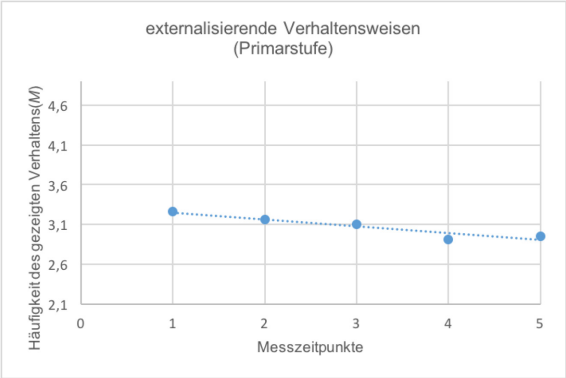

Abbildung 17: Verhaltensveränderungen – Skala EXT (Primarstufe)

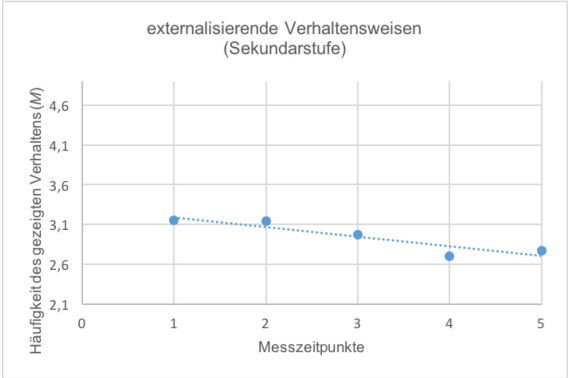

Abbildung 18: Verhaltensveränderungen – Skala EXT (Sekundarstufe)

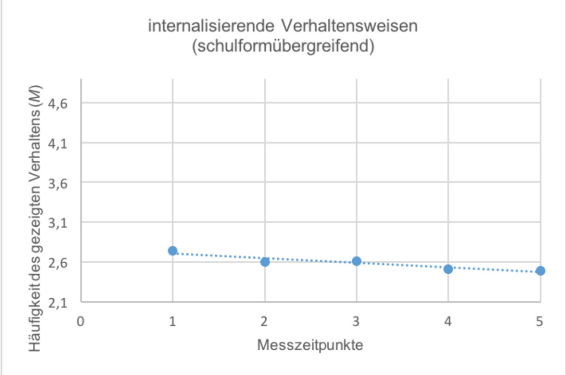

Abbildung 19: Verhaltensveränderungen – Skala INT (schulformübergreifend)

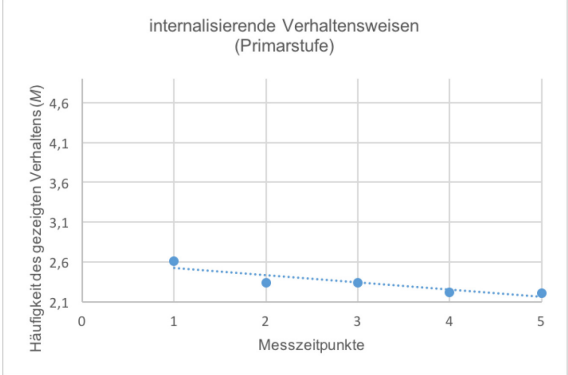

Abbildung 20: Verhaltensveränderungen – Skala INT (Primarstufe)

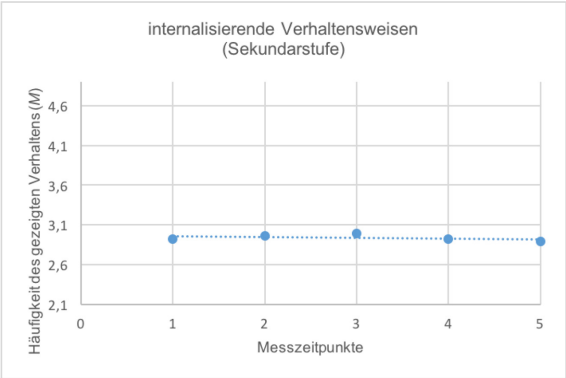

Abbildung 21: Verhaltensveränderungen – Skala INT (Sekundarstufe)

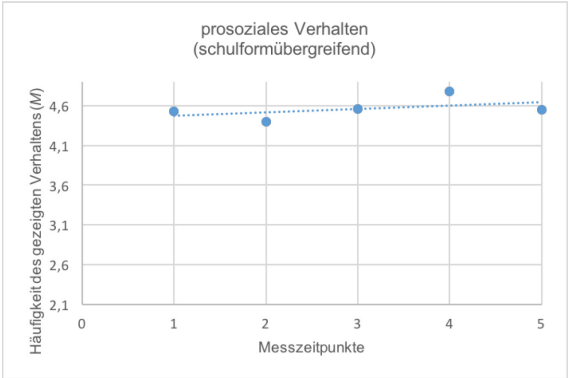

Abbildung 22: Verhaltensveränderungen – Skala PS (schulformübergreifend)

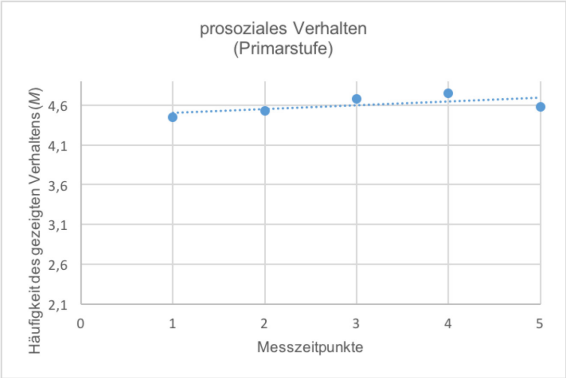

Abbildung 23: Verhaltensveränderungen – Skala PS (Primarstufe)

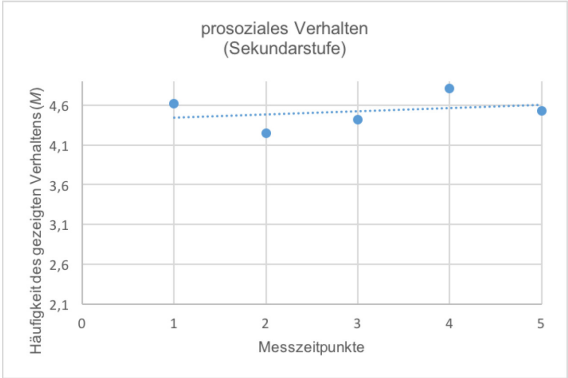

Abbildung 24: Verhaltensveränderungen – Skala PS (Sekundarstufe

Ein Vergleich des schulbezogenen Verhaltens zwischen den Schulformen verdeutlicht, dass sich die Mittelwerte der GrundschülerInnen von den Mittelwerten der GesamtschülerInnen nur im geringen Maße unterscheiden. Zum vierten Messzeitpunkt weisen die GesamtschülerInnen höhere Mittelwerte auf. Zu den anderen vier Messzeitpunkten sind die Mittelwerte der GrundschülerInnen höher. Die Standardabweichungen zeigen, dass zu allen Messzeitpunkten die Werte der GrundschülerInnen im größeren Maße streuen. Durchschnittlich wurde das Verhalten beider Schülergruppen mit einem Skalenwert von 4 beurteilt. Die Trendlinien veranschaulichen, dass die Lehrkräfte schulbezogenes Verhalten sowohl bei den GrundschülerInnen als auch bei den GesamtschülerInnen mit zunehmender Anzahl der Messungen im Durchschnitt tendenziell häufiger beobachteten (vgl. Abbildung 14 & 15). Eine ANOVA mit Messwiederholung mit Huynh-Feldt-Korrektur konnte keinen signifikanten Unterschied zwischen den beiden Schülergruppen im Verhaltensverlauf feststellen (F[3,76, 522,371] = 1.065, p = .371]). Hinsichtlich des schulbezogenen Verhaltens haben sich die beiden Schülergruppen nicht systematisch unterschiedlich verändert.

Für die Skala Externalisierende Verhaltensweisen zeigen sich durchweg geringe Unterschiede in den Mittelwerten der beiden Schülergruppen. Dabei sind die Mittelwerte der GrundschülerInnen zu allen fünf Messzeitpunkten höher und die Standardabweichungen der GrundschülerInnen zu vier Messzeitpunkten größer. Durchschnittlich haben Lehrkräfte der Grundschule und der Gesamtschule das Verhalten ihrer SchülerInnen mit einem Skalenwert von 3 beurteilt. Die Trendlinien der Liniendiagramme 17 und 18 veranschaulichen, dass sowohl die GrundschülerInnen als auch die GesamtschülerInnen im Verlauf der Zeit tendenziell seltener externalisierende Verhaltensweisen zeigten. Eine ANOVA mit Messwiederholung mit Huynh-Feldt-Korrektur bestätigt, dass sich das Auftreten der externalisierenden Verhaltensweisen der beiden Schülergruppen nicht systematisch unterschiedlich verändert. Für diesen Verhaltensbereich besteht kein signifikanter Unterschied in der Verhaltensveränderung der beiden Schülergruppen (F[3,442, 547,203] = .544, p = .676).

Größere Unterschiede zwischen den Mittelwerten der Schulformen sind im Bereich internalisierender Verhaltensveränderungen erkennbar. Zu allen Messzeitpunkten zeigten GesamtschülerInnen häufiger internalisierende Verhaltensweisen als GrundschülerInnen bei gleichzeitig höheren Standardabweichungen. Grundschullehrkräfte beurteilten das Verhalten ihrer SchülerInnen überwiegend mit einem durchschnittlichen Skalenwert von 2. Gesamtschullehrkräfte beurteilten die internalisierenden Verhaltensweisen ihrer SchülerInnen hingegen zu allen Messzeitpunkten mit einem durchschnittlichen Skalenwert von 3. Die Trendlinie der GrundschülerInnen sinkt tendenziell (vgl. Abbildung 20). Die Trendlinien der GesamtschülerInnen zeigt einen kaum erkennbaren Abfall der durchschnittlichen Werte (vgl. Abb 21). Die internalisierenden Verhaltensweisen beider Schülergruppen verändern sich über die Zeit unterschiedlich. Die Berechnung einer ANOVA mit Messwiederholung mit Huynh-Feldt-Korrektur stützt diese Werte. Sie zeigt, dass sich die Veränderungen internalisierender Verhaltensweisen über die Messzeitpunkte zwischen den Schülergruppen signifikant unterschieden ($F[3{,}617, 491{,}891] = 3.864$, $p = .006$).

Ein Vergleich des prosozialen Verhaltens zwischen beiden Schulformen zeigt, dass sich die Mittelwerte der GrundschülerInnen von den Mittelwerten der GesamtschülerInnen unterscheiden. Zum zweiten, dritten und fünften Messzeitpunkt sind die Mittelwerte der GrundschülerInnen größer. Zu den übrigen zwei Messzeitpunkten sind die Mittelwerte der GesamtschülerInnen größer. Durchschnittlich beurteilten LehrerInnen der Grundschule und LehrerInnen der Gesamtschule das prosoziale Verhalten der SchülerInnen mit Skalenwerten von 4 und 5. Über alle Messzeitpunkte zeigten die GrundschülerInnen höhere Standardabweichungen. Bei der Betrachtung der Mittelwerte der GesamtschülerInnen über die Zeit zeigen sich große Unterschiede zwischen den Messzeitpunkten. Tendenziell zeigten die SchülerInnen häufiger prosoziale Verhaltensweisen, wie die Trendlinie im Liniendiagramm 24 veranschaulicht. Die Mittelwerte der GrundschülerInnen unterscheiden sich über die Messzeitpunkte im geringeren Maße. Auch diese Schülergruppe zeigt im Durchschnitt tendenziell häufiger prosoziale Verhaltensweisen (vgl. Abbildung 23). Eine ANOVA mit Messwiederholung mit Huynh-Feldt-Korrektur konnte aufzeigen, dass sich im Vergleich das prosoziale Verhalten der Schülergruppen signifikant unterschiedlich veränderte ($F[3{,}755, 555{,}706] = 2.969$, $p = .022$).

Forschungsfrage 5: Welche Variationsbreite weisen die Beurteilungen der Lehrkräfte der Grund- und Gesamtschulen auf?

Die nachfolgenden Tabellen 11 bis 15 zeigen pro Item die durchschnittliche Variationsbreite zwischen dem maximalen und minimalen Wert einer Verhaltensbeurteilung pro Item. Dazu wurde zunächst pro Fall und Item die Variationsbreite zwischen maximaler und minimaler Verhaltensbeurteilung über die fünf Messzeitpunkte bestimmt. Anschließend wurden pro Item das arithmetische Mittel und die Standardabweichung der Variationsbreiten schulformübergreifend und pro Schulform berechnet. In den Tabellen finden sich außerdem Angaben zu den maximalen und minimalen Variationsbreiten, sowie drei Perzentilangaben (25., 50. und 75. Perzentil). Fehlende Werte führten auch bei diesen Berechnungen zu einem listenweisen Fallausschluss. In diesem Zusammenhang beziehen sich die Angaben auf Fälle, für die zu allen fünf Messzeitpunkten Ratings durchgeführt wurden.

In der Tabelle 11 werden die Werte der Variationsbreite pro Item und schulformübergreifend dargestellt.

Tabelle 11: Variationsbreite der Verhaltensbeurteilungen pro Item und schulformübergreifend

Item	*M (SD)*	*Min*	*Perzentile*	*Max*

			25	*50*	*75*	
SV01 *N* = 151	2,1 (1,22)	0	1	2	3	5
SV02 *N* = 154	1,88 (1,29)	0	1	2	3	5
SV03 *N* = 169	1,99 (1,15)	0	1	2	3	5
SV04 *N* = 168	2,11 (1,21)	0	1	2	3	5
VP05 *N* = 176	1,61 (1,52)	0	0	1	3	6
VP06 *N* = 179	1,72 (1,33)	0	1	1	3	5
VP07 *N* = 174	1,64 (1,47)	0	0	1	3	6
HY08 *N* = 180	1,65 (1,36)	0	1	1	3	5
HY09 *N* = 176	2,06 (1,46)	0	1	2	3	6
HY10 *N* = 176	2,09 (1,36)	0	1	2	3	6
EP11 *N* = 169	1,68 (1,51)	0	0	1	3	6
EP12 *N* = 170	1,16 (1,25)	0	0	1	2	5
EP13 *N* = 170	1,57 (1,53)	0	0	1	2,25	6
PS14 *N* = 166	1,72 (1,26)	0	1	1	2	6
PS15 *N* = 163	1,69 (1,3)	0	1	1	2	6
PS16 *N* = 153	1,67 (1,19)	0	1	2	2	6
VPG17 *N* = 147	1,77 (1,5)	0	1	2	3	6
VPG18 *N* = 150	1,38 (1,34)	0	0	1	2	5
VPG19 *N* = 146	1,21 (1,49)	0	0	1	2	6

Die Tabelle 11 veranschaulicht für die Items insgesamt eine durchschnittliche Variationsbreite zwischen 1,16 (EP11) und 2,11 (SV04). Bei allen Items konnte als kleinste Differenz ein Wert von 0 ermittelt werden. Für die maximale Variationsbreite konnte bei allen Items eine Differenz von 5 bzw. 6 festgestellt werden. In 25% der Fälle sind die Differenzen kleiner als 0 bzw. 1 und in 75% der Fälle kleiner als 2 bzw. 3. Auffällig sind die Werte der Items EP12 und VPG18. Diese weisen einen Maximalwert von 5 auf. Zudem zeigen sich Variationsbreiten, die in 25%

der Fälle einer Differenz von 0, in 50% der Fälle einer Differenz kleiner als 1 und in 75% der Fälle einer Differenz kleiner als 2 entsprechen.
In der Tabelle 12 werden die Werte der Skala Schulbezogenes Verhalten (SV) unter Berücksichtigung der Variable Schulform dargestellt.

Tabelle 12: Variationsbreite der Verhaltensbeurteilungen der Skala SV pro Schulform

Item	Schulform	*M (SD)*	*Min*	*Perzentile*			*Max*
				25.	*50.*	*75.*	
SV01	Grundschule *N* = 84	1,93 (1,2)	0	1	2	2	5
	Gesamtschule *N* = 67	2,31 (1,21)	0	1	2	3	5
SV02	Grundschule *N* = 88	1,63 (1,32)	0	1	1	3	5
	Gesamtschule *N* = 66	2,23 (1,17)	0	1	2	3	5
SV03	Grundschule *N* = 99	1,87 (1,24)	0	1	2	3	5
	Gesamtschule *N* = 70	2,17 (0,99)	0	1	2	3	5
SV04	Grundschule *N* = 97	1,93 (1,21)	0	1	2	3	5
	Gesamtschule *N* = 71	2,37 (1,17)	0	2	2	3	5

Beim Vergleich der Werte der Grundschullehrkräfte und der Gesamtschullehrkräfte, fällt auf, dass sich über alle vier Items der Skala SV durchschnittlich kleinere Variationsbreiten bei den Grundschullehrkräften als bei den Gesamtschullehrkräften zeigen.
Tabelle 13 veranschaulicht die Variationsbreiten pro Schulform für die sechs Items der Skala Externalisierende Verhaltensweisen.

Tabelle 13: Variationsbreite der Verhaltensbeurteilungen der Skala EXT pro Schulform

Item	Schulform	*M (SD)*	*Min*	*Perzentile*			*Max*
				25.	*50.*	*75.*	
VP05	Grundschule *N* = 100	1,56 (1,6)	0	0	1	3	6
	Gesamtschule *N* = 76	1,68 (1,43)	0	1	1	2,75	5
VP06	Grundschule *N* = 99	1,75 (1,35)	0	1	1	3	5
	Gesamtschule *N* = 80	1,68 (1,3)	0	1	1	2	5
VP07	Grundschule *N* = 95	1,58 (1,46)	0	0	1	3	6
	Gesamtschule *N* = 79	1,72 (1,48)	0	0	2	2	6
HY08	Grundschule *N* = 101	1,62 (1,38)	0	0	2	2,5	5
	Gesamtschule *N* = 79	1,68 (1,34)	0	1	1	3	5
HY09	Grundschule *N* = 97	1,97 (1,55)	0	1	2	3	6
	Gesamtschule *N* = 79	2,18 (1,33)	0	1	2	3	6
HY10	Grundschule *N* = 97	1,94 (1,4)	0	1	2	3	6
	Gesamtschule *N* = 71	2,28 (1,29)	0	1	2	3	5

Ein Vergleich der Schulformen zeigt in der Skala EXT für fünf Items (VP05, VP07, HY08, HY09, HY10) durchschnittlich kleinere Differenzen bei Grundschullehrkräften als bei den Gesamtschullehrkräften. Beim sechsten Item (VP06) der Skala sind hingegen durchschnittlich kleinere Differenzen bei den Gesamtschullehrkräften als bei den Grundschullehrkräften erkennbar.
Tabelle 14 zeigt die Werte der Items der Skala Internalisierende Verhaltensweisen pro Schulform.

Tabelle 14: Variationsbreite der Verhaltensbeurteilungen der Skala INT pro Schulform

Item	Schulform	M (SD)	*Min*	*Perzentile*			*Max*
				25.	*50.*	*75.*	
EP11	Grundschule *N* = 93	1,75 (1,55)	0	0	2	3	6
	Gesamtschule *N* = 76	1,59 (1,46)	0	1	1	2	6
EP12	Grundschule *N* = 93	1,02 (1,27)	0	0	1	2	5
	Gesamtschule *N* = 77	1,33 (1,22)	0	0	1	2	5
EP13	Grundschule *N* = 93	1,55 (1,62)	0	0	1	3	6
	Gesamtschule *N* = 77	1,58 (1,43)	0	1	1	2	6
VPG17	Grundschule *N* = 89	1,57 (1,49)	0	0	1	3	6
	Gesamtschule *N* = 58	2,07 (1,46)	0	1	2	3	6
VPG18	Grundschule *N* = 90	1,28 (1,35)	0	0	1	2	5
	Gesamtschule *N* = 60	1,53 (1,32)	0	0,25	1	2	5
VPG19	Grundschule *N* = 86	0,95 (1,35)	0	0	0	1	5
	Gesamtschule *N* = 60	1,58 (1,6)	0	0	1	3	

Bei der Skala INT zeigen sich bei fünf von sechs Items (EP12, EP13, VP17, VPG18, VPG19) durchschnittlich kleinere Variationsbreiten als bei den Gesamtschullehrkräften. Eine Ausnahme stellt in dieser Skala das Item EP11 dar, bei welchen die Gesamtschullehrkräfte durchschnittlich kleinere Differenzen aufweisen als die Grundschullehrkräfte.
Tabelle 15 verdeutlicht die Werte der Skala Prosoziales Verhalten unter Berücksichtigung der Schulformen.

Tabelle 15: Variationsbreite der Verhaltensbeurteilungen der Skala PS pro Schulform

Item	Schulform	*M (SD)*	*Min*	*Perzentile*			*Max*
				25.	*50.*	*75.*	
PS14	Grundschule *N* = 92	1,55 (1,2)	0	1	1	2	6
	Gesamtschule *N* = 74	1,92 (1,31)	0	1	2	3	6
PS15	Grundschule *N* = 89	1,37 (1,18)	0	0,5	1	2	6
	Gesamtschule *N* = 74	2,07 (1,35)	0	1	2	3	6
PS16	Grundschule *N* = 81	1,57 (1,17)	0	1	2	3	6
	Gesamtschule *N* = 72	1,79 (1,21)	0	1	1	2	5

Ein Vergleich der Schulformen zeigt, dass die Grundschullehrkräfte ebenso bei den drei Items der Skala PS durchschnittlich kleinere Variationsbreiten aufweisen als die Gesamtschullehrkräfte.

Insgesamt zeigen die Grundschullehrkräfte bei 17 von 19 Items eine geringere durchschnittliche Variationsbreite als die Gesamtschullehrkräfte. Lediglich bei den beiden Items VP06 und EP11 fällt die durchschnittliche Variationsbreite der Gesamtschullehrkräfte geringer aus.

6.5. Diskussion

Forschungsfrage 1: Wie setzen Lehrkräfte an Grund- und Gesamtschulen die Ratingskala ein? Handelt es sich bei der Ratingskala um ein ökonomisches und nützliches Testinstrument?

Im Sinne einer Feldstudie wurde die Studie im natürlichen schulischen Setting durchgeführt (Döring & Bortz, 2016). Dadurch stimmen die Untersuchungsbedingungen mit den Alltagsbedingungen, in denen die Ratingskala zukünftig eingesetzt werden soll, weitestgehend überein und es lassen sich Ableitungen für den zukünftigen Einsatz der Ratingskala im inklusiven Handlungsfeld der Grund- und Gesamtschule tätigen (ebd.).

Die im Rahmen der ersten Forschungsfrage erhobenen Daten liefern Hinweise auf vorgehensspezifische Aspekte, wie den Erhebungszeitraum und die Beobachtungszeiträume. Diese vorgehensspezifischen Aspekte bedingen die Qualität eines Testverfahrens (Casale et al., 2015a). Sie werden nachfolgend diskutiert. Die erhobenen Daten liefern zudem Hinweise auf die Situations- bzw. Messbedingungen unter denen die Ratingskala eingesetzt wurden. Diese gilt es zu erfassen, da der schulische Kontext, in dem SchülerInnen mit Verhaltensschwierigkeiten unterrichtet werden, eine komplexe Umgebung darstellt und das Auftreten von Verhaltensweisen von unterschiedlichsten Rahmenbedingungen beeinflusst wird (Conroy et al.,

2008; Huber & Rietz, 2015). Eine Untersuchung der Messbedingungen ermöglicht es zu bestimmen, unter welchen Bedingungen eine zuverlässige Erfassung von Verhaltensverläufe möglich ist (Casale et al., 2017). Eine Erhebung der Situationsbedingungen stellt in diesem Zusammenhang eine große Herausforderung dar, da nicht alle Faktoren kontrolliert werden können (Conroy et al., 2008). Dennoch müssen Bedingungsfaktoren und potentiell kritische Variablen, wenn möglich berücksichtigt und identifiziert werden (Casale et al., 2015a; Conroy et al., 2008). Da Befunde zur „'Schwierigkeit' der diagnostischen Situation" (Huber & Rietz, 2015, S. 94) und zum Einsatz von DBR-Verfahren im inklusiven Setting fehlen (Chafouleas et al., 2010), will die vorliegende Pilotierungsstudie einen Beitrag zu diesem Forschungsschwerpunkten leisten. Die Situationsbedingungen, unter denen die Ratingskala in dieser Studie eingesetzt wurde, werden nachfolgend dargestellt.

Die Lehrkräfte wurden gebeten nach jeder Verhaltensbeurteilung Angaben zum Beobachtungssetting, zur Beobachtungssituation und zum Beobachtungszeitraum zu notieren. Auf diese Weise konnten umfangreiche Daten zum situativen Kontext, in dem die Ratingskala eingesetzt wurde, gesammelt werden. Die Daten wurden mittels Häufigkeitsanalysen ausgewertet.

Die Ratingskala sollten von den Lehrkräften an aufeinanderfolgenden Tagen eingesetzt werden. Dieses hochfrequente Vorgehen, eignet sich, um in kurzen Zeiträumen eine hohe Anzahl von Messwerten zu erheben und folglich die Verhaltensentwicklungen der SchülerInnen im Verlauf abzubilden (Casale et al., 2017; Huber & Rietz, 2015). Zur Berechnung des Erhebungszeitraums wurden Schultage von Montag bis Freitag berücksichtigt. Keine Beachtung fanden Wochenenden, Ferientage und Feiertage. Hinsichtlich des Erhebungszeitraums zeigt sich, dass die Grundschullehrkräfte fünf aufeinanderfolgende Beurteilungen in weniger Schultagen durchführen konnten als die Gesamtschullehrkräfte. Durchschnittlich benötigten die Grundschullehrkräfte sieben Schultage für fünf Einsätze der Ratingskala. Gesamtschullehrkräfte benötigten hingegen durchschnittlich 12 Schultage. Dies entspricht bei den Grundschullehrkräften einem Zeitraum von etwas mehr als einer Woche und bei den Gesamtschullehrkräften einem Zeitraum von etwas mehr als zwei Wochen. 76,6% aller Grundschullehrkräfte war es sogar möglich fünf Beurteilungen an fünf unmittelbar aufeinanderfolgenden Schultagen einzusetzen. Unter den Gesamtschullehrkräften war dies nur 22,5% der Lehrkräfte möglich. Grundschullehrkräften liegt folglich bei einem einmaligen täglichen Einsatz zu einem früheren Zeitpunkt eine vergleichbare Anzahl von Datenpunkten vor. Infolgedessen können pädagogisch relevante Entscheidungen zur Förderung der SchülerInnen zu einem früheren Zeitpunkt getroffen werden. Grundsätzlich besteht die Möglichkeit Verhaltensbeurteilungen auch mehrmals täglich durchzuführen. Bei einem mehrmals täglichen Einsatz der Ratingskala sind zuverlässige Interpretationen der Messwerte zu einem früheren Zeitpunkt möglich. Chafouleas et al. (2007) konnten beispielsweise aufzeigen, dass ein Reliabilitätskoeffizient von .70 sowohl

bei je einem Rating an sieben Tagen als auch bei je zwei Ratings an vier Tagen erreicht werden kann. Grundsätzlich ist die Anzahl der Ratings wichtiger als die Anzahl der Beobachtungstage. Unter Berücksichtigung der Forschungsarbeiten von Riley-Tillman et al. (2011) und Volpe und Briesch (2012), die herausstellen konnten, dass die Bildung von Mittelwerten über mehrere Messungen zu höheren Reliabilitätswerten und damit zu zuverlässigeren Messungen und stabileren Werten führen, erscheint es sinnvoll der Anregung von Casale et al. (2017) zu folgen und einen mehrmals täglichen Einsatz der Ratingskala in beiden Schulformen zu empfehlen, um zeitnah und zugleich zuverlässig pädagogisch relevante Entscheidungen treffen zu können (Casale et al., 2017).

Die Ergebnisse zum Beobachtungszeitraum, auf den sich die Verhaltensbeurteilungen beziehen, zeigen, dass Grundschullehrkräfte häufiger als Gesamtschullehrkräfte längere Zeiträume beobachteten. Am häufigsten wählten Grundschullehrkräfte den Zeitraum von einem Schultag (65,2%). Diesen wählten Gesamtschullehrkräfte in keinem der Fälle. Sie beobachteten am häufigsten den Zeitraum einer Schulstunde (82,9%). Diese Verteilung ist darauf zurückzuführen, dass die SchülerInnen der Grundschule überwiegend nach dem Klassenlehrerprinzip unterrichtet werden, damit sie sich auf wenige Bezugspersonen fokussieren können, wohingegen die SekundarstufenschülerInnen überwiegend nach dem Fachlehrerprinzip und dementsprechend von einer Vielzahl von LehrerInnen unterrichtet werden (KMK, 2017). Dies impliziert, dass die Lehrkräfte der Gesamtschule insgesamt weniger Stunden pro Woche und die Grundschullehrkräfte insgesamt mehr Stunden pro Woche in einer Klasse unterrichten. Für Grundschullehrkräfte bieten sich insgesamt längere Beobachtungszeiträume zur Beurteilung des Verhaltens einzelner SchülerInnen. Eine gemeinsame Betrachtung der Erhebungszeiträume und der Beobachtungszeiträume, lässt darauf schließen, dass es Grundschullehrkräften möglich ist, häufiger an einem Tag und auch häufiger innerhalb einer Woche Ratings durchzuführen und somit früher zuverlässige Entscheidungen von pädagogischer Relevanz treffen zu können. An Gesamtschulen könnte dieser vermeintliche Nachteil durch eine angemessene Kooperation und Kommunikation zwischen verschiedenen Lehrkräften und pädagogischen Fachkräften ausgeglichen werden. Grundsätzlich sollte beachtet werden, dass Verhaltensbeurteilungen subjektiv geprägt sind und folglich mit numerisch varianten Verhaltensbeurteilungen zu rechnen ist (Casale, 2017; Schmidt-Atzert et al., 2012). Werden die Verhaltensbeurteilungen jedoch nicht hinsichtlich der numerischen Beurteilungen verglichen, sondern hinsichtlich der Trendentwicklungen, so sind auch parallele Beurteilungen interpretierbar. Werden die Verhaltenstrends identisch beurteilt, liegt eine strukturelle Invarianz vor (Casale, 2017). Forschungsarbeiten konnten diesbezüglich zeigen, dass die Genauigkeit und Zuverlässigkeit von Verhaltensbeurteilungen erhöht werden kann, wenn parallele Beurteilungen von Schülerverhalten in derselben Situation und zum selben Zeitpunkt von mehreren Lehrkräften bzw. pädagogischen Fachkräften durchgeführt werden (Casale et al., 2015c). Akzeptable Werte konnten

bereits ab einer Anzahl von zwei parallel beurteilenden Personen festgestellt werden (ebd.). Folglich kann bei entsprechenden personellen Ressourcen ein paralleler Einsatz der Ratingskala von kooperierenden Lehrkräften zu zuverlässigeren und genaueren Ergebnissen in der Verhaltensbeurteilung und somit zu einer verbesserten Passung der individuellen Fördermaßnahmen zu den individuellen Bedürfnissen führen. Generell gilt, dass alle Ratings eine bedeutende Datengrundlage für wichtige Kommunikationen zwischen Lehrkräften und KollegInnen, Eltern oder SchülerInnen darstellen (Chafouleas et al., 2009a). Ob dies auch für die vorliegende Ratingskala gilt müssen aufgrund der Flexibilität der DBR Methode weitere Forschungsarbeiten zeigen.

Eine Auswertung hinsichtlich der Kontinuität der Beobachtungszeiträume zeigt, dass die Grundschullehrkräfte durchweg denselben Zeitraum zur Beobachtung wählten. Die Gesamtschullehrkräfte beobachteten hingegen nur in 87% der Fälle denselben Zeitraum. Es zeigt sich eine größere Varianz in der Länge der Beobachtungen bei den Gesamtschullehrkräften.

Beobachteten die Lehrkräfte einen Zeitraum von einer Schulstunde, so wurde zudem erfragt, im Rahmen welches Unterrichtsfaches das Verhalten beobachtet wurde. Die Anzahl aller beobachteten Schulfächer ist mit 16 bei den Gesamtschullehrkräften höher als bei den Grundschullehrkräften, welche das Verhalten in sieben verschiedenen Fächern beobachteten. Ein Vergleich über die verschiedenen Messzeitpunkte zeigt, dass ca. ¾ aller Gesamtschullehrkräfte (74,2%) und auch ca. ¾ aller Grundschullehrkräfte (76,7%) das Verhalten der SchülerInnen in mehreren Schulfächern und somit vor dem Hintergrund zusätzlich variierender Situationsbedingungen beobachteten. Es liegen keine bedeutenden Unterschiede zwischen den Schulformen vor, jedoch sollte der jeweilige prozentuale Anteil der Lehrkräfte, die eine Schulstunde beobachteten, beachtet werden. Bei den Gesamtschullehrkräfte sind es 82,9% der Lehrkräfte, die einen Beobachtungszeitraum von einer Schulstunde wählten, bei den Grundschullehrkräften hingegen nur 30,4%. Dies lässt auf eine größere Varianz der Beobachtungssituationen im Setting der Gesamtschule schließen. Eine Auswertung der Beobachtungssituationen zeigt, dass Grundschullehrkräfte (89,5%) ihre Beobachtungen häufiger auf Situationen im Klassenverband bezogen als es bei den Gesamtschullehrkräften der Fall war (77,1%). 80,4% der Grundschullehrkräfte beobachteten dabei über alle Messzeitpunkte dieselbe Beobachtungssituation, ebenso wie 93,8% der Gesamtschullehrkräfte. Hier zeigt sich eine größere Varianz in der Beobachtungssituation bei den Grundschullehrkräften. Unterschiedliche Beobachtungssettings können dabei zu Unterschieden im Verhalten über die Zeit führen (Chafouleas et al., 2010). Als weitere Beobachtungssituationen wurde sowohl in der Primar- als auch in der Sekundarstufe das Setting der Kleingruppe oder des Einzelunterrichts angegeben. Diese Settings basieren auf Konzepten der äußeren Differenzierung und fallen in den Bereich der sonderpädagogischen Förderung (Koch & Textor, 2015, S. 119). Im Sinne eines inklusiven

Unterrichts gelten sie als wenig sinnvoll da das Ziel schulischer Inklusion ein Abbau exkludierender Bedingungen und Prozesse ist (Hennemann et al., 2018; Koch & Textor, 2015). Da es sich bei dem Begriff der Inklusion um einen dimensionalen Begriff handelt, kann ein inklusives Setting u.a. für verschiedene Schulfächer oder unter Berücksichtigung der besonderen pädagogischen Bedürfnisse der SchülerInnen unterschiedlich gestaltet werden (Gebhardt et al., 2014). Methoden der äußeren Differenzierung können dann mit den Grundgedanken und Zielen der Inklusion in Einklang gebracht werden, wenn sie flexibel eingesetzt werden und potentiell von allen SchülerInnen der Klasse genutzt werden können (Koch & Textor, 2015). Ein Einsatz von Methoden der äußeren Differenzierung sollte nicht zu einer Herabsetzung einzelner SchülerInnen führen (ebd.). Ein gemeinsamer Unterricht von SchülerInnen mit und ohne Verhaltensschwierigkeiten im Klassenverband, wie er im Sinne der Inklusion angestrebt wird, scheint auf Basis der Ergebnisse dieser Studie in Grundschulen häufiger gegeben als in Gesamtschulen. Diese Ergebnisse werden durch die Forschungsliteratur gestützt. So konnte aufgezeigt werden, dass der Anteil inklusiv beschulter SchülerInnen in Deutschland in Grundschulen (46,9%) deutlich höher ist als der Anteil inklusiv beschulter SchülerInnen in Gesamtschulen (33,4%) (Bertelsmann-Stiftung, 2015). Insbesondere für Lehrkräfte der Sekundarstufe stellt die Umsetzung einer inklusiven Bildung eine große Herausforderung dar, da sie zum einen mangelnde diagnostische Kompetenzen aufweisen, die auf eine marginale Behandlung inklusiv-didaktischer Inhalte im Lehramtsstudium zurückgeführt werden können (vgl. Kapitel 2). Zum anderen erzeugen inklusive Konzepte in der leistungsorientierten Gesamtschule, die auf eine Selektion und Homogenisierung der SchülerInnen abzielt, ein Spannungsfeld, welches ebendiese Umsetzung erschweren kann (Amrhein, 2015).

In den im Rahmen dieser Studie durchgeführten Interviews sollte ermittelt werden, ob der Einsatz der Ratingskala einen sinnvollen Beitrag zur inklusiven Beschulung von SchülerInnen mit Verhaltensschwierigkeiten leisten kann. Die vier Interviews wurden mit je zwei sonderpädagogischen und zwei Regelschullehrkräften der Grund- und Gesamtschulen durchgeführt. Sie liefern basierend auf der Fachexpertise der Lehrkräfte umfangreiche und komplexe Informationen zum Einsatz der Ratingskala im jeweiligen schulischen Handlungsfeld.

Die Lehrkräfte der Sekundarstufe schildern, dass ein Einsatz der Ratingskala problemlos möglich ist, wenn eine angemessene Schülerzahl beobachtet werden soll. Im Interview gab der Sonderpädagoge der Gesamtschule an, dass er Verhaltensbeurteilungen für eine Anzahl von fünf SchülerInnen für möglich hält, jedoch eine Anzahl von zwei bis drei SchülerInnen bevorzugen würde. Die Regelschullehrerin der Grundschule hingegen gab an, dass sie eine Anzahl von fünf SchülerInnen für zu umfangreich erachtet und die Beurteilung des Verhaltens eines Schülers bzw. einer Schülerin für umsetzbar hält. Grundsätzlich ist es üblich nicht die ganze Klasse, sondern lediglich eine ausgewählte Anzahl von SchülerInnen zu beobachten (Cha-

fouleas et al., 2010). Wie hoch die Anzahl der ausgewählten SchülerInnen bei parallelen Beobachtungen sein sollte, ist in der Forschungsliteratur nicht eindeutig definiert. Chafouleas et al. (2010) weisen darauf hin, dass fokussierte und genaue Verhaltensbeurteilungen lediglich bis zu einer Schüleranzahl von sieben SchülerInnen möglich sind. Die Ergebnisse dieser Studie liefern Hinweise darauf, dass die Anzahl der SchülerInnen niedriger als fünf und mitunter sogar niedriger als drei sein sollte. Die Anzahl der SchülerInnen sollte in Abhängigkeit zu den weiteren Messbedingungen und den weiteren Anforderungen des Schulalltages ausgewählt werden. Die Regelschullehrkraft der Grundschule schilderte beispielsweise, dass die zusätzlichen Anforderungen des Schulalltages, wie z.B. das Unterrichten, die genaue Beurteilung des Verhaltens erschwerten. Sie äußerte, dass es ihr leichter gefallen wäre, das Verhalten der SchülerInnen zu beurteilen, wenn sie die SchülerInnen lediglich hätte beobachten können und nicht zeitgleich hätte unterrichten müssen. Generell sind im schulischen Setting kurze und schnelle Instrumente erforderlich, die „wenig personelle Ressourcen erfordern“ (Casale et al., 2015a, S. 40). Dementsprechend wäre es ungünstig, wenn ein Einsatz der Ratingskala neben einer unterrichtenden Lehrkraft einen unabhängigen Beobachter erfordern würde. Die Forschungsliteratur zeigt, dass dies nicht zwingend notwendig ist. So konnten Chafouleas et al. (2010) in ihrer Studie aufzeigen, dass Lehrkräfte ein DBR-Verfahren auf mindestens genauso reliable Weise einsetzen können, wie externe Beobachter, die mit weniger Aufgaben und Verantwortung in der jeweiligen Beobachtungssituation konfrontiert sind. Diese Ergebnisse unterstreichen die Angaben der Regelschullehrkraft der Gesamtschule, welche angab, dass sie die Ratingskala problemlos auch während des Unterrichts, z.B. in Stillarbeitsphasen, ausfüllen konnte. Sie fügte hinzu: „Und wir beobachten ja unsere Kinder sowieso. Deswegen war das angemessen.“ (vgl. Transkript Anhang 3.9). Die Beobachtungen und Beurteilungen des Verhaltens stellten für die Regelschullehrkraft der Gesamtschule im Gegensatz zur Regelschullehrkraft keine zusätzliche Anforderung dar. Die Unsicherheiten der Regelschullehrkraft der Grundschule in der Verhaltensbeurteilung werden darauf zurückgeführt, dass die Lehrkraft zuvor keine Erfahrungen im Einsatz von Ratingskalen sammeln konnte. Im Allgemeinen stellen individuelle Lehrereigenschaften, wie z.B. diagnostische Kompetenzen, Einflussfaktoren auf Verhaltensbeurteilungen dar, welche es zu beachten gilt (Chafouleas et al., 2010; Südkamp et al., 2012). In den Standards für die Lehrerbildung – Bildungswissenschaften führt die Ständige Konferenz der Kultusminister (KMK) den Kompetenzbereich des Beurteilens auf (KMK, 2004). Sie formulieren, dass Lehrkräfte dazu in der Lage sein müssen Lernvoraussetzungen und Lernprozesse kompetent zu diagnostizieren (ebd.). Weiß, Kollmannsberger und Kiel (2013) haben in ihrer Forschungsarbeit, die sich auf Gruppendiskussionen stützt, die Anforderungsprofile von Lehrkräften verschiedener Schulformen erhoben. Sie führen diagnostische Kompetenzen lediglich im Anforderungsprofil von Förderschullehrkräften, nicht jedoch im Anforderungsprofil von Primar- und Sekundarstufenlehrkräften auf. Es ist davon auszugehen,

dass Regelschullehrkräfte weniger Erfahrungen im Tätigkeitsbereich des Diagnostizierens aufweisen als sonderpädagogische Lehrkräfte (Hesse & Latzko, 2017). Im Zuge der Weiterentwicklung des deutschen Schulsystems hin zu einem inklusiven Schulsystem sind diagnostische Kompetenzen jedoch zunehmend von größerer Bedeutung und somit endgültig auch im Anforderungsprofil von Grund- und Sekundarstufenlehrkräften zu verankern (Hesse & Latzko, 2017; Linderkamp 2015). Schulformübergreifend gaben lediglich ein Drittel aller Lehrkräfte dieser Studie an, Erfahrungen in der Arbeit mit Ratingskalen zu haben. Im Vergleich verfügten anteilig mehr Grundschullehrkräfte als Gesamtschullehrkräfte über Erfahrungen in der Arbeit mit Ratingskalen (Grundschule: 34,78%; Gesamtschule: 31,25%). Dies kann u.a. darauf zurückgeführt werden, dass anteilig mehr SchülerInnen mit sonderpädagogischem Förderbedarf in Grundschulen (46,9%) unterrichtet werden, als an Gesamtschulen (33,4%) (Bertelsmann-Stiftung, 2015). Insgesamt ist davon auszugehen, dass diagnostische Tätigkeiten für Lehrkräfte im inklusiven Schulsetting häufig noch eine Herausforderung darstellen. Grundsätzlich konnten Forschungsarbeiten jedoch aufzeigen, dass Lehrkräfte, die regelmäßig mit den betreffenden SchülerInnen arbeiten und im natürlichen Setting der Beobachtungssituation beispielsweise als Klassenlehrer agieren, reliablere Beurteilungen vornehmen können als geschulte externe Beobachter (Chafouleas et al., 2010). Um den mangelnden diagnostischen Kompetenzen entgegen wirken zu können, sollten den Lehrkräften angemessene Instruktionen zum Einsatz der Ratingskala vorliegen, die sie dazu befähigen die Ratingskala im Sinne einer verantwortungsvollen Diagnostik einzusetzen (Bühner, 2011). Im Allgemeinen sollten Instruktionen im Rahmen der Testkonstruktion überprüft werden, da sie mögliche Fehler bei der Testkonstruktion darstellen und mitunter missverständlich sein können (Schmidt-Atzert et al., 2012). In den Interviews wurde erfragt, ob die Lehrkräfte die Instruktion als verständlich bezeichnen. Dabei ist zu beachten, dass drei der interviewten Lehrkräfte eine mündliche Schulung erhielten, während der vierten Lehrkraft ausschließlich eine schriftliche Instruktion vorlag. Die Inhalte zum Einsatz und zur Durchführung der Ratingskala waren in der mündlichen Schulung und der schriftlichen Instruktion identisch. Der einzige Unterschied bestand darin, dass die Lehrkräfte, die eine mündliche Schulung erhielten, Verständnisschwierigkeiten unmittelbar im Gespräch äußern konnten. Der schriftlichen Instruktion konnten die Kontaktdaten der TestleiterInnen entnommen werden, sodass Verständnisschwierigkeiten telefonisch oder per Email geklärt werden konnten. Dieses Angebot wurde nicht wahrgenommen. Alle Lehrkräfte gaben an, dass die Instruktion verständlich war. Folglich scheint ein Einsatz der Ratingskala auf Basis mündlicher Schulungen als auch auf Basis schriftlicher Instruktionen möglich zu sein. Ähnliche Ergebnisse finden sich in der Forschungsliteratur, so konnte festgestellt werden, dass bereits Laienbeobachter, die eine kurze Schulung erhalten, zufriedenstellende Verhaltensbeurteilungen vornehmen können (Christ et al., 2011; Huber & Rietz, 2015). Ob eine mündliche Schulung oder eine schriftliche Instruktion zu Unterschieden im Einsatz der Ratingskala führt,

konnte im Zuge dieser Studie nicht erhoben werden. Die aktuelle Forschungslage deutet jedoch daraufhin, dass intensive Schulungen dann notwendig erscheinen, wenn wenig eindeutige Verhaltensweisen in qualitativ mittlerer Ausprägung beurteilt werden müssen (Huber & Rietz, 2015). Die Regelschullehrerin der Grundschule berichtete, dass ihr extreme Verhaltensweisen besser in Erinnerung blieben und diese teilweise weniger auffälliges Verhalten überdeckten. Sie fragte sich, ob ihre Beurteilungen am Ende des langen Zeitraums eines Schultages, welcher zudem mit vielfältigen Eindrücken aller Art einhergeht, den tatsächlichen Beobachtungen entsprachen. Diese Erfahrung der Lehrkraft entspricht einem bekannten Urteilsfehler, welcher in der Literatur als „Baseline-Error“ (Döring & Bortz, 2016, S. 255) beschrieben wird. Die Beurteilungen von Ereignissen beziehen sich dabei nicht auf die objektive Häufigkeit, sondern auf die sogenannte Baseline, wobei irrtümlicherweise prägnante, typische oder im Gedächtnis soeben verfügbare Ereignisse für sehr wahrscheinlich erachtet werden und zur Beurteilung herangezogen werden (ebd.). Dieser Urteilsfehler ist nach Beurteilungen längerer Zeiträume, wie z.B. einem Schultag, sehr wahrscheinlich. In langen Beobachtungszeiträumen konnten die Lehrkräfte große Schwankungen im Verhalten beobachten. Lewis (2016) führt auf, dass SchülerInnen mit Verhaltensstörungen Verhaltensweisen entlang eines Kontinuums von extremen externalisierenden bis extremen internalisierenden Verhaltensweisen zeigen (Lewis, 2016). In welchem Ausmaß die Verhaltensweisen der SchülerInnen schwanken, wird tiefergehend im Zuge der vierten Forschungsfrage diskutiert (vgl., S. 97.). Die Lehrkraft schilderte des Weiteren, dass es ihr schwer fiel die Beurteilungen auf die Länge eines Schultages zu beziehen. Deswegen wird vermutet, dass Verhaltensbeurteilungen mit Ratingskalen, die direkt im Anschluss an kürzere Beobachtungszeiträume durchgeführt werden, zuverlässigere Verhaltensbeurteilungen liefern. Die Studie von Riley-Tillman et al. (2011) stützt diese Annahme. Die Autoren konnten für längere Zeiträume (20 Minuten) im Vergleich zu kürzeren Zeiträumen (10 Minuten) bessere Test-Retest-Reliabilitätswerte feststellen (ebd.). Sie zeigten darüber hinaus, dass sich die Reliabilitätswerte erhöhten, wenn ein langer Beobachtungszeiträume (20 Minuten) in mehrere Beobachtungssequenzen (vier Sequenzen von je einer Länge von 5 Minuten) unterteilt und eine entsprechende Anzahl von Verhaltensbeurteilungen durchgeführt wurde (ebd.). Auch Volpe und Briesch (2012) konnten aufzeigen, dass die Bildung eines Mittelwertes über mehrere Messungen zur Erhöhung der Reliabilität führt. Huber und Rietz (2015) schlussfolgern, dass ein Mittelwert aus mehreren Einzelmessungen als Kennwert robuster ist als ein einzelner erhobener Wert (ebd.). Es erscheint sinnvoll die Beurteilungen mithilfe der Ratingskala auf kürzere Zeiträume (z.B. eine Schulstunde) zu beziehen und Verhaltensbeurteilungen (z.B. mehrmals am Tag) häufiger durchzuführen. Häufigere Verhaltensbeurteilungen könnten die Lehrkräfte entlasten und zugleich zu reliableren Werten in der Verhaltensbeurteilung führen.

Im Zuge der ersten Forschungsfrage wurde des Weiteren erhoben, ob die Ratingskala den Kriterien der Ökonomie und Nützlichkeit entspricht.

Das Kriterium der Ökonomie ist dann erfüllt, wenn die Ratingskala bei geringem Materialaufwand in einer kurzen, angemessenen Zeitspanne durchgeführt sowie schnell und mit wenig Aufwand ausgewertet werden kann (Bundschuh & Winkler, 2014; Bühner, 2011). Es sollte durchweg eine einfache Handhabung möglich sein (Bundschuh & Winkler, 2014). Zusätzlich sollte die Durchführungszeit eine angemessene Länge aufweisen und im Verhältnis zur Bedeutung des diagnostischen Zwecks stehen (Bühner, 2011).

Alle Lehrkräfte gaben an, dass die Ratingskala schnell und häufig durchgeführt werden konnte. Lediglich erste Einarbeitungen und Durchführungen sowie die Reflexion der Ratingskala waren mit einem größeren Aufwand verbunden. Der Aufwand erschien den Lehrkräften aber als angemessen.

Der Aspekt der schnellen und wenig aufwändigen Auswertung konnte im Zuge dieser Studie nicht untersucht werden, da noch kein Tool zur Auswertung der Skala entwickelt wurde und die Lehrkräfte die Ergebnisse der Verhaltensbeurteilungen nicht auswerteten.

Der sonderpädagogische Lehrer der Gesamtschule wünschte sich ausdrücklich ein computerbasiertes Verfahren, welches eine schnelle Auswertung der SchülerInnendaten ermöglichte und eine Grafik des Verhaltensverlaufs bereitstellen sollte (vgl. schriftliche Dokumentation Anhang 3.10). Hartke (2017) gibt an, dass erhobene Schülerdaten anschaulich visualisiert werden sollten, damit sie für Kommunikationsprozesse zwischen Lehrkräften und Eltern genutzt werden können. Denkbar ist eine ähnliche computergestützte Auswertung und Darstellung der Daten wie sie die Onlineplattform LEVUMI bereits für die lernverlaufsdiagnostischen Testverfahren anbietet (Gebhardt et al., 2016).

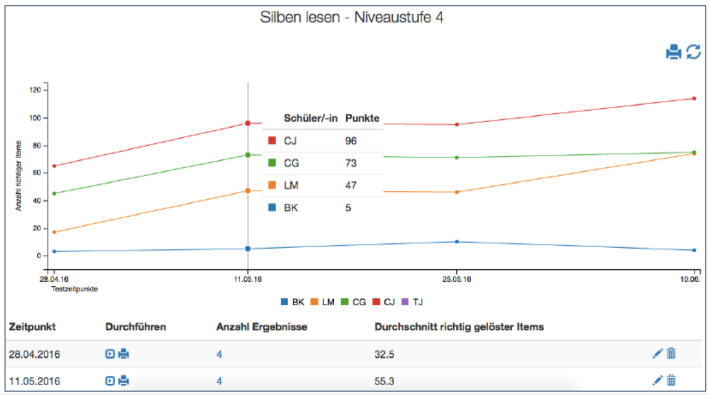

Abbildung 25: Darstellung der Schülerdaten mittels Entwicklungsgraphen – Individueller Lernfortschritt der SchülerInnen (Gebhardt et al., 2016, S. 16).

Eine solche computergestützte Auswertung der Schülerdaten kann sowohl auf Grundlage digitalisierter als auch papierbasierter Ratingskalen realisiert werden (ebd.). Erfolgt eine unmittelbare digitale Erfassung der Verhaltensbeurteilungen, so werden alle weiteren notwendigen

Aufgaben zur Analyse, Auswertung und Darstellung automatisch computergestützt durchgeführt (ebd.). Eine papiergestützte Verhaltensbeurteilung erfordert ein zusätzliches Übertragen der Daten in den Computer (ebd.), was die Ökonomie der Auswertung etwas einschränkt. Sie bietet jedoch den Vorteil, dass nicht zwingend ein digitales Gerät im Rahmen der Beobachtungssituation vorhanden sein muss, um die Verhaltensbeurteilungen durchzuführen zu können. Ist ein digitales Gerät nicht im jeweiligen Setting vorhanden, so könnte potentiell das Kriterium der Direktheit eingeschränkt sein. Jedoch stellt die Direktheit eine wesentliche Eigenschaft von DBR-Verfahren dar (Christ et al., 2009). Diese sollten zeitgleich bzw. unmittelbar nach dem Auftreten des Verhaltens und am gleichen Ort durchgeführt werden (ebd.). Zur Gewährleistung des Kriteriums der Direktheit und des Kriteriums der Ökonomie, scheint es sinnvoll die Ratingskala sowohl in papiergestützter als auch computergestützter Version im Rahmen der Onlineplattform LEVUMI anzubieten. Auf diese Weise haben die Lehrkräfte die Möglichkeit das Medium mit Bezug zur Beobachtungssituation zu wählen.

Ferner erfüllt die Ratingskala dahingehend das Kriterium der Ökonomie, als dass sie unter geringem Materialaufwand durchführbar ist. Pro Verhaltensbeurteilung liegt den Lehrkräften lediglich eine DIN-A4 Seite vor. Die Ratingskala verfügt somit über ein kompaktes Format, ähnlich wie der SDQ (Goodman, 2001). Die beiden Lehrkräfte der Gesamtschule merkten an, dass sie die Aussagekraft der Ratingskala im Vergleich zu weiteren testdiagnostischen Instrumenten aufgrund der Länge für eingeschränkt halten. Sie schlugen vor, die Ratingskala zu einer zweiseitigen Ratingskala zu erweitern (vgl. Anhang 3.9 & 3.10). Dabei ist zu beachten, dass die Generalisierbarkeit und Zuverlässigkeit der Verhaltensbeurteilungen bei einer höheren Itemanzahl lediglich minimal steigt (Casale et al., 2015c). Zugleich steigt die Komplexität bzw. Schwierigkeit einer Bewertungsaufgabe mit zunehmender Anzahl von Items (Huber & Rietz, 2015). Dies könnte wiederum die Ökonomie der Ratingskala einschränken. Da eine Direct Behavior Ratingskala nicht wie ein Screeninginstrument zur umfassenden Diagnostik eines breit gefassten Konstrukts eingesetzt werden soll erscheint eine Erweiterung der Ratingskala in dem Maße, wie es die Lehrkräfte vorschlagen, wenig sinnvoll. Vielmehr sollen hochfrequente Messungen zur änderungssensitiven Erfassung eines eng gefassten Konstrukts ermöglicht werden (Grosche, 2014).

Die Regelschullehrkraft der Grundschule gab an, dass in ihrem Fall ein ökonomischer Einsatz der Ratingskala von dem Wechsel der Valenz der Skalen beeinträchtigt wurde. Die Ratingskala beginnt mit positiv formulierten Items, die ein erwünschtes schulbezogenes Verhalten erfassen. Darauf folgen negativ formulierte Items zu den Skalen VP, HY, EP, die unerwünschte Ausprägungen dieser Verhaltensweisen erheben. Die nächste Skala PS ist wiederum positiv formuliert. Zum Schluss finden sich schließlich negativ formulierte Items zur Skala VPG. Forschungsergebnisse zur Valenz von Items liegen zu drei Verhaltensweisen vor. Die

Studien empfehlen für die verschiedenen Verhaltensweisen jeweils eine unterschiedliche Valenz. Für die Verhaltensweise „Teilnahme am Unterricht“ sollten Items positiv formuliert sein, für die Verhaltensweisen „Störendes Verhalten“ und „Respektvolles Verhalten“ hingegen negativ (Chafouleas et al., 2013; Christ et al., 2011; Riley-Tillman et al., 2009). Eine einheitliche positive oder negative Formulierung der Items der gesamten Skala scheint unter Berücksichtigung dieser Studienergebnisse nicht sinnvoll. Viel eher könnten die Skalen blockweise entsprechend der Valenz angeordnet werden, sodass nur ein einmaliges Umdenken zwischen der Valenz der Skalen notwendig wird. Außedem könnte eine schriftliche Anmerkung auf diesen Wechsel hinweisen und extreme Verhaltensausprärungen konnten als Ankerbeispiele eine Orientierung bieten (Döring & Bortz, 2016). Ob dies einen positiven Einfluss auf die Ökonomie der Ratingskala hat, müssten weitere Studien zeigen.

Basierend auf den vorhergegangenen Auswertungen wird geschlussfolgert, dass ein ökonomischer Einsatz der Ratingskala im inklusiven Kontext der Grund- und Gesamtschule möglich ist. Die Ratingskala erfüllt damit ein weiteres Kriterium verlaufsdiagnostischer Testverfahren und zuteilen die dritte Forderung, die Voß und Gebhardt (2017a) an verlaufsdiagnostische Instrumente stellen. Ein häufiger und kontinuierlicher Einsatz der Ratingskala zur Verlaufsdiagnostik scheint damit gewährleistet.

Das Kriterium der Nützlichkeit wurde nicht explizit durch eine Interviewfrage erhoben. Da es sich um halbstrukturierte Interviews handelte, war es den Lehrkräften unabhängig der Interviewfragen möglich, eigene Anliegen, Anregungen oder Themen vorzubringen. Die Lehrkräfte machten Angaben zur Nützlichkeit der Ratingskala im jeweiligen schulischen Setting. Dementsprechend erfolgt nachfolgend eine Evaluation der Ratingskala im Hinblick auf das Gütekriterium der Nützlichkeit. Dieses Kriterium ist von großer Bedeutung. Es erfasst zum einen, ob die ausgewählten Verhaltensbereiche im sonderpädagogischen Kontext von Relevanz sind. Zum anderen erhebt es, ob das Instrument die Förderung der SchülerInnen in sinnvoller Weise unterstützen kann (Bundschuh & Winkler, 2014; Bühner 2011; Sikora, 2015a). Ersterer Aspekt wird von Casale et al. (2015a) unter dem Kriterium der sozialen Validität aufgegriffen. Die Auswahl der Verhaltensbereiche ist relevant für die Verlaufsdiagnostik von Schülerverhalten (ebd.). Eine Festlegung der Items führt dazu, dass die Ratingskala nur in spezifischen Kontexten von Relevanz ist (vgl. Kapitel 6.3.1). Mithilfe der Ratingskala sollten Verhaltensweisen erfasst werden können, die unmittelbar von Bedeutung für das Verhalten im schulischen Kontext sind (Voß & Gebhardt, 2017a). Ist das Kriterium der sozialen Validität nicht erfüllt, dann ist es zwar möglich Verhaltensveränderungen abzubilden, jedoch wären diese Erhebungen für das schulische Setting nicht von Relevanz. Folglich würde das verlaufsdiagnostische Instrument nur selten eingesetzt werden (Casale et al., 2015a). Die Sonderpädagogin der Grundschule wies darauf hin, dass die Verhaltensbereiche der Ratingskala für das Setting des Klassenverbandes inhaltlich von Relevanz sind. Die Items der Skala PS seien jedoch für den

sonderpädagogischen Kontext weniger ausschlaggebend, da FörderschülerInnen im Zuge inklusiver Settings an Grundschulen häufiger in Einzelunterrichtsphasen unterrichtet werden würden. Wie bereits zuvor diskutiert sind Settings wie der Einzelunterricht oder die Kleingruppe, in der die SchülerInnen im Sinne der äußeren Differenzierung aus dem Unterricht genommen werden, im Grunde nicht mit den Zielen der Inklusion vereinbar, da sie exkludierenden Bedingungen und Prozessen entsprechen (Koch & Textor, 2015). Werden sie jedoch flexibel und offen für alle SchülerInnen angeboten, können sie den Zielen der schulischen Inklusion zuträglich sein (Gebhardt et al., 2014; Koch & Textor, 2015). Inwieweit diese Bedingungen in diesem Fall erfüllt werden, kann auf Basis der Erhebungen nicht beurteilt werden. Einige Lehrkräfte gaben zwar an, die Verhaltensbeurteilungen auf Basis von Situationen, wie Einzelunterricht oder Kleingruppenarbeit durchgeführt zu haben, jedoch bezog sich die Mehrheit der Beobachtungen auf Situationen im Klassenverband. Die SchülerInnen mit Verhaltensstörungen nahmen somit im überwiegenden Maße am gemeinsamen Unterricht teil. Schlussfolgernd werden die Items der Skala PS werden als relevant für das inklusive Setting der Grund- und Gesamtschulen angesehen. Die Sonderpädagogin der Grundschule gab des Weiteren an, dass sonderpädagogische Lehrkräfte häufiger mit SchülerInnen mit externalisierenden Verhaltensschwierigkeiten als mit internalisierenden Verhaltensschwierigkeiten arbeiten würden. So habe sie auch im Rahmen der Studie entsprechende Verhaltensweisen kaum beobachten können und folglich „fast immer ‚Nie'" (vgl. Transkript Anhang 3.8.) angekreuzt. Wie die Ergebnisse der vierten Forschungsfrage zeigen, fallen die Mittelwerte der SchülerInnen in der Skala INT am niedrigsten aus (vgl., S. 97). Es zeigt sich, dass die internalisierenden Verhaltensweisen im Vergleich zu den anderen Verhaltensbereichen sowohl im inklusiven Setting der Grundschule als auch im inklusiven Setting der Gesamtschule seltener beobachtet wurden. Fraglich ist an dieser Stelle, worauf die niedrigen Werte der Verhaltensbeurteilungen zurückzuführen sind. Es stellt sich die Frage, ob internalisierende Verhaltensweisen tatsächlich seltener als andere Verhaltensweisen gezeigt wurden, wie die Sonderpädagogin der Grundschule angibt, oder ob sie lediglich seltener identifiziert werden. Blumenthal, Hartke und Vrban (2017) geben an, dass SchülerInnen mit internalisierenden Verhaltensschwierigkeiten im Vergleich zu SchülerInnen mit externalisierenden Verhaltensschwierigkeiten weniger auffallen. Zudem werden internalisierende Verhaltensweisen seltener als störend empfunden und dementsprechend seltener identifiziert als externalisierende Verhaltensweisen (Blumenthal et al., 2017; Dever et al., 2015). Grundsätzlich zeigen sich jedoch sowohl im Grundschulalter als auch im Jugendalter emotionale Probleme, wie Angststörungen oder Depressionen (Steinhausen, 2016). Ihre Bedeutung sollte nicht unterschätzt werden (Ihle & Esser, 2002). Hölling et al. (2014) berichten auf Basis der Ergebnisse der KiGGS-Studie, dass internalisierende Verhaltensschwierigkeiten, also emotionale Probleme und Probleme mit Gleichaltrigen,

sowohl im Grundschulalter als auch im Jugendalter im Vergleich zu externalisierenden Verhaltensweisen geringer ausgeprägt sind. Es wird jedoch davon ausgegangen, dass die erhobenen Werte aufgrund der Schwierigkeit diese identifizieren zu können unterschätzt werden (Robert Koch-Institut, 2018a). Dementsprechend ist es von Bedeutung internalisierende Verhaltensweisen mithilfe dieser Ratingskala im schulischen Setting zu erfassen. DeVries et al. (2018) konnten aufzeigen, dass Verhaltensprobleme mit Gleichaltrigen, als ein Aspekt internalisierender Verhaltensweisen, einen starken negativen Einfluss auf die schulischen Leistungen der SchülerInnen haben kann. Der Einsatz von präventiven Maßnahmen, wie der vorliegenden Ratingskala, zur frühzeitigen Erfassungen dieser Verhaltensweisen, ist dementsprechend von immenser Bedeutung (Hölling et al., 2014). Es zeigt sich, dass die Ratingskala einen sinnvollen Beitrag zur Erfassung von internalisierenden Verhaltensweisen leisten kann. Darüber hinaus merkten die Lehrkräfte an, dass die SchülerInnen im Verlauf der Zeit große Schwankungen im Verhalten zeigen und deshalb ein Einsatz der Ratingskala zur Erfassung dieser Verhaltensveränderungen nützlich erscheint. Wie bereits erläutert, zeigen SchülerInnen mit Verhaltensschwierigkeiten Verhaltensweisen unterschiedlichster Art (Lewis, 2016). Extrem externalisierende und extrem internalisierende Verhaltensweisen stellen dabei die beiden Pole eines Kontinuums dar (ebd.). Die Regelschullehrkraft der Gesamtschule bezeichnete die Verhaltensbeurteilungen mit der Ratingskala als nützlich, da sie die Beobachtungen, die ohnehin durchgeführt würden, ergänze. Die sonderpädagogische Lehrkraft der Grundschule gab an, dass die Daten der Ratings potentiell als Informationen für kollegiale Gespräche genutzt werden können. Im Allgemeinen entstehen in der pädagogischen Arbeit eine Vielzahl von Kommunikationsanlässen (Hartke, 2017). Effektive Kommunikationen können nur auf Grundlage von Schülerdaten erfolgen (ebd.). Die Ratingskala kann in diesem Zusammenhang die Kommunikation zwischen den am Förderprozess beteiligten Personen als nützliches Werkzeug unterstützen (Volpe & Fabiano, 2013). Zeitnahe, relevante und positive Kommunikationen können einen bedeutenden Beitrag zum Lernerfolg der SchülerInnen leisten (Chafouleas, Reschly, Chaffee & Briesch, 2016). Folglich kann die Ratingskala als nützliches Instrument einen Beitrag zur individuellen Förderung der SchülerInnen im schulischen Setting leisten.

Des Weiteren kann auf Basis der Erhebungen das Kriterium der Inferenz untersucht werden. Mittels des Kriteriums der Inferenz wird überprüft, wie aufwändig oder komplex es ist, schlussfolgernd ein Item zu beantworten (Casale et al., 2015a). Die Befragung der Lehrkräfte sollte erste Hinweise auf den Grad der Inferenz der Items liefern. Generell sollten Items in verlaufsdiagnostischen Instrumenten eine wenig aufwändige Einschätzung der Items ermöglichen, also keine hohe Inferenz aufweisen, da sie sonst kaum ökonomisch einsetzbar sind (ebd.). Geben die Lehrkräfte an, dass sie länger über die Beantwortung eines Items nachdenken mussten, spricht dies für „ein höheres Maß an schlussfolgernden Kognitionen“ und somit für eine hohe Inferenz des Items (ebd., S. 41f.). Die Lehrkräfte gaben an, dass die Items SV01,

SV02, SV03, VP05, HY08, HY09, PS16, VPG17, VPG19 aufgrund konkreter Operationalisierungen schnell beurteilt werden konnten und diese eindeutige Aussagen ermöglichten. Die sonderpädagogische Lehrkraft der Grundschule führte an, dass es ihr sehr schwer fiel die Items der Skala EP zu beurteilen, da sie zum Beispiel nicht beurteilen könne, ob sich ein Schüler fürchte. Es wurde in allen drei Items der Skala EP die Formulierung „wirkt“ gewählt, da sich ein Rater im Kontext subjektiver Beurteilungen „nie sicher sein kann, ob die beobachtenden Lernenden sich tatsächlich etwas nervös oder ängstlich verhalten oder es nur so auf den Betrachter wirkt“ (Sauerland, i.D., S.81). Emotionen bestehen aus zwei verschiedenen Ebenen, der objektiven und der subjektiven Ebene (Hennemann et al., 2017). Die subjektive Ebene entspricht den empfundenen Gefühlen. Diese können nicht beobachtet werden (ebd.). Es ist hingegen jedoch möglich die objektive Ebene, den Ausdruck oder die körperliche Komponente (wie z.B. den Gesichtsausdruck) zu beobachten (ebd.). Die Lehrkraft merkt an, dass es ihr leichter fallen würde einen Gesichtsausdruck zu beurteilen (Vgl. Anhang 3.8). Sie scheint schwierig zu sein, die der Skala zugrundliegenden Angststörungen und depressiven Störungen basierend auf der vorliegenden Formulierung zu beurteilen. Es wäre möglich die Items, insbesondere Item EP12 („Wirkt ängstlich/ fürchtet sich“) durch Beispiele in Form spezifischer Merkmale und Symptome dieser Störungen (z.B. Anspannung, Zittern, Erregtheit, Besorgtheit, Energielosigkeit) zu ergänzen. Dabei könnten zusätzlich Begriffe der körperlichen Komponente wie „Gesichtsausdruck“ oder „Körperhaltung“ berücksichtigt werden. Außerdem wäre es denkbar, die Formulierung „wirkt“ durch die Formulierung „verhält sich“ zu ersetzen, um die objektive Ebene der Emotion verstärkt in der Formulierung der Items zu berücksichtigen. Ob eine veränderte Formulierung die Bearbeitung der Items der Skala EP erleichtert, sollte in zukünftigen Studien untersucht werden.

Die Lehrkräfte führten zudem an, dass mehrdeutige Items (SV04, VP05, VP06, VP07, VPG18), wie zum Beispiel das Item SV04 „Arbeitet ruhig am Platz und verweigert nicht die Mitarbeit“ die Beurteilung von zwei Verhaltensweisen erfordern, schwer auszufüllen waren. Bei der Auswertung der Anmerkungen und Kommentare, die die Lehrkräfte an einige Items geschrieben haben, zeigte sich, dass die Lehrkräfte u.a. einen Teil von mehrdeutigen Items wegstrichen (vgl. Anhang 3.12). Es wird vermutet, dass die Lehrkräfte kenntlich machen wollten, worauf sie sich inhaltlich beziehen. Grundsätzlich ist zu beachten, dass mehrdeutige Items von verschiedenen Ratern möglicherweise unterschiedlich interpretiert werden (Schmidt-Atzert et al., 2012). Es erscheint demzufolge sinnvoll vermeintlich mehrdeutige Items hinsichtlich des Gütekriteriums der Eindimensionalität zu überprüfen. Zukünftige Forschungsarbeiten sollten untersuchen, ob nachfolgende Veränderungen basierend auf der Prüfung der Eindimensionalität die Inferenz der Items verringern und dementsprechend die Ökonomie der Ratingskala erhöhen.

Forschungsfrage 2: Wie ist die interne Konsistenz der Skalen des 4-Faktormodells pro

Messzeitpunkt und über die Messzeitpunkte?

Da es sich bei dem DBR um eine besonders flexible Methode handelt), ist die Überprüfung der Reliabilität, also der „Zuverlässigkeit über verschiedene Messbedingungen“ (Casale et al., 2017, S. 145) von großer Bedeutung. Unter welchen Messbedingungen die Studie durchgeführt wurde, konnte bereits zuvor im Zuge der Beantwortung der ersten Forschungsfrage (vgl. ab S.) erläutert werden. Zur Schätzung der Reliabilität wurde die interne Konsistenz nach Cronbach’s Alpha berechnet. Es wurden zufriedenstellende Werte der internen Konsistenz für die Skalen SV und INT und hohe Werte der internen Konsistenz für die Skalen EXT und PS ermittelt. Diese Werte weisen auf eine hohe Reliabilität hin. Mithilfe der Ratingskala können die vier Verhaltensbereiche schulbezogenes und prosoziales Verhalten, externalisierende und internalisierende Verhaltensweisen reliabel im inklusiven Setting der Grund- und Gesamtschule erhoben werden. Auf Basis der Skalen erfasste Unterschiede liegen scheinen nicht zufällig vorzuliegen, sondern auf die gemessenen Konstrukte zurückzugehen (Bühner, 2011). Sie erfüllt folglich die erste von Voß und Gebhardt (2017a) aufgestellte Forderung an verlaufsdiagnostische Instrumente. Die Ratingskala kann Verhalten konsistent erfassen (Hintze, 2005; Voß, 2014).

Die Überprüfung der Gültigkeit der Skalierung ist bei Verfahren, die zur Einschätzung von Verhalten verwendet werden von grundlegender Bedeutung (Casale et al., 2015a). Um das Gütekriterium der Skalierung zu überprüfen wurde die Trennschärfe der Items bestimmt. Die Trennschärfe liefert Hinweise darauf, in welchem Maße ein Item Messungen des angestrebten Konstrukts ermöglicht (Bühner, 2011). Für drei Items der Skala SV (SV02, SV03, SV04), alle Items der Skala EXT (VP05, VP06, VP07; HY08, HY09, HY10), drei Items der Skala INT (EP11, EP12, EP13) und alle Items der Skala PS (PS14, PS15, PS16) konnten für alle fünf Messzeitpunkte hohe Trennschärfen berechnet werden. Mittelmäßige Trennschärfen wurden für die drei Items VPG17 (zu einem Messzeitpunkt), VPG18 und VPG19 (jeweils zu allen fünf Messzeitpunkten) festgestellt. Diese Items ermöglichen folglich Messungen der jeweiligen Konstrukte schulbezogenes, externalisierendes, internalisierendes und prosoziales Verhalten. Für diese trennscharfen Items gilt, fällt der Kennwert der Person im jeweiligen Item niedrig bzw. hoch aus, so fällt auch der zugehörige Testwert der Skala für die Person niedrig bzw. hoch aus (vgl. Kapitel 6.3.1). Das Item SV01 weist zu drei Messzeitpunkten mittelmäßige und zu zwei Messzeitpunkten sogar nur niedrige Itemtrennschärfen auf. Niedrige Trennschärfen, wie bei dem Item SV01, müssen nicht zwingend zur Entfernung des Items aus der Skala führen (Bühner, 2011). Es ist jedoch fraglich, ob die mittels der Verrechnungsvorschrift gebildeten Testwerte inhaltlich von Wert und interpretierbar sind (Casale et al., 2015a). Grundsätzlich gilt, dass bei in etwa gleichen Trennschärfen innerhalb einer Skala gültige Verrechnungsvorschriften, wie z.B. das Aufsummieren der Items einer Skala, verwendet werden können. Eine ent-

sprechende Verrechnungsvorschrift kann pro Skala die Verschiedenheiten im Verhalten abbilden und zur Interpretation der Konstrukte bzw. Verhaltensbereiche herangezogen werden (Bühner, 2011; Casale et al., 2015a).

Bei der Skala SV ist es aufgrund der niedrigen bis mittelmäßigen Trennschärfe des Items SV01 im Vergleich zu den übrigen Items der Skala fraglich, ob die mittels Verrechnungsvorschrift gebildeten Skalenwerte inhaltlich von Wert sind und für Interpretationen herangezogen werden können. Das Item SV01 („Meldet sich im Unterricht“) operationalisiert die Beteiligung des Schülers am Unterricht und ist inhaltlich dem Konstrukt des Arbeitsverhaltens zuzuordnen (Henning et al., 2017). Eine verlaufsdiagnostische Erfassung des Arbeitsverhaltens ist notwendig, um Verhaltensschwierigkeiten zeitnah feststellen zu können (ebd.). Diese können die schulischen Leistungen von SchülerInnen potentiell negativ beeinflussen (ebd.). Dementsprechend sollte das Arbeitsverhalten verlaufsdiagnostisch erfasst werden, um zeitnah Verhaltensschwierigkeiten feststellen zu können (ebd.). Experteninterviews zeigen zudem eine hohe Relevanz dieses Items für das schulische Setting an. Das Entfernen des Items aus der Skala Schulbezogenes Verhalten erscheint folglich wenig sinnvoll. Ob eine Verrechnungsvorschrift zur Bildung eines Skalenwertes verwendet werden kann und der Skalenwert die Verschiedenheiten im schulbezogenen Verhalten abbilden kann, ist fraglich. In einer weiteren Studie sollte untersucht werden, ob eine Umformulierung des Items die Trennschärfe des Items verbessern könnten. Denkbar wäre beispielsweise die Formulierung „Beteiligt sich aktiv und konstruktiv am Unterricht“. Für dieses Item konnten Henning et al. (2017) eine hohe Trennschärfe und somit hohe Korrelation mit der Gesamtskala zum Arbeits- und Sozialverhalten feststellen.

Für die Items der Skala EXT konnten ausnahmslos hohe Trennschärfen berechnet werden. Für diese Skala kann eine gültige Verrechnungsvorschrift zur Bildung eines Skalenwertes verwendet werden und die daraus folgenden numerischen Unterschiede zwischen den SchülerInnen können die Verschiedenheiten in den externalisierenden Verhaltensweisen abbilden.

Für die Skala INT kann basierend auf den mittelmäßigen bis hohen Itemtrennschärfen das Fazit gezogen werden, dass die Verwendung einer Verrechnungsvorschrift zur Bildung des Skalenwertes möglich ist und dass die Unterschiede zwischen den SchülerInnen in den Skalenwerten die Abbildung unterschiedlicher Ausprägungen internalisierender Verhaltensweisen ermöglichen.

Die Items der Skala PS gehen ebenfalls mit ausnahmslos hohen Trennschärfen einher, sodass ein Skalenwert mittels Verrechnungsvorschrift gebildet werden und dieser die Verschiedenheiten im prosozialen Verhalten zwischen den SchülerInnen abbilden kann.

Die Ratingskala entspricht in den Skalen EXT, INT und PS einem weiteren verlaufsdiagnostisch relevanten Gütekriterium, dem Kriterium der Skalierung. Interpretationen des Testwertes der Skala SV sind aufgrund der niedrigen bis mittelmäßigen Trennschärfe des Items SV01 nur

eingeschränkt möglich. Es sind weitere Studien notwendig, die untersuchen, ob designspezifische Veränderungen, wie z.B. eine Umformulierung des Items, zur Erhöhung der Trennschärfe beitragen können.

Forschungsfrage 3: Wie korrelieren die Werte der fünf Messzeitpunkte der vier Skalen unter Berücksichtigung der Schulform miteinander?

Auf den vorherigen Seiten konnte dargelegt werden, dass die Ratingskala psychometrische Eigenschaften der Statusdiagnostik aufweist. Unklar ist aber noch, ob die Ratingskala die zweite Forderung nach Voß und Gebhardt (2017a) erfüllt und sich für verlaufsdiagnostische Zwecke eignet. Es stellt sich die Frage, ob mithilfe der Ratingskala Veränderungen von Merkmalen über die Zeit erfasst werden können. Dazu wurde im Rahmen der zweiten Fragestellung die Änderungssensibilität der Ratingskala überprüft.

Die Korrelationsergebnisse zeigen, dass für alle Skalen und schulformübergreifend ein Reliabilitätsabfall der Verhaltensbeurteilungen über die fünf Messzeitpunkte verzeichnet werden kann. Klauer (2011) gibt an, dass ein Reliabilitätsabfall über die Zeit nachweist, dass es sich bei dem jeweiligen Testinstrument um ein änderungssensibles Instrument handelt. Folglich erfüllt die vorliegende Ratingskala das Gütekriterium der Änderungssensibilität. Sowohl im schulischen Kontext der Grundschule als auch im schulischen Kontext der Gesamtschule können mithilfe der Ratingskala in kurzen Zeiträumen kleine und relevante Veränderungen abgebildet werden. Die Ratingskala erfüllt folglich auch die zweite Forderung an verlaufsdiagnostische Instrumente nach Voß und Gebhardt (2017a). Nach Klauer (2011) können entsprechende Ergebnisse zudem so interpretiert werden, dass sich das Verhalten der Grund- und GesamtschülerInnen im Verlauf der Zeit sowohl schulformintern als auch schulformübergreifend unterschiedlich verändert. Eine tiefergehende Untersuchung der Verhaltensentwicklungen der SchülerInnen erfolgt im Zuge der vierten Forschungsfrage (vgl., S. 97).

Die Ratingskala ermöglicht folglich die Analyse der individuellen Verhaltensveränderungen in den Bereichen des schulbezogenen und prosozialen Verhaltens sowie in den Bereichen der externalisierenden und internalisierenden Verhaltensweisen. Die Verhaltensentwicklungen einzelner SchülerInnen können abgebildet und Entscheidungen über individuell passende Fördermaßnahmen getroffen werden. Zudem sind zeitnah Aussagen hinsichtlich der Wirksamkeit dieser Fördermaßnahmen möglich (Casale et al., 2015b). Ein Einsatz der Ratingskala in gestuften Fördersystemen wie dem RTI-Ansatz scheint folglich möglich. Ferner kann die Ratingskala einen Beitrag zur evidenzbasierten sonderpädagogischen Praxis, insbesondere der Evidenzbasierung im Einzelfall leisten (ebd.). Es kann geschlussfolgert werden, dass die Ratingskala ein hilfreiches Instrument in der Umsetzung inklusiver Konzepte darstellt.

Insgesamt kann basierend auf den vorherigen Auswertungen und unter Berücksichtigung einer 4-faktoriellen Struktur eine akzeptable Testgüte festgestellt werden. Folglich ist die Ra-

tingskala zur Erfassung von schulbezogenen, externalisierenden, internalisierenden und prosozialen Verhaltensweisen einsetzbar. Für den SDQ konnte zudem eine 5-faktorielle und eine 2-faktorielle Interpretationsstruktur nachgewiesen werden (vgl. z.B. (Goodman et al., 2010).Unter Berücksichtigung der hinzugefügten Skala SV, wäre im ersten Fall eine 6-faktorielle Interpretationsstruktur für die Direct Behavior Ratingskala denkbar. Im zweiten Fall bliebe es bei einer 2-faktoriellen Interpretationsstruktur, wenn ebenfalls eine Problemwertskala bzw. ein Gesamtproblemwert (VP, HY, EP, VPG) gebildet und die Skalen PS und SV zu einer übergeordneten Skala zusammengefasst werden würden. Diese Skala könnte beispielsweise als „Kompetenzskala“ bezeichnet werden. Weitere Studien sollten untersuchen, ob für diese Interpretationsstrukturen ebenfalls eine gute psychometrische Güte nachgewiesen werden kann und ob entsprechende Verrechnungsvorschriften gebildet werden können. Unabhängig der faktoriellen Strukturen der Ratingskala können die Lehrkräfte im Schulalltag zudem einzelne Skalen oder Items auswählen und diese unabhängig der übrigen Skalen und Items zur Erfassung des entsprechenden Schülerverhaltens heranziehen.

Forschungsfrage 4: Welche Mittelwerte weisen die Grund- und GesamtschülerInnen pro Messzeitpunkt und über die Messzeitpunkte auf?

Zuvor konnte nachgewiesen werden, dass es sich bei der Ratingskala um ein reliables und änderungssensitives Instrument handelt. Im Rahmen der vierten Forschungsfrage soll nun erhoben werden, welche Verhaltensausprägungen und –veränderungen Grund- und GesamtschülerInnen im Vergleich über die Zeit aufweisen.

Bei der Interpretation der Mittelwerte ist zu beachten, dass aufgrund der unterschiedlichen Valenz der Skalen hohe Werte in den Skalen Schulbezogenes und Prosoziales Verhalten für das häufige Auftreten angemessener Verhaltensweisen stehen. Im Gegensatz dazu repräsentieren hohe Werte in den Skalen Externalisierende und Internalisierende Verhaltensweisen ein häufiges Auftreten unangemessener Verhaltensweisen. Grundsätzlich ist zu beachten, dass eine Beurteilung von Schülerverhalten im Hinblick auf den Aspekt der Angemessenheit durchweg subjektiv geprägt ist (Casale et al., 2015a). Um die gezeigten Verhaltensweisen nicht nur den beiden Dimensionen „angemessen“ bzw. „unangemessen“ zuordnen zu können, sondern differenziertere Aussagen hinsichtlich der Häufigkeit der gezeigten Verhaltensweisen treffen zu können, werden die sieben Skalenpunkte der Likert-Skala in dieser Pilotierungsstudie mit den folgenden verbalen Marken verknüpft: 1 – nie, 2 – sehr selten, 3 – selten, 4 – gelegentlich, 5 – oft, 6 – sehr oft, 7 – immer (vgl. Kapitel 6.3.2). Die Skalenpunkte werden dabei als Stufen eines Merkmalskontinuums aufgefasst und die Abstände zwischen den Stufen als gleich groß interpretiert (Döring & Bortz, 2016).

Ein Nachteil abstrakter Häufigkeitsratingskalen, wie der vorliegenden, ist, dass aufgrund der abstrakten Begriffe ein gewisser Interpretationsspielraum besteht und unklar bleibt, welche Bedeutung die beurteilenden Personen einer Skalenstufe zuschreiben (ebd.). Dies gilt es zu

berücksichtigen. Eine Konkretisierung der Zeiträume, wie sie von Döring und Bortz (2016) vorgeschlagen wird, erscheint nur dann sinnvoll, wenn interindividuelle Vergleiche vorgenommen werden sollen. Sollen Verhaltensveränderungen hingegen intraindividuell verglichen werden, so erscheinen abstrakte Häufigkeitsratingskalen sinnvoller, da sie den Vorteil bieten, universell über verschiedene Kontexte anwendbar zu sein (ebd.).

Tabelle 10 veranschaulicht, dass die Lehrkräfte der Grund- und Gesamtschule schulformübergreifend das schulbezogene Verhalten der SchülerInnen durchschnittlich mit einem Wert von 4 beurteilten. Die Verhaltensweisen wurden entsprechend der verbalen Marken gelegentlich beobachtet. Die numerische Marke 4 repräsentiert die Skalenmitte. Diese Werte fallen im Vergleich zur Forschungsliteratur etwas niedriger aus. So konnten Henning et al. (2017) für die Gesamtgruppe ihrer Testpersonen für Items zum Arbeits- und Sozialverhalten auf Basis einer Prozentskala Mittelwerte zwischen 56 und 86 feststellen. Ein durchschnittlicher Mittelwert von 3 über alle fünf Messzeitpunkte zeigt, dass Lehrpersonen beider Schulformen externalisierende Verhaltensweisen selten beobachteten. Noch seltener als externalisierende Verhaltensweisen wurden internalisierende Verhaltensweisen schulformübergreifend beobachtet. Prosoziales Verhalten wurde im Vergleich zu den anderen Verhaltensweisen am häufigsten berichtet. Die durchschnittlichen Werte verteilen sich auf die Skalenwerte 4 und 5. Die Verhaltensweisen wurden gelegentlich bis oft beobachtet. Die Werte zu den Skalen EXT, INT und PS werden von der Forschungsliteratur gestützt. So zeigten sich in der KiGGS-Studie (Welle 1) die höchsten Werte im Prosozialen Verhalten (M = 8,3) (Hölling et al., 2014, S. 812). Ebenfalls wurden externalisierende Verhaltensweisen (Verhaltensprobleme M = 2,2; Hyperaktivität M = 3,2) häufiger berichtet als internalisierende Verhaltensweisen (Emotionale Probleme M = 2,0; Peer-Probleme M = 1,4) (ebd., S. 812). Es zeigt sich in Übereinstimmung mit der Forschungsliteratur, dass internalisierende Verhaltensweisen seltener beobachtet wurden als externalisierende Verhaltensweisen (Klasen et al., 2017). Geringere Werte in der Skala internalisierende Verhaltensweisen können mitunter auch darauf zurückgeführt werden, dass diese eher bei Mädchen als bei Jungen berichtet werden (Costello, Copeland & Angold, 2011; Robert Koch-Institut, 2018a). In der Studie wurden jedoch schulformübergreifend mehr Jungen (70,7%) als Mädchen (29,3%) beobachtet. Bei Jungen werden hingegen häufiger externalisierende Verhaltensweisen beobachtet (Costello et al., 2011; Robert Koch-Institut, 2018a). Vermutlich wird das Auftreten von internalisierenden Verhaltensschwierigkeiten jedoch unterschätzt, da diese Auffälligkeiten schwerer zu identifizieren sind (Dever et al., 2015, Robert Koch-Institut, 2018a). Internalisierende Verhaltensschwierigkeiten fallen im Gegensatz zu externalisierenden Verhaltensschwierigkeiten weniger auf und werden seltener als störend empfunden (Blumenthal et al., 2017). Dies sollte jedoch nicht darüber hinwegtäuschen, dass SchülerInnen im Kindes- und Jugendalter von internalisierenden Verhaltensproblemen betroffen sein können (Steinhausen, 2016). Den Verhaltensschwierigkeiten muss mit angemessenen Präventions-

und Interventionsmaßnahmen begegnet werden, denn internalisierende Verhaltensschwierigkeiten, wie die Verhaltensprobleme mit Gleichaltrigen, können einen negativen Einfluss auf schulischen Leistungen von Kindern und Jugendlichen haben (DeVries et al., 2018; Hölling et al., 2014). Sie sollten in ihrer Bedeutung nicht unterschätzt werden (Ihle & Esser, 2002). Die Ergebnisse der KiGGS-Studie weisen darauf hin, dass internalisierende Verhaltensauffälligkeiten von Eltern niedriger eingeschätzt werden als von den betroffenen Kindern selbst (Klasen et al., 2016). Externalisierende Verhaltensweisen werden hingegen häufiger auf Basis von Fremdurteilen erhoben als auf Basis von Selbsturteilen (ebd.). Klasen et al. (2016) empfehlen auf Basis dieser Ergebnisse, dass internalisierende Verhaltensweisen stets von den betroffenen Kindern und Jugendlichen selbst beurteilt werden sollten. Diesbezüglich sollten auch Kinder unter elf Jahren berücksichtigt werden, da sie dieses Verhalten bereits valide beurteilen können (ebd.). Ziel der Förderung von SchülerInnen mit einem emotionalen und sozialen Förderschwerpunkt ist die langfristige Stabilisation der Steuerungsfähigkeit ihres Verhaltens (KMK, 2000). Dies kann nur dann realisiert werden, wenn SchülerInnen aktiv in den Prozess der Förderung einbezogen werden. Es ist wichtig, dass den SchülerInnen die Ziele der Förderung transparent gemacht werden. Dazu ist eine angemessene Darstellung der Schülerdaten notwendig (Hartke, 2017). Auf Basis der Ratingskala und einer angemessenen Darstellung der Auswertung der Schülerdaten kann den SchülerInnen ihr aktueller Entwicklungsstand visualisiert und es können gemeinsam Förderziele festgelegt werden. Es kann beispielsweise ein Liniendiagramm erstellt werden, indem die aktuellen Verhaltensentwicklungen abgebildet werden und eine Ziellinie die erwünschte Verhaltensausprägungen visualisiert (Johnson et al., 2016). Positive Bewertungen des eigenen Verhaltens können dabei zu mehr Motivation und Lernbereitschaft führen (Lohbeck et al., 2015a). Zukünftige Forschungsarbeiten sollten folglich neben der Weiterentwicklung der vorliegenden Ratingskala zur Fremdbeurteilungen auch eine Weiterentwicklung der Ratingskala zur Selbstbeurteilungsversion für SchülerInnen der Primar- und Sekundarstufe in den Blick nehmen.

Ein Vergleich der Mittelwerte zwischen den Schulformen bzw. Altersstufen der SchülerInnen zeigt, dass sich die Mittelwerte der SchülerInnen im Bereich des schulbezogenen Verhaltens nur wenig unterscheiden. Zu vier von fünf Messzeitpunkten weisen die GesamtschülerInnen leicht höhere Werte auf. Eine mögliche Begründung für die Unterschiede kann unter Berücksichtigung entwicklungspsychologischer Ansätze in den unterschiedlichen Bezugsrahmen der SchülerInnen zu finden sein (Voß & Gebhardt, 2017a). So sind GesamtschülerInnen mit anderen Erwartungen hinsichtlich angemessener Verhaltensweisen konfrontiert als GrundschülerInnen. Während GrundschülerInnen, insbesondere Erstklässler, angemessene Verhaltensweisen im schulischen Setting erst erlernen und einüben müssen, werden diese Verhaltensweisen in Setting der Gesamtschule bereits vorausgesetzt (ebd.). Unter Berücksichtigung der Tatsache, dass GesamtschülerInnen im Gegensatz zu Schulanfängern seit mehreren Jahren

mit den Anforderungen und Erwartungen des schulischen Settings konfrontiert sind (Chafouleas et al., 2010), ist zu erwarten, dass die GesamtschülerInnen stärkere Ausprägungen im schulbezogenen Verhalten aufweisen. Diese Annahme wird von den deskriptiven Ergebnissen dieser Studie zu vier von fünf Messzeitpunkten gestützt.

Ein deskriptiver Vergleich der Ausprägungen der externalisierenden Verhaltensweisen zwischen den Schulformen zeigt, dass die Mittelwerte der GrundschülerInnen über alle fünf Messzeitpunkte höher sind als die Mittelwerte der GesamtschülerInnen. Diese Ergebnisse werden von der Forschungsliteratur gestützt. Die Ergebnisse der KiGGS-Studie zeigen, dass externalisierende Verhaltensschwierigkeiten mit zunehmenden Alter abnehmen (Robert Koch-Institut, 2018a). Für die 7- bis 10-Jährigen sind in den Skalen Verhaltensprobleme (M = 2,2) und Hyperaktivität (M = 3,5) höhere Mittelwerte berichtet worden als bei den 11- bis 13-Jährigen (Verhaltensprobleme M = 2,1; Hyperaktivität M = 3,2) sowie den 14- bis 17-Jährigen (Verhaltensprobleme M = 2,0; Hyperaktivität M = 2,7) (Hölling et al., 2014, S. 813).

Die erhobenen Mittelwerte der SchülerInnen zu internalisierenden Verhaltensschwierigkeiten unterscheiden sich im größeren Maße. Zu allen Messzeitpunkten sind internalisierende Verhaltensschwierigkeiten bei den GesamtschülerInnen stärker ausgeprägt als bei den GrundschülerInnen. Diese deskriptiven Ergebnisse können zuteilen von den Ergebnissen der KiGGS-Studie gestützt werden. Während emotionale Probleme mit zunehmenden Alter in stärkeren Ausprägungen auftreten (7 bis 10 Jahre M = 2,1; 11 bis 13 Jahre M = 2,2; 14 bis 17 Jahre M = 2,1), sind Verhaltensprobleme mit Gleichaltrigen mit zunehmenden Alter seltener berichtet worden (7 bis 10 Jahre M = 1,5; 11 bis 13 Jahre M = 1,5; 14 bis 17 Jahre M = 1,4) (ebd., S. 813).

Ein Vergleich der Mittelwerte der prosozialen Verhaltensweisen von Grund- und GesamtschülerInnen zeigt für drei von fünf Messzeitpunkten höhere Ausprägungen prosozialen Verhaltens bei den GrundschülerInnen. In der KiGGS-Studie sind ähnliche Werte erhoben worden. Während für die 7- bis 10-Jährgen und die 11- bis 13-Jährigen mit einem Wert von M = 8,4 hohe Ausprägungen prosozialen Verhaltens berichtet wurden, weisen die 14- bis 17-Jährigen mit einem Wert von M = 8,2 niedrigere Ausprägungen prosozialen Verhaltens auf (ebd., S. 813). Niedrigere Werte im prosozialen Verhalten bei den GesamtschülerInnen können möglicherweise auf die komplexeren Entwicklungsaufgaben der SchülerInnen in der Pubertät zurück geführt werden (Robert Koch-Institut, 2018a).

Die Verläufe über die fünf Messzeitpunkte in dieser Studie veranschaulichen, dass sowohl GrundschülerInnen als auch GesamtschülerInnen in den Skalen schulbezogenen und prosozialen Verhaltens zunehmend häufiger angemessene Verhaltensweisen gezeigt haben. In den Skalen externalisierende und internalisierende Verhaltensweisen wurden zunehmend seltener unangemessene Verhaltensweisen beobachtet. Dies wird auf den positiven Einfluss des Einsatzes einer Verlaufsdiagnostik zurückgeführt (vgl. Kapitel 3.1). In diesem Sinne unterstützen

verlaufsdiagnostische Methoden die Arbeit von Lehrkräften positiv, als dass die Aufmerksamkeit der Lehrkräfte im Hinblick auf die betreffenden Verhaltensweisen aufrechterhalten wird (Volpe & Fabiano, 2013). Auf der Grundlage verlaufsdiagnostischer Instrumente erhalten die Lehrkräfte fortlaufend ein Feedback zur Entwicklung der SchülerInnen. Zahlreiche Forschungsarbeiten konnten aufzeigen, dass sich ein Feedback zum Lernstand und zur Entwicklung der SchülerInnen positiv auf die Effizienz eingesetzter pädagogischer Fördermaßnahmen auswirkt (vgl. z.B. Hattie et al., 2013; Stecker et al., 2005). Dieses Feedback kann die Effizienz eingesetzter Fördermaßnahmen steigern, indem die Passung zwischen den individuellen Bedürfnissen der SchülerInnen und den individuellen Fördermaßnahmen überprüft und somit die Wirksamkeit der Fördermaßnahmen bestimmt wird. Hierbei ist zu beachten, dass Lehrkräfte erst dann Veränderungen im pädagogischen Handeln vornehmen können, wenn sie erstens die Verhaltensentwicklungen der SchülerInnen identifizieren können und zweitens auf Grundlage dieser Daten Rückschlüsse auf ihr unterrichtliches Handeln und die eingesetzten Fördermaßnahmen ziehen können (Voß & Gebhardt, 2017b). Positive Entwicklungen der SchülerInnen sind nicht allein auf den Einsatz verlaufsdiagnostischer Verfahren zurückzuführen, sondern viel eher auf die Modifikationen unterrichtlichen Handelns oder den Einsatz von Feedbackmethoden unter Berücksichtigung der dokumentierten Entwicklungsverläufe (Hattie & Timperley, 2007; Stecker et al., 2005). Es wird davon ausgegangen, dass die positiven Verhaltensentwicklungen der SchülerInnen auf Handlungen, Einflussnahmen und Interventionen der Lehrkräfte zurückzuführen sind. Folglich scheinen die Lehrkräfte dazu in der Lage zu sein, mittels der Ratingskala die Entwicklungen der SchülerInnen identifizieren und Ableitungen für ihr pädagogisches Handeln treffen zu können. Dies stellt eine wichtige Voraussetzung für die Auswahl und Modifikation passender Fördermaßnahmen in Abhängigkeit zur Entwicklung der SchülerInnen dar (Voß & Gebhardt, 2017b). Dabei ist zu berücksichtigen, dass keine einheitlichen bzw. systematischen Interventionen im Rahmen dieser Studie vorgenommen wurden. Ob die Lehrkräfte ihr unterrichtliches Handeln verändert, Entwicklungsfeedbacks oder konkrete Interventionen eingesetzt haben, kann basierend auf den Daten dieser Studie nicht geklärt werden. Weitere Forschungsarbeiten sollten systematisch erheben, inwiefern Lehrkräfte auf Basis der Ratingskala Rückschlüsse auf ihr eigenes pädagogisches Handeln ziehen können.

Können Unterschiede im Verhalten der SchülerInnen über die Zeit festgestellt werden, so liefert dies Hinweise auf die Änderungssensitivität von DBR-Verfahren (Chafouleas et al., 2010). Es ließen sich Verhaltensveränderungen über die Zeit für alle Verhaltensbereiche und beide Schulformen feststellen. Diese Ergebnisse stützen die Ergebnisse der zweiten Fragestellung und untermauern, dass die Ratingskala das Gütekriterium der Änderungssensibilität erfüllt.

Ferner konnte im Rahmen dieser Studie erhoben werden, dass sich das Verhalten der GrundschülerInnen im Vergleich zum Verhalten der GesamtschülerInnen unterschiedlich verändert.

Insbesondere in den Skalen internalisierende Verhaltensstörungen und prosoziales Verhalten weisen Grund- und GesamtschülerInnen signifikante Unterschiede in der Verhaltensveränderung über die Zeit auf, wie eine deskriptive Auswertung der Mittelwerte und eine inferenzstatistische Analyse mittels Varianzanalysen zeigt. Diese Ergebnisse werden von der Forschungsliteratur gestützt. Beispielsweise konnten Chafouleas et al. (2010) ebenfalls für eine Ratingskala zu den Verhaltensweisen Teilnahme am Unterricht und störendes Verhalten einen signifikanten Varianzanteil auf die Facetten Person und Tag zurückführen. Eine Betrachtung der Ausprägungen der internalisierenden Verhaltensschwierigkeiten der Grund- und GesamtschülerInnen über die fünf Messzeitpunkte im Vergleich zeigt, dass die Mittelwerte der GrundschülerInnen in einem Bereich zwischen M = 2,21 und M = 2,61, die Mittelwerte der GesamtschülerInnen hingegen nur in einem Bereich von M = 2,89 und M = 2,99 liegen (vgl. Tabelle 10). Die unterschiedlichen Veränderungen in der Skala internalisierende Verhaltensweisen können darauf zurückgeführt werden, dass internalisierende Verhaltensweisen und insbesondere emotionale Probleme in der Kindheit labiler und kürzer auftreten und mit stärkeren Entwicklungsprozessen einhergehen als in der Adoleszenz (Robert Koch-Institut, 2018a; Steinhausen, 2016). Die signifikanten Unterschiede in der Veränderung prosozialer Verhaltensweisen zwischen den beiden Schulformen lassen sich ebenfalls anhand der Mittelwerte veranschaulichen. Während für die GesamtschülerInnen Werte zwischen M = 4,25 und M = 4,81 berichtet werden können, wurden für die GrundschülerInnen Werte zwischen M = 4,45 und M = 4,75 erhoben. Die größeren Unterschiede im Verhaltensverlauf bei den GesamtschülerInnen können darauf zurückgeführt werden, dass die SchülerInnen in dieser Altersphase neben beobachteten Verhaltensproblemen mit zunehmend komplexeren Entwicklungsaufgaben konfrontiert sind (Robert Koch-Institut, 2018b; Wettstein et al., 2011). Die festgestellten Unterschiede im Verhalten über die Zeit könnten zudem auch auf die unterschiedlichen Beobachtungssettings zurück geführt werden (Chafouleas et al., 2010), denn Verhaltensstörungen stellen Reaktionen auf bestimmte Situationen dar und treten situationsabhängig auf. Wie die Auswertungen der ersten Forschungsfrage darlegt, beobachteten Gesamtschullehrkräfte im Rahmen der fünf Erhebungen häufiger Situationen unterschiedlicher Länge und verschiedener Schulfächer, wohingegen Grundschullehrkräfte Verhalten häufiger zunächst im Rahmen des Klassenverbandes und dann im Rahmen anderer Situationen wie Einzelunterricht beobachteten oder umgekehrt. Diese wechselnden Beobachtungsstationen und Beobachtungszeiträume können mitunter auch einen Einfluss auf die Unterschiede in den Verhaltensänderungen zwischen den Schulformen haben. Ein weiterer Grund für unterschiedliche Verhaltensveränderungen zwischen unterschiedlichen Verhaltensbereichen könnte hingegen auf die Art, die Formulierung oder die Spezifität der Zielverhaltensweisen zurückgeführt werden (Chafouleas et al., 2010).

Die Ergebnisse dieser Studie zeigen, dass schulbezogenes und prosoziales Verhalten sowie

externalisierenden und Verhaltensschwierigkeiten für die SchülerInnen der Stichprobe auf Basis der Ratingskala in ähnlicher Ausprägung wie in der Forschungsliteratur berichtet werden können. Insgesamt zeigen rund 20% aller SchülerInnen psychische Auffälligkeiten, welche mit hohen Persistenzraten einhergehen (Ihle & Esser, 2002; Klasen et al., 2017; Robert Koch-Institut, 2018a). Diese hohen Anteile sowie der Aspekt, dass mangelndes schulbezogenes und prosoziales Verhalten sowie externalisierende und internalisierende Verhaltensschwierigkeiten schulische Leistungen negativ beeinflussen können, zeigt auf, wie wichtig eine frühzeitige Identifikation von Verhaltensschwierigkeiten ist (DeVries et al., 2018) (Reinke, Herman, Petras & Ialongo, 2008). Schulbezogene und prosoziale Verhaltensweisens stellen zugleich eine wichtige Voraussetzung für erfolgreiche schulische Leistungen dar und sollten entsprechend gefördert werden (DeVries et al., 2018; Henning et al., 2017). Auch aus dem Grund, dass verfestigte Verhaltensstörungen nur schwer behandelt werden können sollten Verhaltensschwierigkeiten frühzeitig erkannt und präventive Fördermaßnahmen eingeleitet werden, um Manifestationen entgegen wirken zu können (Petermann & Lehmkuhl, 2010; Voß & Gebhardt, 2017a). Wie die positiven Verhaltensentwicklungen der SchülerInnen über die Zeit zeigen, kann die Ratingskala dazu einen sinnvollen Beitrag leisten, indem die Lehrkräfte auf Basis dieser Verhaltensauffälligkeiten identifizieren und in Anlehnung daran geeignete pädagogische Handlungsmöglichkeiten ableiten. Ein solcher Ansatz, wie zum Beispiel der RTI-Ansatz bietet die Möglichkeit betreffende SchülerInnen frühzeitig zu identifizieren und angemessene Modifikationen des pädagogischen Handelns im Hinblick auf die individuellen Bedürfnisse der SchülerInnen vorzunehmen (Blumenthal, 2016). Dazu werden weitere verlaufsdiagnostische Methoden benötigt (Blumenthal et al., 2017). Ein vielversprechender Ansatz ist dabei das Direct Behavior Rating (ebd.).

Forschungsfrage 5: Welche Variationsbreite weisen die Beurteilungen der Lehrkräfte der Grund- und Gesamtschulen auf?

Im Zusammenhang der Beantwortung der fünften Fragestellung gilt ebenfalls, die Skalenpunkte werden als Stufen eines Merkmalskontinuums und die Abstände zwischen den Stufen als gleich groß interpretiert (Döring & Bortz, 2016).

Die Ergebnisse veranschaulichen, dass die Lehrkräfte bei allen Items im Durchschnitt einen Bereich von einem bzw. zwei Skalenpunkten genutzt haben, um das Verhalten zu beurteilen. Bei allen Items zeigt die minimale Variationsbreite von 0, dass vereinzelte Lehrkräfte keine Verhaltensveränderungen über die fünf Messzeitpunkte beobachten konnten. Ein Maximalwert von 6 verdeutlicht, dass bei diesen Items (VP05, VP07, HY09, HY10, EP11, EP13, PS14, PS15, PS16, VPG17, VPG19) die gesamte Skalenbreite zur Beurteilung der Verhaltensveränderungen über die Zeit genutzt wurde. Auffällig ist, dass die Lehrkräfte bei sieben Items (VP05, VP07, EP11, EP12, EP13, VPG18, VPG19) sogar in 25% der Fälle keine Verhaltensverände-

rungen beobachten konnten und das Verhalten der SchülerInnen konstant mit demselben Skalenwert beurteilten. Bei dem Item VP05 entspricht dies beispielsweise einer Anzahl von 44 Fällen. In 75% der Fälle beurteilten die Lehrkräfte das Verhalten der SchülerInnen mit einer Variationsbreite von zwei bzw. drei Skalenpunkten. Die Werte deuten darauf hin, dass für keines der Items eine mangelnde Differenzierung vorliegt. Die Maximalwerte zeigen, dass die Lehrkräfte mitunter die gesamte Skalenbreite zur Beurteilung des Schülerverhaltens verwendet haben. Unter Berücksichtigung des Aspektes, dass traditionelle Ansätze zur Verhaltensbeurteilung danach streben den Prozentsatz der Varianz zu maximieren, der der Person zugeschrieben werden kann, sind die Ergebnisse der Pilotierungsstudie als zufriedenstellend zu bezeichnen (Chafouleas et al, 2010). Insgesamt erscheint eine Neukonstruktion der Skala nicht notwendig. Die 7-stufige Skala eignet sich folglich zur Erhebung von Verhaltensveränderungen über fünf Messzeitpunkte. Da die Ratingskala im schulischen Kontext zukünftig jedoch zu mehr als fünf Messzeitpunkten eingesetzt werden soll, müssen weitere Forschungsarbeiten zeigen, ob eine 7-stufife Skala bei mehr Messzeitpunkten weiterhin eine angemessene Breite aufweist. Die geringsten Differenzierungen zeigen sich bei den beiden Items EP12 und VPG18. Hier haben die Lehrkräfte zum einen nicht die gesamte Skalenbreite zur Beurteilung verwendet und zum anderen in 25% der Fälle das Verhalten über fünf Messzeitpunkte mit demselben Skalenwert beurteilt. Des Weiteren zeigen die Lehrkräfte in 50% der Fälle lediglich eine Differenzierung von kleiner als einem Skalenwert und in 75% der Fälle eine Differenzierung von kleiner als zwei Skalenwerten.

Ein Vergleich der Schulformen zeigt, dass Grundschullehrkräfte bei 17 von 19 Items geringere Differenzierungen zur Beurteilung der Verhaltensweisen vornehmen als Gesamtschullehrkräfte. Dies wird auf den Urteilsfehler „Baseline-Error“ (Döring & Bortz, 2016, S. 255) zurückgeführt. Im Allgemeinen treten Urteilsfehler im Zuge der Verhaltensbeurteilung häufig auf. Diese sind grundsätzlich bei der Interpretation erhobener quantitativer Daten zu berücksichtigen (ebd.). Der Urteilsfehler „Baseline-Error“ tritt insbesondere dann auf, wenn größere Zeiträume beobachtet werden. Dies ist, wie die schulformspezifische Auswertung der Beobachtungszeiträume dieser Studie zeigt, bei der Mehrheit der Grundschullehrkräfte der Fall. Bei dem Basline-Error orientieren sich die Verhaltensbeurteilungen der SchülerInnen eher an typischen, gerade im Gedächtnis verfügbaren oder extremen, besonders einprägsamen Verhaltensweisen. Es ist davon auszugehen, dass dieser Urteilsfehler bei den Grundschullehrkräften aufgrund der Länge des Beobachtungszeitraums gehäuft aufgetreten ist. Dieser lange Beobachtungszeitraum scheint mit einer geringeren Differenzierung in den Antworttendenzen einherzugehen. Entsprechendes schilderte auch die interviewte Regelschullehrerin der Grundschule. Sie gab an, dass es ihr am Ende eines Schultages schwer fiel das Verhalten der SchülerInnen differenziert zu beurteilen: „So habe ich viel aus dem Bauchgefühl gemacht“ (Vgl. Transkript Anhang 3.7). Dabei ist zu beachten, dass SchülerInnen mit Verhaltensstörungen

Verhaltensweisen entlang eines Kontinuums von extremen externalisierenden bis extremen internalisierenden Verhaltensweisen zeigen (Lewis, 2016). Die Lehrerin gab an, Verhalten in verschiedenen Ausprägungen über den Zeitraum eines Schultages beobachten zu können. Am Ende des Tages fiel es ihr dann schwer eine zusammenfassende Beurteilung zu tätigen: „War es denn wirklich den ganzen Tag so und so?“ (Vgl. Transkript Anhang 3.7). Somit scheinen differenziertere Beurteilungen auf Basis längerer Beobachtungszeiträume, wie z.B. einem Schultag, schwieriger zu sein als kürzere Beobachtungszeiträume, wie eine Schulstunde. Eine Möglichkeit differenzierte Beurteilungen über einen Zeitraum von Schultag zu erlangen, könnte die Untergliederung dieses Zeitraums in verschiedene kürzere Zeiträume sein. Entsprechendes wurde bereits im Zuge der Beantwortung der ersten Forschungsfrage diskutiert (vgl., S. 80).

Kritische Reflexion

Zusammenfassend kann festgestellt werden, dass eine Beantwortung der entwickelten Forschungsfragen basierend auf den erhobenen Daten hinreichend möglich war. Die diskutierten Ergebnisse der Studie ermöglichen Ableitungen für zukünftige Einsätze der Ratingskala im inklusiven Setting der Grund- und Gesamtschulen. Dennoch gilt es zu berücksichtigen, dass es sich bei dem DBR um eine sehr flexible Methode handelt (Casale et al, 2017). Die Ergebnisse dieser Studie sollten demnach nicht ohne weiteres auf andere Settings oder andere Direct Behavior Ratingskalen übertragen werden. Zudem ist bei allen Interpretationen zu beachten, dass es sich bei der vorliegenden Pilotierungsstudie um eine Feldstudie handelt. Potentielle Störvariablen können nicht im vollen Maße kontrolliert werden (Döring & Bortz, 2016). Die Lehrkräfte setzten die Ratingskala eigenverantwortlich ein. Dabei kann es potenziell zu Durchführungsfehlern kommen, die sich nicht offensichtlich in den Daten der Verhaltensbeurteilungen zeigen. Mitunter können intrapersonelle Faktoren die Verhaltensbeurteilungen beeinflussen. So haben beispielsweise die Motivation und die momentane geistige Verfassung der Lehrkräfte Auswirkungen auf die Ratings (Schmidt-Atzert et al., 2012). Sind die Lehrkräfte an verschiedenen Tagen der fünf Verhaltensbeurteilungen unterschiedlich motiviert die Ratingskalen auszufüllen oder unterschiedlich stark durch andere Einflüsse des Schulsettings belastet, kann dies Auswirkungen auf die Genauigkeit der Beurteilungen haben. Entsprechendes wurde beispielsweise von der Regelschullehrkraft der Grundschule berichtet. Am Ende des langen Beobachtungszeitraums von einem Schultag fiel es ihr schwer das Verhalten der SchülerInnen mit hoher Genauigkeit zu beurteilen. Unter Berücksichtigung des zuvor erläuterten Halo-Bias ist folglich insbesondere bei den Lehrkräften, die als Grundlage für die Beobachtungen einen langen Beobachtungszeitraum (wie z.B. einen Schultag) gewählt haben mit Verzerrungen in den Beobachtungen zu rechnen. In zukünftigen Studien könnten die Auswirkungen dieses Durchführungsfehlers begrenzt werden, indem nicht nur eine Verhaltensbeurtei-

lung am Ende, sondern mehrere Verhaltensbeurteilungen über den gesamten Zeitraum durchgeführt werden, denn eine Verrechnung der Werte zu einem Mittelwert führt zu höheren Reliabilitätswerten (Riley-Tillman et al., 2011; Volpe & Briesch, 2012). Zudem lag der Mehrheit der Lehrkräfte lediglich eine schriftliche Instruktion vor. Eine missverständliche Instruktion kann dabei ebenfalls zu Durchführungsfehlern führen (Schmidt-Atzert et al., 2012). Zwar gaben die vier interviewten Lehrkräfte an, dass die Instruktion verständlich sei. Inwiefern die Instruktion aber tatsächlich umgesetzt und wie die Ratingskala tatsächlich eingesetzt wurde, konnte im Rahmen dieser Studie nicht nachvollzogen werden. Weitere Forschungsarbeiten sollten im Speziellen untersuchen, ob auf Basis der schriftlichen Instruktion ein adäquater Einsatz der Ratingskala erfolgt oder ob Schulungen notwendig sind. Die aktuelle Forschungslage liefert Hinweise darauf, dass genaue Verhaltensbeurteilungen mittels DBR Ratingskalen auf Basis kurzer Schulungen möglich sind (Huber & Rietz, 2015). Hinsichtlich schriftlicher Instruktionen liegen keine Forschungsbefunde vor. Ein weiterer Durchführungsfehler wurde bei der Auswertung der Ratingskalen deutlich. Einige der Lehrkräfte missachteten, dass es sich bei den Items einer Ratingskala um festgelegte und vorgegebene Antwortkategorien handelt (Bühner, 2011). Sie notierten Anmerkungen und Kommentare an die Items, strichen einzelne Wörter durch oder gaben Beispiele zu den Items an (vgl. Anhang 3.12). Teilweise wurde durch diese Veränderungen die inhaltliche Bedeutung des Items verändert, sodass entsprechende Daten von der Auswertung ausgeschlossen werden mussten. Auf Basis der vorliegenden Informationen kann nicht abschließend geklärt werden, warum die Lehrkräfte diese Veränderungen vornahmen. Möglicherweise können sie auf Fehler in der Testkonstruktion, wie mehrdeutige Items, zurückgeführt werden. Mehrdeutige Items erschweren die Beurteilung eines Items, da sie unterschiedliche Interpretationen zulassen. Zukünftige Forschungsarbeiten sollten dementsprechend die betreffenden Items (SV03, SV04, VP05, VP06, VP07, HY08, HY09, HY10, EP11, EP12, EP13, PS15, PS16, VPG17, VPG18) hinsichtlich ihrer Eindimensionalität, einem relevanten Gütekriterium verlaufsdiagnostischer Instrumente, überprüfen (Casale et al., 2015a). Eine Limitation dieser Studie stellt des Weiteren der Aspekt dar, dass mögliche Datenpunkte fehlen. Dieses Problem kann im Allgemeinen bei Feldstudien im schulischen Setting erwartet werden (Chafouleas et al., 2010). Ein Teil der fehlenden Werte kann in dieser Studie jedoch zudem auf die Instruktion zurückgeführt werden. Darin wurden die Lehrkräfte dazu aufgefordert, Verhaltensweisen, die sie in der jeweiligen Situation nicht beobachten konnten, auszulassen. Diese Instruktion wurde den Lehrkräften gegeben, da sich Verhaltensbeurteilungen im DBR Ansatz auf die konkreten beobachteten Situationen beziehen. Kann ein Verhalten jedoch nicht in der Situation beobachtet werden, besteht die Gefahr, dass Lehrkräfte Ratings auf Grundlage allgemeiner Eindrücke zum Schülerverhalten durchführen. Dies würde die Ergebnisse verzerren. Darüber hinaus wird in dieser Studie auf das Ausschlussverfahren des Fallweisen Löschens zurückgegriffen, da Unterschiede im Verhalten erst bei einer Anzahl von fünf

Messwerten auf die beobachtete Person zurückgeführt werden können (Huber & Rietz, 2015). Dabei ist zu beachten, dass Ausschlussverfahren zu unbemerkten Verzerrungen in den Ergebnissen führen können und die Aussagekraft der erhobenen Ergebnisse eingeschränkt sein kann (Döring & Bortz, 2016).

7. Fazit und Ausblick

Die vorliegende Arbeit zielte darauf ab, eine im Rahmen des Forschungsprojektes LEVUMI neu konstruierte Direct Behavior Ratingskala im inklusiven Handlungsfeld allgemeinbildender Schulen zu erproben. Es sollte untersucht werden, inwieweit ein Einsatz dieser Ratingskala unter gegebenen Messbedingungen möglich ist. Dazu wurden Lehrkräfte an Grund- und Gesamtschulen gebeten, mithilfe der Ratingskala das Verhalten der SchülerInnen wiederholend zu fünf Messzeitpunkten zu beurteilen. In diesem Zuge wurden die Verhaltensausprägungen und Verhaltensveränderungen der SchülerInnen über die Zeit erfasst. Ziel war es zu überprüfen, ob mithilfe der Ratingskala Verhaltensveränderungen im Verlauf reliabel und änderungssensibel erfasst werden können. Auf Basis der erhobenen Daten sollte außerdem überprüft werden, ob die Ratingskala die verlaufsdiagnostischen Testgütekriterien der Skalierung, Ökonomie und Nützlichkeit bzw. der sozialen Validität erfüllt.

Darüber hinaus war von Interesse, inwiefern und unter welchen Messbedingungen die Lehrkräfte die Ratingskala einsetzten.

Zu Beginn der Arbeit findet sich eine theoretische Einführung in die Thematik schulrelevanter Verhaltensweisen unter Berücksichtigung der Inklusion. Darauffolgend wurde das Konzept der Verlaufsdiagnostik unter Einbeziehung der zwei Ansätze Lernverlaufsdiagnostik und Verhaltensverlaufsdiagnostik sowie den beiden gängigen verhaltensdiagnostischen Methoden systematische direkte Verhaltensbeobachtung und Verhaltensbeurteilung mit Ratingskalen dargestellt. Mit dem SDQ und dem DBR wird im Anschluss daran zum einen ein weit verbreitetes Instrument der Verhaltensbeurteilung mit Ratingskalen (SDQ) und zum anderen ein neuer vielversprechender Ansatz der Verhaltensverlaufsdiagnostik (DBR) vorgestellt. Es schließt sich der empirische Teil der Arbeit mit der Erläuterung des methodischen Vorgehens und der Ergebnisdarstellung der fünf Forschungsfragen an. Zum Abschluss wurden die Ergebnisse der Studie unter Berücksichtigung der Forschungsliteratur diskutiert und kritisch reflektiert. Nachfolgend werden die zentralen Ergebnisse der Pilotierungsstudie vorgestellt.

Da es sich bei der vorliegenden Studie um eine Feldstudie handelt, liegen Übereinstimmungen zwischen den Untersuchungsbedingungen und den Alltagsbedingungen, in denen die Ratingskala zukünftig eingesetzt werden soll, vor. Folglich sind unter Berücksichtigung der Messbedingungen Ableitungen für einen Einsatz der Ratingskala im inklusiven Handlungsfeld der Grund- und Gesamtschulen möglich.

Eine Auswertung der Erhebungszeiträume und der Beobachtungszeiträume zeigt, dass es Grundschullehrkräften häufiger möglich ist Verhaltensbeurteilungen durchzuführen als Gesamtschullehrkräften. Dies wird auf die unterschiedlichen Lehrerprinzipien an Grund- und Gesamtschulen zurückgeführt. Während Lehrkräfte an Grundschulen nach dem Klassenlehrerprinzip eingesetzt werden, unterrichten Lehrkräfte an Gesamtschulen nach dem Fachlehrerprinzip. Diese Ergebnisse deuten darauf hin, dass es Lehrkräften an Grundschulen zu einem früheren Zeitpunkt möglich ist zuverlässige Entscheidungen von pädagogischer Relevanz zu treffen. Weitere Forschungsarbeiten sollten insbesondere für das Setting der Gesamtschule klären, ob mehrere Ratings innerhalb eines Beobachtungszeitraums entsprechende Entscheidungen zu einem früheren Zeitpunkt ermöglichen.

Hinsichtlich der Beobachtungssituationen wird deutlich, dass Grundschullehrkräfte häufiger als Gesamtschullehrkräfte Situationen im Klassenverband beobachteten. Als weitere Beobachtungssituationen wurden in beiden Schulformen das Setting der Kleingruppe und der Einzelunterricht angegeben. Die Ergebnisse deuten darauf hin, dass inklusive Konzepte in der Grundschule fortschrittlicher umgesetzt werden als in der Gesamtschule. Forschungsbefunde stützen diese Ergebnisse: Anteilig werden mehr GrundschülerInnen als GesamtschülerInnen inklusiv beschult (Bertelsmann-Stiftung, 2015). Insbesondere im Bereich der Sekundarstufe bedarf es besserer Ausbildungs- und Weiterbildungsmöglichkeiten, die den Erwerb notwendiger diagnostischer Kompetenzen ermöglichen.

Der Einsatz der Ratingskala ist im inklusiven Setting der Grund- und Gesamtschulen problemlos möglich, wenn die Anzahl parallel zu beobachtender SchülerInnen niedriger als fünf bzw. drei ist. Lange Beobachtungszeiträume (z.B. ein Schultag) erschweren die Beurteilungen. Das Auftreten extremer Verhaltensweisen hingegen erleichtert die Beurteilungen. Dies kann möglicherweise auf den bekannten Urteilsfehler „Baseline-Error“ (Döring & Bortz, 2016, S. 255) zurückgeführt werden. Folglich erscheint es sinnvoll Verhaltensbeurteilungen nicht auf lange, sondern auf kurze Zeiträume zu beziehen und wenn möglich ein Rating mehrmals täglich durchzuführen. Zudem ist es zuträglich, wenn die Lehrkräfte über diagnostische Kompetenzen verfügen. Zukünftige Forschungsarbeiten sollten untersuchen, ob kurze Schulungen zu einem Aufbau diagnostischer Kompetenzen beitragen können und infolgedessen die Genauigkeit der Verhaltensbeurteilungen mittels vorliegender Ratingskala erhöht werden kann.

Ein ökonomischer Einsatz der Ratingskala ist abgesehen von einer umfangreichen Einarbeitung und Reflexion möglich. An dieser Stelle zeigt sich die Notwendigkeit eines ökonomischen Auswertungstools. Denkbar wäre beispielsweise eine Auswertung und Darstellung der Daten ähnlich zu den lernverlaufsdiagnostischen Testinstrumenten der Onlineplattform LEVUMI. In weiteren Forschungsarbeiten sollte ein entsprechendes Tool entwickelt werden. Zusätzlich müssen zukünftige Forschungsarbeiten zeigen, ob eine Anordnung der Skalen unter Berücksichtigung der Valenz zu einer Erhöhung der Ökonomie führen kann.

Insgesamt zeigt sich, dass die Ratingskala eine hohe Nützlichkeit für die individuelle Förderung der SchülerInnen im inklusiven Setting der Grund- und Gesamtschulen aufweist. Zum einen gaben die Lehrkräfte an, dass die SchülerInnen große Schwankungen im Verhalten über die Zeit zeigen. Mithilfe der Ratingskala ist es möglich diese Schwankungen zu erfassen. Zum anderen kann die Ratingskala Beobachtungen ergänzen, die ohnehin im Unterricht durchgeführt werden. Die in diesem Zusammenhang erhobenen Daten können dann als Grundlage für eine effektive Kommunikation zwischen den Lehrkräften, Eltern und SchülerInnen dienen.
In Anlehnung an den SDQ sind verschiedene Ansätze zur Interpretation der Ratingskala denkbar. Die vorliegende Studie fokussiert eine 4-faktorielle Interpretationsstruktur. Unter Berücksichtigung dieser Struktur konnten Schätzungen der internen Konsistenz nach Cronbach's Alpha für die Skalen SV und INT zufriedenstellende Reliabilitätswerte und für die Skalen EXT und PS hohe Werte zeigen. Insgesamt ermöglicht die Ratingskala reliable Verhaltensbeurteilungen unter den zuvor erläuterten Messbedingungen. Erfasste Unterschiede scheinen nicht zufällig vorzuliegen.
Auswertungen hinsichtlich des Gütekriteriums der Skalierung zeigen in drei Skalen (EXT, INT, PS) mittelmäßige bis hohe Trennschärfen. Für diese Skalen können gültige Verrechnungsvorschriften verwendet werden. Für ein Item der Skala SV (SV01) konnten nur niedrige bis mittelmäßige Trennschärfen ermittelt werden. Aufgrund der hohen Relevanz des Items für das schulische Setting erscheint ein Ausschluss des Items aus der Skala wenig sinnvoll. Es ist jedoch fraglich, ob für die Skala SV eine gültige Verrechnungsvorschrift verwendet werden kann. Weitere Forschungsarbeiten müssen zeigen, ob beispielsweise eine Umformulierung des Items unter Berücksichtigung des zu operationalisierenden Konstrukts „Beteiligung am Unterricht" zu einer höheren Trennschärfe führt.
Korrelationsberechnungen nach Pearson zwischen dem ersten Messzeitpunkt und den weiteren Messzeitpunkten zeigen für alle Skalen und schulformübergreifend einen Reliabilitätsabfall über die Zeit. Folglich können mithilfe der Ratingskala Verhaltensveränderungen im Verlauf änderungssensibel erfasst werden.
Die erhobenen Mittelwerte der SchülerInnen zeigen ähnlich hohe erwünschte (SV, PS), unerwünschte (EXT) und ähnlich niedrige (INT) Verhaltensausprägungen im Vergleich zur Forschungsliteratur (Henning et al., 2017; Hölling et al., 2014). Niedrige Ausprägungen in den internalisierenden Verhaltensschwierigkeiten können darauf zurückgeführt werden, dass sie im Gegensatz zu externalisierenden Verhaltensweisen weniger auffallen und seltener als störend empfunden werden. Da sie die schulischen Entwicklungen der SchülerInnen negativ beeinflussen können (DeVries et al., 2018), sind sie von hoher Relevanz und sollten dementsprechend mithilfe der Ratingskala erfasst werden. Tendenziell werden internalisierende Ver-

haltensweisen häufiger im Zuge von Selbsteinschätzungen als im Rahmen von Fremdeinschätzungen berichtet (Klasen et al., 2016). Dies zeigt die Notwendigkeit der Entwicklung einer Selbstbeurteilungsversion der Ratingskala auf.

Auswertungen hinsichtlich der Verhaltensveränderungen der SchülerInnen verdeutlichen, dass die SchülerInnen dieser Studie zunehmend häufiger angemessene Verhaltensweisen in den Skalen SV und PS gezeigt haben und zunehmend seltener unangemessene Verhaltensweisen in den Skalen EXT und INT. Dies kann auf den positiven Einfluss verlaufsdiagnostischer Instrumente zurückgeführt werden, welcher den Lehrkräften eine Grundlage zur Evaluation unterrichtlichen Handelns bietet (Hattie et al., 2013; Stecker et al., 2005). Weiter ist aufzuführen, dass sich das Verhalten der GrundschülerInnen im Vergleich zum Verhalten der GesamtschülerInnen unterschiedlich und in den Skalen INT und PS sogar signifikant unterschiedlich veränderte. Diese Ergebnisse werden von Befunden der Forschungsliteratur gestützt und liegen in den Entwicklungsabläufen der Kinder und Jugendlichen begründet (Voß & Gebhardt, 2017a; Robert Koch-Institut, 2018a; Steinhausen, 2016). Diese Ergebnisse liefern einen weiteren Hinweis darauf, dass es sich bei der Ratingskala um ein änderungssensibles verlaufsdiagnostisches Instrument handelt.

Eine Untersuchung der Skalennutzung durch die Lehrkräfte liefert Hinweise darauf, dass die Breite der 7-stufigen Likertskala angemessen ist, um Verhaltensveränderungen über fünf Ratings zu beurteilen. Weitere Forschungsarbeiten müssen aufzeigen, ob sich diese Skalenbreite auch zur Erfassung von Verhaltensveränderungen über längere Zeiträume (z.B. ein Schuljahr) eignet.

Insgesamt konnte festgestellt werden, dass die Ratingskala alle drei von Voß und Gebhardt (2017a) formulierten Forderungen an verlaufsdiagnostische Instrumente erfüllt. Bei der Ratingskala handelt es sich um ein reliables, änderungssensibles, ökonomisches und nützliches Instrument zur Verhaltensverlaufsdiagnostik, welches einen Beitrag zur Umsetzung der Inklusion im inklusiven Handlungsfeld der Grund- und Gesamtschule leisten kann. Mithilfe der Ratingskala kann die Frage beantwortet werden, „wie Interventionen bei Verhaltensproblemen analog, engmaschig und ökonomisch evaluiert werden können“ (Huber & Rietz, 2015, S. 77). Das Instrument ist insbesondere für die Arbeit mit SchülerInnen mit Verhaltensauffälligkeiten von Nutzen, da die aufwändigen Präventions- und Fördermaßnahmen frühzeitig hinsichtlich ihrer Wirksamkeit evaluiert und bei Bedarf zeitnah an die individuellen Bedürfnisse der SchülerInnen angepasst werden können. Dies kann wiederum die akademischen Leistungen der SchülerInnen positiv beeinflussen und zu Lernerfolgen führen (ebd.). Die Ratingskala kann Lehrkräfte darin unterstützen inklusive Konzepte, wie z.B. den RTI-Ansatz, umzusetzen. Es zeigt sich zudem, dass die Ratingskala zur „Überprüfung des Fördererfolgs im Einzelfall geeignet“ (Casale, 2017, S. 1) ist. Sie kann die Umsetzung einer evidenzbasierten pädagogischen Praxis ermöglichen.

Ziel einer intensiven Erforschung der Methode ist es, dass auf Basis der Gesamtheit aller Forschungen grundsätzliche Ableitungen zur Testgüte von DBR-Verfahren unabhängig der beurteilenden Personen möglich werden (Huber & Rietz, 2015). Die vorliegende Arbeit liefert diesbezüglich Hinweise auf die Konstruktion und den Einsatz von MI-Skalen unter Berücksichtigung der Eigenschaften von Lehrern und Schülern der Grund- und Gesamtschulen. Grundsätzlich sollten die Ergebnisse dieser Studie aufgrund der Flexibilität der Methode und der Abhängigkeit spezifischer Verhaltensweisen von vielfältigen Faktoren vornehmlich mit Bezug zu den erhobenen Messbedingungen interpretiert werden (Casale, 2017; Huber & Rietz, 2015). Die Befunde der vorliegenden Arbeit bestätigen, dass es sich bei dem Direct Behavior Rating um einen vielversprechenden verlaufsdiagnostischen Ansatz handelt.

V. Literaturverzeichnis

Altendorfer-Kling, U., Ardelt-Gattinger, E. & Thun-Hohenstein, L. (2007). Der Selbstbeurteilungsbogen des SDQ anhand einer österreichischen Feldstichprobe. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *35*(4), 265–271. Verfügbar unter https://doi.org/10.1024/1422-4917.35.4.265 [26.04.18]

Amrhein, B. (2015). Professionalisierung für Inklusion - Impulse für die Lehrer/-innenbildung der Sekundarstufe. In E. Kiel, E. Fischer, U. Heimlich, J. Kahlert & R. Lelgemann (Hrsg.), *Inklusion im Sekundarbereich* (S. 140–164). Stuttgart: Verlag W. Kohlhammer.

Autorengruppe Bildungsberichterstattung (Hrsg.). (2018). *Bildung in Deutschland 2018: ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung*. Bielefeld: wbv Media.

Balt, M., Ehlert, A. & Fritz, A. (2017). Theoriegeleitete Testkonstruktion dargestellt am Beispiel einer Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht. *Empirische Sonderpädagogik*, (2), 165–183.

Barkmann, C. & Schulte-Markwort, M. (2007). Psychische Störungen im Kindes- und Jugendalter: Epidemiologie und Diagnostik. *Monatsschrift Kinderheilkunde*, 155(10), 906–914. Verfügbar unter https://doi.org/10.1007/s00112-007-1588-4 [13.05.18]

Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderungen. (2017). *Die UN-Behindertenrechtskonvention. Übereinkommen über die Rechte von Menschen mit Behinderungen.* Verfügbar unter https://www.behindertenbeauftragter.de/SharedDocs/Publikationen/UN_Konvention_deutsch.pdf?__blob=publicationFile&v=2 [17.04.18]

Beelmann, A. (2008). Prävention im Schulalter. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 442–464). Göttingen: Hogrefe.

Beelmann, A. & Raabe, T. (2007). *Dissoziales Verhalten von Kindern und Jugendlichen: Erscheinungsformen, Entwicklung, Prävention und Intervention*. Göttingen: Hogrefe.

Bertelsmann-Stiftung. (2015). *Inklusion in Deutschland. Daten und Fakten.* Gütersloh.

Bierhoff, H.-W. (2010). Prosoziales Verhalten in der Schule. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl.) (S. 671–677). Weinheim: Beltz.

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. Verfügbar unter https://doi.org/10.1080/0969595980050102 [26.04.18]

Blumenthal, Y. (2016). Ein Rahmenkonzept mit mehreren Förderebenen - Response to Intervention (RTI). In B. Hartke (Hrsg.), *Der Response-to-Intervention-Ansatz in der Praxis: Evaluationsergebnisse zum Rügener Inklusionsmodell* (S. 20–32). Münster New York: Waxmann.

Blumenthal, Y., Hartke, B. & Vrban, R. (2017). Schulbasierte Interventionen bei Verhaltensproblemen in der Sekundarstufe nach dem Response-to-Intervention-Ansatz. In A. Methner, K. Popp & B. Seebach (Hrsg.), *Verhaltensprobleme in der Sekundarstufe: Unterricht - Förderung - Intervention* (S. 123–144). Stuttgart: Verlag W. Kohlhammer.

Brezinka, V. (2003). Zur Evaluation von Präventivinterventionen für Kinder mit Verhaltensstörungen. *Kindheit und Entwicklung*, *12*(2), 71–83. Verfügbar unter https://doi.org/10.1026//0942-5403.12.2.71 [26.04.18]

Briesch, A. M., Chafouleas, S. M. & Riley-Tillman, T. C. (2010). Generalizability and Dependability of Behavior Assessment Methods to Estimate Academic Engagement: A Comparison of Systemativ Direct Observation and Direct Behavior Rating. *School Psychology Review*, *39*(3), 408–421.

Briesch, A. M., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C. & Christ, T. J. (2013). The Influence of Alternative Scale Formats on the Generalizability of Data Obtained From Direct Behavior Rating Single-Item Scales (DBR-SIS). *Assessment for Effective Intervention*, *38*(2), 127–133. Verfügbar unter https://doi.org/10.1177/1534508412441966 [26.04.18]

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Bundschuh, K. & Winkler, C. (2014). *Einführung in die sonderpädagogische Diagnostik* (8. Aufl). München: Reinhardt.

Casale, G. (2017). „Nützt es was oder nützt es nichts?“ Direct Behavior Rating (DBR) als diagnostische Methode zur zeitnahen Überprüfung des Fördererfolgs bei unterrichtlichen Schülerinnen- und Schülerverhalten. *Postdamer Zentrum für empirische Inklusionsforschung (ZEIF)*, 2017, Nr. 1. Verfügbar unter https://www.uni-potsdam.de/fileadmin01/projects/inklusion/PDFs/ZEIF-Blog/Casale_2017_Direct_Behavior_Rating.pdf [13.05.18]

Casale, G., Grosche, M., Volpe, R. J. & Hennemann, T. (2017). Zuverlässigkeit von Verhaltensverlaufsdiagnostik über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen. *Empirische Sonderpädagogik*, (2), 143–164.

Casale, G., Hennemann, T. & Grosche, M. (2015b). Zum Beitrag der Verlaufsdiagnostik für eine evidenzbasierte sonderpädagogische Praxis am Beispiel des Förderschwerpunkts der emotionalen und sozialen Entwicklung. *Zeitschrift für Heilpädagogik*, *66*, 325–334.

Casale, G., Hennemann, T., Huber, C. & Grosche, M. (2015a). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung*, *41*(1), 37–54.

Casale, G., Hennemann, T., Volpe, R. J., Briesch, A. M. & Grosche, M. (2015c). Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse. *Empirische Sonderpädagogik*, (3), 258–268.

Chafouleas, S. M. (2011). Direct Behavior Rating: A Review of the Issues and Research in Its Development. *Education and treatment of children*, *34*(4), 575–591.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C. & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology*, *48*, 219–246.

Chafouleas, S. M., Christ, T. J. & Riley-Tillman, T. C. (2009b). Generalizability of Scaling Gradients on Direct Behavior Ratings. *Educational and Psychological Measurement*, *69*(1), 157–173. Verfügbar unter https://doi.org/10.1177/0013164408322005 [03.06.18]

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M. & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review*, *36*(1), 63–79.

Chafouleas, S. M., Hagermoser Sanetti, L. M., Jaffery, R. & Fallon, L. M. (2012b). An Evaluation of a Classwide Intervention Package Involving Self-Management and a Group Contingency on Classroom Behavior of Middle School Students. *Journal of Behavioral Education*, *21*(1), 34–57. Verfügbar unter https://doi.org/10.1007/s10864-011-9135-8 [26.04.18]

Chafouleas, S. M., Jaffery, R., Riley-Tillman, T. C., Christ, T. J. & Sen, R. (2013). The Impact of Target, Wording, and Duration on Rating Accuracy for Direct Behavior Rating. *Assessment for Effective Intervention*, *39*(1), 39–53. Verfügbar unter https://doi.org/10.1177/1534508413489335 [26.04.18]

Chafouleas, S. M., Kilgus, S. P. & Hernandez, P. (2009b). Using Direct Behavior Rating (DBR) to Screen for School Social Risk: A Preliminary Comparison of Methods in a Kindergarten Sample. *Assessment for Effective Intervention*, *34*(4), 214–223. Verfügbar unter https://doi.org/10.1177/1534508409333547 [26.04.18]

Chafouleas, S. M., Reschly, A. L., Chaffee, R. & Briesch, A. M. (2016). Using DBR to Communicate across Contexts. In A. M. Briesch (Hrsg.), *Direct behavior rating: linking assessment, communication, and intervention* (S. 38–57). New York, NY: The Guilford Press.

Chafouleas, S. M., Riley-Tillman, T. C. & Christ, T. J. (2009a). Direct Behavior Rating (DBR): An Emerging Method for Assessing Social Behavior Within a Tiered Intervention System.

Assessment for Effective Intervention, *34*(4), 195–200. Verfügbar unter https://doi.org/10.1177/1534508409340391 [26.04.18]

Chafouleas, S. M., Riley-Tillman, T. C. & McDougal, J. L. (2002). Good, bad, or in-between: How does the daily behavior report card rate? *Psychology in the Schools*, *39*(2), 157–169. Verfügbar unter https://doi.org/10.1002/pits.10027 [03.06.18]

Chafouleas, S. M., Sanetti, L. M. H., Kilgus, S. P. & Maggin, D. M. (2012a). Evaluating Sensitivity to Behavioral Change Using Direct Behavior Rating Single-Item Scales. *Council for Exceptional Children*, *78*(4), 491–505.

Christ, T. J., Riley-Tillman, T. C. & Chafouleas, S. M. (2009). Foundation for the Development and Use of Direct Behavior Rating (DBR) to Assess and Evaluate Student Behavior. *Assessment for Effective Intervention*, *34*(4), 201–213. Verfügbar unter https://doi.org/10.1177/1534508409340390 [03.06.18]

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M. & Boice, C. H. (2010). Direct Behavior Rating (DBR): Generalizability and Dependability Across Raters and Observations. *Educational and Psychological Measurement*, *70*(5), 825–843. Verfügbar unter https://doi.org/10.1177/0013164410366695 [05.06.18]

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M. & Jaffery, R. (2011). Direct Behavior Rating: An Evaluation of Alternate Definitions to Assess Classroom Behaviors. *School Psychology Review*, *40*(2), 181–199.

Conley, L., Marchant, M. & Caldarella, P. (2014). A Comparison of Teacher Perceptions And Research-Based Categories of Student Behavior Difficulties. *Education*, *134*, 439–451.

Conroy, M. A., Stichter, J. P., Daunic, A. & Haydon, T. (2008). Classroom-Based Research in the Field of Emotional and Behavioral Disorders: Methodological Issues and Future Research Directions. *The Journal of Special Education*, *41*(4), 209–222. Verfügbar unter https://doi.org/10.1177/0022466907310369 [26.04.18]

Costello, E. J., Copeland, W. & Angold, A. (2011). Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when adolescents become adults? Trends in psychopathology across the adolescent years. *Journal of Child Psychology and Psychiatry*, *52*(10), 1015–1025. Verfügbar unter https://doi.org/10.1111/j.1469-7610.2011.02446.x [13.05.18]

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G. & Angold, A. (2003). Prevalence and Development of Psychiatric Disorders in Childhood and Adolescence. *Archives of General Psychiatry*, *60*(8), 837. Verfügbar unter https://doi.org/10.1001/archpsyc.60.8.837 [03.06.18]

Daniels, B., Volpe, R. J., Briesch, A. M. & Gadow, K. D. (2017). Dependability and Treatment Sensitivity of Multi-Item Direct Behavior Rating Scales for Interpersonal Peer Conflict. *Assessment for Effective Intervention*, *43*(1), 48–59. Verfügbar unter

https://doi.org/10.1177/1534508417698456 [13.05.18]

Deno, S. L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education*, *37*(3), 184–192. Verfügbar unter https://doi.org/10.1177/00224669030370030801 [03.06.18]

Dever, B. V., Dowdy, E., Raines, T. C. & Carnazzo, K. (2015). Stability and change of behavioral and emotional scores: Screening Stability. *Psychology in the Schools*, *52*(6), 618–629. Verfügbar unter https://doi.org/10.1002/pits.21825 [05.06.18]

DeVries, J. M., Gebhardt, M. & Voß, S. (2017). An assessment of measurement invariance in the 3- and 5-factor models of the Strengths and Difficulties Questionnaire: New insights from a longitudinal study. *Personality and Individual Differences*, *119*, 1–6. Verfügbar unter https://doi.org/10.1016/j.paid.2017.06.026 [13.05.18]

DeVries, J. M., Rathmann, K. & Gebhardt, M. (2018). How Does Social Behavior Relate to Both Grades and Achievement Scores? *Frontiers in Psychology*, *9*. Verfügbar unter https://doi.org/10.3389/fpsyg.2018.00857 [13.05.18]

Diehl, K. & Hartke, B. (2011). Zur Reliabilität und Validität des formativen Bewertungssystems IEL-1: Inventar zur Erfassung der Lesekompetenz von Erstklässlern. *Empirische Sonderpädagogik*, (2), 121–146.

DiPerna, J. C. & Elliott, S. N. (2002). Promoting Academic Enablers to Improve Student Achievement: An Introduction to the Mini-Series. *School Psychology Review*, *31*(3), 293–297.

Di Riso, D., Salcuni, S., Chessa, D., Raudino, A., Lis, A. & Altoè, G. (2010). The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children. *Personality and Individual Differences*, *49*(6), 570–575. Verfügbar unter https://doi.org/10.1016/j.paid.2010.05.005 [03.06.18]

Döpfner, M. & Petermann, F. (2012). *Diagnostik psychischer Störungen im Kindes- und Jugendalter* (3. Aufl). Göttingen: Hogrefe.

Döring, N. & Bortz, J. (2016a). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin Heidelberg: Springer.

Dorsch, F., Wirtz, M. A. & Strohmer, J. (Hrsg.). (2017). *Dorsch - Lexikon der Psychologie* (18. Aufl.). Bern: Hogrefe.

Dresing, T. & Pehl, T. (2017). *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitativ Forschende* (7. Aufl.). Marburg: Eigenverlag.

Ellinger, S. & Stein, R. (2012). Effekte inklusiver Beschulung: Forschungsstand im Förderschwerpunkt emotionale und soziale Entwicklung. *Empirische Sonderpädagogik*, (2), 85–109.

Essau, C. A., Olaya, B., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., O'callaghan, J. & Ollendick, T. H. (2012). Psychometric properties of the Strengths and Difficulties Questionnaire from five European countries: The Strengths and Difficulties Questionnaire. *International Journal of Methods in Psychiatric Research*, *21*(3), 232–245. Verfügbar unter https://doi.org/10.1002/mpr.1364 [05.06.18]

Fingerle, M. (2008). Einführung in die Entwicklungspsychopathologie. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 67–80). Göttingen: Hogrefe.

Forness, S. R., Kim, J. & Walker, H. M. (2012). Prevalence of Students with EBD: Impact on General Education. *Beyond Behavior*, *21*(2), 3–10.

Förster, N., Kuhn, J.-T. & Souvignier, E. (2017). Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, (2), 116–122.

Förster, N. & Souvignier, E. (2015). Effects of Providing Teachers With Information About Their Students' Reading Progress. *School Psychology Review*, *44*(1), 60–75.

Fox, J. & Conroy, M. (1995). Setting events and behavioral disorders of children and youth: An interbehavioral field analysis for research and practice. *Journal of Emotional & Behavioral Disorders*, (3), 130–141.

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review*, *33*(2), 188–192.

Gebhardt, M. (2015). Gemeinsamer Unterricht von Schülerinnen und Schülern mit und ohne sonderpädagogischen Förderbedarf – Ein empirischer Überblick. In E. Kiel, E. Fischer, U. Heimlich, J. Kahlert & R. Lelgemann (Hrsg.), *Inklusion im Sekundarbereich* (S. 39–52). Stuttgart: Verlag W. Kohlhammer.

Gebhardt, M., Casale, G., Jungjohann, J., DeVries, J. (i.D.). Lern-Verlaufs-Monitoring. LEVUMI. Lehrerhandreichung. SDQ, DBR & PIQ. Technische Universität Dortmund.

Gebhardt, M., Diehl, K. & Mühling, A. (2016a). Online-Lernverlaufsmessung für alle Schülerinnen und Schüler in inklusiven Klassen. *Zeitschrift für Heilpädagogik*, *66*, 444–453.

Gebhardt, M., Diehl, K. & Mühling, A. (2016b). *Lern-Verlaufs-Monitoring LEVUMI Lehrerhandbuch*. Technische Universität Dortmund. http://dx.doi.org/10.17877/DE290R-17792

Gebhardt, M., Heine, J.-H., Zeuch, N. & Förster, N. (2015a). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen. *Empirische Sonderpädagogik*, (3), 206–222.

Gebhardt, M. & Mühling, A. (n.d.). Verfügbare Tests. Verfügbar unter https://www.levumi.de/tests [13.05.18]

Gebhardt, M., Sälzer, C. & Tretter, T. (2014). Die gegenwärtige Umsetzung des gemeinsamen Unterrichts in Deutschland. *Heilpädagogische Forschung*, *40*(1), 22–31.

Goodman, A., Lamping, D. L. & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–1191. Verfügbar unter https://doi.org/10.1007/s10802-010-9434-x [03.06.18]

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. Verfügbar unter https://doi.org/10.1111/j.1469-7610.1997.tb01545.x [05.06.18]

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345. Verfügbar unter https://doi.org/10.1097/00004583-200111000-00015 [05.06.18]

Goodman, R., Ford, T., Simmons, H., Gatward, R. & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, *177*(06), 534–539. Verfügbar unter https://doi.org/10.1192/bjp.177.6.534 [03.06.18]

Goodman, R. & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is Small Beautiful? *Journal of Abnormal Child Psychology*, *27*(1), 17–24.

Grosche, M. (2014). Fördermaßnahmen im Prozess überprüfen. Das Konzept der Lernverlaufsdiagnostik. In *Fördern* (S. 113–115). Seelze: Friedrich Verlag.

Haller, A.-C., Klasen, F., Petermann, F., Barkmann, C., Otto, C., Schlack, R. & Ravens-Sieberer, U. (2016). Langzeitfolgen externalisierender Verhaltensauffälligkeiten: Ergebnisse der BELLA-Kohortenstudie. *Kindheit und Entwicklung*, *25*(1), 31–40. Verfügbar unter https://doi.org/10.1026/0942-5403/a000186 [05.06.18]

Hartke, B. (Hrsg.). (2017). *Handlungsmöglichkeiten schulische Inklusion: das Rügener Modell kompakt*. Stuttgart: Verlag W. Kohlhammer.

Hattie, J., Beywl, W. & Zierer, K. (2013). *Lernen sichtbar machen*. Baltmannsweiler: Schneider-Verl. Hohengehren.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. Verfügbar unter https://doi.org/10.3102/003465430298487 [05.06.18]

Hennemann, T., Hövel, D., Casale, G., Hagen, T. & Fitting-Dahlmann, K. (2017). *Schulische Prävention im Bereich Verhalten* (2. Aufl.). Stuttgart: Verlag W. Kohlhammer.

Hennemann, T., Ricking, H. & Huber, C. (2018). Organisationsformen inklusiver Förderung im

Bereich emotional-sozialer Entwicklung. In R. Stein & T. Müller (Hrsg.), *Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung* (2. Aufl.) (S. 115–149). Stuttgart: Verlag W. Kohlhammer.

Henning, T., Schramm, S. A. & Linderkamp, F. (2017). Einschätzung des Arbeits- und Sozialverhaltens durch Lehrkräfte - eine Validierungsstudie. *Empirische Sonderpädagogik*, (1), 52–65.

Hesse, I. & Latzko, B. (2017). Pädagogisch-psychologische Diagnostik - unverzichtbares Werkzeug für gelingende Lehr- und Erziehungstätigkeit aller Lehrkräfte. In A. Methner, K. Popp & B. Seebach (Hrsg.), *Verhaltensprobleme in der Sekundarstufe: Unterricht - Förderung - Intervention* (S. 52–73). Stuttgart: Verlag W. Kohlhammer.

Hillenbrand, C. (2008a). Begriffe und Theorien im Förderschwerpunkt soziale und emotionale Entwicklung - Versuch einer Standortbestimmung. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 5–24). Göttingen: Hogrefe.

Hillenbrand, C. (2008b). *Einführung in die Pädagogik bei Verhaltensstörungen: mit 25 Abbildungen, 6 Tabellen und 45 Übungsaufgaben* (4. Aufl.). München Basel: Ernst Reinhardt Verlag.

Hintze, J. M. (2005). Psychometrics of Direct Observation. *School Psychology Review*, *34*(4), 507–519.

Hölling, H., Schlack, R., Petermann, F., Ravens-Sieberer, U. & Mauz, E. (2014). Psychische Auffälligkeiten und psychosoziale Beeinträchtigungen bei Kindern und Jugendlichen im Alter von 3 bis 17 Jahren in Deutschland – Prävalenz und zeitliche Trends zu 2 Erhebungszeitpunkten (2003–2006 und 2009–2012): Ergebnisse der KiGGS-Studie – Erste Folgebefragung (KiGGS Welle 1). *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *57*(7), 807–819. Verfügbar unter https://doi.org/10.1007/s00103-014-1979-3 [03.06.18]

Huber, C. & Rietz, C. (2015). Direct Behavior Rating (DBR) als Methode zur Verhaltensverlaufsdiagnostik in der Schule: Ein systematisches Review von Methodenstudien. *Empirische Sonderpädagogik*, *7*(2), 75–98.

Ihle, W. & Esser, G. (2002). Epidemiologie psychischer Störungen im Kindes- und Jugendalter: *Psychologische Rundschau*, *53*(4), 159–169. Verfügbar unter https://doi.org/10.1026//0033-3042.53.4.159 [05.06.18]

Ihle, W. & Esser, G. (2008). Epidemiologie psychischer Störungen des Kindes- und Jugendalters. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 49–62). Göttingen: Hogrefe.

Johnson, A. H., Riley-Tillman, T. C. & Chafouleas, S. M. (2016). Summarizing DBR Data for

Interpretation and Decision Making. In A. M. Briesch (Hrsg.), *Direct behavior rating: linking assessment, communication, and intervention* (S. 213–235). New York, NY: The Guilford Press.

Jungjohann, J. & Gebhardt, M. (2018). Lernverlaufsdiagnostik im inklusiven Anfangsunterricht Lesen. Verschränkung von Lernverlaufsdiagnostik, Förderplanung und Wochenplanarbeit. In: F. Hellmich, G. Görel & M.F. Löper (Hrsg.), Inklusive Schul- und Unterrichtsentwicklung (S. 160-173). Stuttgart: Kohlhammer.

Jungjohann, J., Gebhardt, M., Diehl, K. & Mühling, A. (2017). *Förderansätze im Lesen mit LEVUMI*. Technische Universität Dortmund. Verfügbar unter https://eldorado.tu-dortmund.de/bitstream/2003/36024/1/Förderansätze_Lehrerhandbuch%20LEVUMI.PDF [05.06.18]

Jungjohann, J., DeVries, J. M., Gebhardt, M. & Mühling, A. (2018). Levumi: A Web-Based Curriculum-Based Measruement to Monitor Learning Progress in Inclusive Classrooms. In: K. Miesenberger, G. Kouroupetroglou & P. Penaz (Eds.), Computers Helping People with Special Needs. 16th International Conference, ICCHP 2018, Linz, Austria, July 2018, Proceedings (pp. 369–378). Wiesbaden: Springer. https://doi.org/10.1007/978-3-319-94277-3_58

Jurkowski, S. & Hänze, M. (2014). Diagnostik sozialer Kompetenzen bei Kindern und Jugendlichen: Entwicklung und erste Validierung eines Fragebogens. *Diagnostica*, *60*(4), 167–180. Verfügbar unter https://doi.org/10.1026/0012-1924/a000104 [03.06.18]

Kazdin, A. E. (2005). Evidence-Based Assessment for Children and Adolescents: Issues in Measurement Development and Clinical Application. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 548–558. Verfügbar unter https://doi.org/10.1207/s15374424jccp3403_10 [05.06.18]

Kern, L., Evans, S. W., Lewis, T. J., State, T. M., Weist, M. D. & Wills, H. P. (2015). CARS Comprehensive Intervention for Secondary Students With Emotional and Behavioral Problems: Conceptualization and Development. *Journal of Emotional and Behavioral Disorders*, *23*(4), 195–205. Verfügbar unter https://doi.org/10.1177/1063426615578173 [05.06.18]

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R. & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*(1), 64–75. Verfügbar unter https://doi.org/10.1177/0165025415570647 [03.06.18]

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C. & Welsh, M. E. (2012). Direct behavior rating scales as screeners: A preliminary investigation of diagnostic accuracy in elementary school. *School Psychology Quarterly*, *27*(1), 41–50. Verfügbar unter https://doi.org/10.1037/a0027150 [05.06.18]

Kingston, N. & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. Verfügbar unter https://doi.org/10.1111/j.1745-3992.2011.00220.x [16.04.18]

Klasen, F., Meyrose, A.-K., Otto, C., Reiss, F. & Ravens-Sieberer, U. (2017). Psychische Auffälligkeiten von Kindern und Jugendlichen in Deutschland: Ergebnisse der BELLA-Studie. *Monatsschrift Kinderheilkunde*, *165*(5), 402–407. Verfügbar unter s://doi.org/10.1007/s00112-017-0270-8 [03.06.18]

Klasen, F., Petermann, F., Meyrose, A.-K., Barkmann, C., Otto, C., Haller, A.-C., Schlack, R.; Schulte-Markwort, M.; Ravens-Sieberer, U. (2016). Verlauf psychischer Auffälligkeiten von Kindern und Jugendlichen: Ergebnisse der BELLA-Kohortenstudie. *Kindheit und Entwicklung*, *25*(1), 10–20. Verfügbar unter s://doi.org/10.1026/0942-5403/a000184 [16.04.18]

Klasen, H., Woerner, W., Rothenberger, A. & Goodman, R. (2003). Die deutsche Fassung des Strengths and Difficulties Questionnaire (SDQ-Deu) – Übersicht und Bewertung erster Validierungs- und Normierungsbefunde. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *52*(7), 491–502.

Klauer, K. J. (2011). Lernverlaufsdiagnostik - Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik*, (3), 207–224.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 1–17). Göttingen Bern Wien Paris: Hogrefe.

Kóbor, A., Takács, Á. & Urbán, R. (2013). The Bifactor Model of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment*, *29*(4), 299–307. Verfügbar unter s://doi.org/10.1027/1015-5759/a000160 [16.04.18]

Koch, B. & Textor, A. (2015). Spielräume nutzen - Perspektiven inkusiver Schulentwicklung. In E. Kiel (Hrsg.), *Inklusion im Sekundarbereich* (S. 97–139). Stuttgart: Verlag W. Kohlhammer.

Koglin, U., Barquero, B., Mayer, H., Scheithauer, H. & Petermann, F. (2007). Deutsche Version des Strengths and Difficulties Questionnaire (SDQ-Deu): Psychometrische Qualität der Lehrer-/Erzieherversion für Kindergartenkinder. *Diagnostica*, *53*(4), 175–183. Verfügbar unter https://doi.org/10.1026/0012-1924.53.4.175 [05.06.18]

Jungjohann, J., Gebhardt, M., Diehl, K. & Mühling, A. (2017). Förderansätze im Lesen mit LEVUMI. http://dx.doi.org/10.17877/DE290R-18042

Jungjohann, J. & Gebhardt, M. (2018). Lernverlaufsdiagnostik im inklusiven Anfangsunterricht Lesen. Verschränkung von Lernverlaufsdiagnostik, Förderplanung und Wochenplanarbeit. In: F. Hellmich, G. Görel & M.F. Löper (Hrsg.), *Inklusive Schul- und Unterrichtsentwicklung* (S. 160-173). Stuttgart: Kohlhammer.

Jungjohann, J., DeVries, J. M., Gebhardt, M. & Mühling, A. (2018). Levumi: A Web-Based Curriculum-Based Measruement to Monitor Learning Progress in Inclusive Classrooms. In: K. Miesenberger, G. Kouroupetroglou & P. Penaz (Eds.), *Computers Helping People with Special Needs. 16th International Conference*, ICCHP 2018, Linz, Austria, July 2018, Proceedings (pp. 369–378). Wiesbaden: Springer. https://doi.org/10.1007/978-3-319-94277-3_58

Landscheidt, K. (2001). Das Lehrerurteil bei der Früherkennung von Kindern mit Verhaltensstörungen. *Psychologie in Erziehung und Unterricht*, *48*, 107–119.

LeBel, T. J., Kilgus, S. P., Briesch, A. M. & Chafouleas, S. (2010). The Impact of Training on the Accuracy of Teacher-Completed Direct Behavior Ratings (DBRs). *Journal of Positive Behavior Interventions*, *12*(1), 55–63. Verfügbar unter https://doi.org/10.1177/1098300708325265 [16.04.18]

Lewis, T. J. (2016). Does the Field of EBD Need a Distinct Set of “Intensive” Interventions or More Systemic Intensity Within a Continuum of Social/Emotional Supports? *Journal of Emotional and Behavioral Disorders*, *24*(3), 187–190. Verfügbar unter https://doi.org/10.1177/1063426616652866 [05.06.18]

Linderkamp, F. (2007). Diagnostik von Verhaltensstörungen. In F. Linderkamp & M. Grünke (Hrsg.), *Lern- und Verhaltensstörungen: Genese - Diagnostik - Intervention* (S. 121–129). Weinheim Basel: BeltzPVU.

Linderkamp, F. (2015). Anforderungen an wirksames Handeln von Lehrkräften in Inklusionsschulen. In R. Krüger & C. Mähler (Hrsg.), *Gemeinsames Lernen in inklusiven Klassenzimmern: Prozesse der Schulentwicklung gestalten* (S. 109–119). Köln: Carl Link.

Linderkamp, F. & Grünke, M. (2007). Lern- und Verhaltensstörungen: Klassifikation, Prävalenz & Prognostik. In F. Linderkamp & M. Grünke (Hrsg.), *Lern- und Verhaltensstörungen: Genese - Diagnostik - Intervention* (S. 14–28). Weinheim Basel: BeltzPVU.

Link, P.-C. (2018). Inklusion und Verhaltensstörunen: zu Grundsatzfragen. In R. Stein & T. Müller (Hrsg.), *Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung* (2. Aufl.) (S. 225–247). Stuttgart: Verlag W. Kohlhammer.

Lohbeck, A., Nitkowski, D., Petermann, F. & Petermann, U. (2014a). Erfassung von Schülerselbsteinschätzungen zum schulbezogenen Sozial- und Lernverhalten – Validierung der Schülereinschätzliste für Sozial- und Lernverhalten (SSL). *Zeitschrift für Erziehungswissenschaft*, *17*(4), 701–722. Verfügbar unter https://doi.org/10.1007/s11618-014-0582-6 [16.04.18]

Lohbeck, A., Petermann, F. & Petermann, U. (2014b). Reaktive und proaktive Aggression bei Kindern und Jugendlichen – welche Rolle spielen sozial-emotionale Kompetenzen? *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *62*(3), 211–218. Verfügbar unter

https://doi.org/10.1024/1661-4747/a000197 [05.06.18]

Lohbeck, A., Petermann, F. & Petermann, U. (2015a). Selbsteinschätzungen zum Sozial- und Lernverhalten von Grundschulkindern der vierten Jahrgangsstufe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *47*(1), 1–13. Verfügbar unter https://doi.org/10.1026/0049-8637/a000118 [16.04.18]

Lohbeck, A., Schultheiß, J., Petermann, F. & Petermann, U. (2015b). Die deutsche Selbstbeurteilungsversion des Strengths and Difficulties Questionnaire (SDQ-Deu-S): Psychometrische Eigenschaften, Faktorenstruktur und Grenzwerte. *Diagnostica*, *61*(4), 222–235. Verfügbar unter https://doi.org/10.1026/0012-1924/a000153 [05.06.18]

Makarova, E., Herzog, W. & Schönbächler, M.-T. (2014). Wahrnehmung und Interpretation von Unterrichtsstörungen aus Schülerperspektive sowie aus Sicht der Lehrpersonen. *Psychologie in Erziehung und Unterricht*, *61*(2), 127. Verfügbar unter https://doi.org/10.2378/peu2014.art11d [05.06.18]

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12. Aufl.). Weinheim Basel: Beltz.

Methner, A. & Popp, K. (2017). Einführung in die Problematik von Verhaltensstörung in der Sekundarstufe. In A. Methner, K. Popp & B. Seebach (Hrsg.), *Verhaltensprobleme in der Sekundarstufe: Unterricht - Förderung - Intervention* (S. 19–26). Stuttgart: Verlag W. Kohlhammer.

Meuser, M. & Nagel, U. (2010). ExpertInneninterview: Zur Rekonstrukton spezialisierten Sonderwissens. In R. Becker, B. Kortendiek & B. Budrich (Hrsg.), *Handbuch Frauen- und Geschlechterforschung: Theorie, Methoden, Empirie* (3., erw. und durchges. Aufl, S. 376–379). Wiesbaden: VS, Verl. für Sozialwiss.

Mühling, A., Gebhardt, M. & Diehl, K. (2017). Formative Diagnostik durch die Onlineplattform LEVUMI. *Informatik-Spektrum*, *40*(6), 556–561. Verfügbar unter https://doi.org/10.1007/s00287-017-1069-7 [16.04.18]

Myschker, N. & Stein, R. (2014). *Verhaltensstörungen bei Kindern und Jugendlichen: Erscheinungsformen - Ursachen - hilfreiche Maßnahmen* (7. Aufl.). Stuttgart: Verlag W. Kohlhammer.

Nevermann, C. (2008). Angst. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 258–275). Göttingen: Hogrefe.

Niclasen, J., Skovgaard, A. M., Andersen, A.-M. N., Sømhovd, M. J. & Obel, C. (2013). A Confirmatory Approach to Examining the Factor Structure of the Strengths and Difficulties Questionnaire (SDQ): A Large Scale Cohort Study. *Journal of Abnormal Child Psychology*, *41*(3), 355–365. Verfügbar unter https://doi.org/10.1007/s10802-012-9683-y [05.06.18]

Orthmann Bless, D. (2015). Deskriptivstatistik und Inferenzstatistik. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 106–112). Göttingen: Hogrefe.

Petermann, F. (2005). Zur Epidemiologie psychischer Störungen im Kindes- und Jugendalter: Eine Bestandsaufnahme. *Kindheit und Entwicklung*, *14*(1), 48–57. Verfügbar unter https://doi.org/10.1026/0942-5403.14.1.48 [05.06.18]

Petermann, F. & Lehmkuhl, G. (2010). Prävention von Aggression und Gewalt. *Kindheit und Entwicklung*, *19*(4), 239–244. Verfügbar unter https://doi.org/10.1026/0942-5403/a000031 [16.04.18]

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. Verfügbar unter https://doi.org/10.1016/S0001-6918(99)00050-5 [05.06.18]

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (Hrsg.). (2014a). *Quantitative Methoden 1: Einführung in die Statistik für Psychologen und Sozialwissenschaftler.* (4. Aufl., Bd. 1). Berlin Heidelberg: Springer.

Rasch, B., Friese, M., Naumann, E. & Hofmann, W. (Hrsg.). (2014b). *Quantitative Methoden 2: Einführung in die Statistik für Psychologen und Sozialwissenschaftler.* (4. Aufl, Bd. 2). Berlin: Springer.

Ravens-Sieberer, U., Wille, N., Erhart, M., Bettge, S., Wittchen, H.-U., Rothenberger, A., … Döpfner, M. (2008). Prevalence of mental health problems among children and adolescents in Germany: results of the BELLA study within the National Health Interview and Examination Survey. *European Child & Adolescent Psychiatry*, *17*(S1), 22–33. Verfügbar unter https://doi.org/10.1007/s00787-008-1003-2 [16.04.18]

Reicher, H. & Rossmann, P. (2008). Depression. In B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (S. 243–257). Göttingen: Hogrefe.

Reinke, W. M., Herman, K. C., Petras, H. & Ialongo, N. S. (2008). Empirically Derived Subtypes of Child Academic and Behavior Problems: Co-Occurrence and Distal Outcomes. *Journal of Abnormal Child Psychology*, *36*(5), 759–770. Verfügbar unter https://doi.org/10.1007/s10802-007-9208-2 [05.06.18]

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M. & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly*, *24*(1), 1–12. Verfügbar unter

https://doi.org/10.1037/a0015248 [16.04.18]

Riley-Tillman, T. C., Christ, T. J., Chafouleas, S. M., Boice-Mallach, C. H. & Briesch, A. (2011). The Impact of Observation Duration on the Accuracy of Data Obtained From Direct Behavior Rating (DBR). *Journal of Positive Behavior Interventions*, *13*(2), 119–128. Verfügbar unter https://doi.org/10.1177/1098300710361954 [05.06.18]

Robert Koch-Institut. (2018a). Der Verlauf psychischer Auffälligkeiten bei Kindern und Jugend lichen – Ergebnisse der KiGGS-Kohorte. *Journal of Health Monitoring*, *3*(1), 60–65. Verfügbar unter https://doi.org/10.17886/rki-gbe-2018-011 [16.04.18]

Robert Koch-Institut. (2018b). Editorial: Neues von und über KiGGS. *Journal of Health Monitoring*, *3*(1), 3–7. Verfügbar unter https://doi.org/10.17886/rki-gbe-2018-003 [05.06.18]

Rothenberger, A., Erhart, M., Wille, N., Ravens-Sieberer, U. & die BELLA Arbeitsgruppe. (2008). Psychometric properties of the parent strengths and difficulties questionnaire in the general population of German children and adolescents: results of the BELLA study. *European Child & Adolescent Psychiatry*, *17*(S1), 99–105. Verfügbar unter https://doi.org/10.1007/s00787-008-1011-2 [05.06.18]

Rütter, H. & Lühn, A. (2017). *Einstellung und Selbstwirksamkeit zur Inklusion von Lehramtsstudierenden. Eine vergleichende quantitative Erhebung an der Technischen Universität Dortmund.* Technische Universität, Dortmund.

Saile, H. (2007). Psychometrische Befunde zur Lehrerversion des “Strengths and Difficulties Questionnaire“ (SDQ-L). *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *39*(1), 25–32. Verfügbar unter https://doi.org/10.1026/0049-8637.39.1.25 [17.04.18]

Sander, A. (2006). Liegt Inklusion im Trend? *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, (1), 51–53.

Sauerland, A. (i.D.). Konstruktion eines Direct Behavior Ratings: Adaption des Strengths and Difficulties Questionnaire und Pilotierung im Feld. Technische Universität Dortmund.

Scheithauer, H., Mehren, F. & Petermann, F. (2003). Entwicklungsorientierte Prävention von aggressiv-dissozialem Verhalten und Substanzmissbrauch. *Kindheit und Entwicklung*, *12*(2), 84–99. Verfügbar unter https://doi.org/10.1026//0942-5403.12.2.84 [05.06.18]

Schmidt-Atzert, L., Amelang, M., Fydrich, T., Moosbrugger, H. & Zielinski, W. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin Heidelberg: Springer.

Sedlmeier, P. & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (2. Aufl.). München Harlow Amsterdam Madrid Boston San Francisco Don Mills Mexico City Sydney: Pearson.

Seitz, W. & Stein, R. (2010). Verhaltensstörungen. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl.) (S. 919–927). Weinheim: Beltz.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2000). Empfehlungen zum Förderschwerpunkt emotionale und soziale Entwicklung. Beschluss der Kultusministerkonferenz vom 10.03.2000. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2000/2000_03_10-FS-Emotionale-soziale-Entw.pdf [17.04.18]

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2004). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. Verfügbar unter https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf [05.06.18]

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2017). Das Bildungswesen in der Bundesrepublik Deutschland 2014/2015. Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa. Bonn. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-dt-pdfs/dossier_de_ebook.pdf [17.04.18]

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. Verfügbar unter https://doi.org/10.3102/0034654307313795 [05.06.18]

Sikora, S. (2015a). Messinstrumente. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 76–80). Göttingen: Hogrefe.

Sikora, S. (2015b). Operationalisierung. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 68–75). Göttingen: Hogrefe.

Sinner, D. & Kuhl, J. (2015a). t-Test. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 153–158). Göttingen: Hogrefe.

Sinner, D. & Kuhl, J. (2015b). Varianzanalyse. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 159–165). Göttingen: Hogrefe.

Stecker, P. M., Fuchs, L. S. & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*(8), 795–819. Verfügbar unter https://doi.org/10.1002/pits.20113 [17.04.18]

Steege, M. W., Davin, T. & Hathaway, M. (2001). Reliability and Accuracy of a Performance-Based Behavioral Recording Procedure. *School Psychology Review*, *30*(2), 252–261.

Steinhausen, H.-C. (2016). *Psychische Störungen bei Kindern und Jugendlichen: Lehrbuch der Kinder- und Jugendpsychiatrie und -psychotherapie: mit 39 Abbildungen und 73 Tabellen sowie 75 aktuellen Original-Fragebögen und Skalen* (8. Aufl.). München: Elsevier, Urban & Fischer.

Stein, R. & Stein, A. (2014). *Unterricht bei Verhaltensstörungen: ein integratives didaktisches Modell* (2. Aufl). Bad Heilbrunn: Klinkhardt.

Stockheim, D. (2015). Korrelationsanalysen. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 137–144). Göttingen: Hogrefe.

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *42*(2), 111–122. Verfügbar unter https://doi.org/10.1026/0049-8637/a000011 [05.06.18]

Strengths and Difficulties Questionnaire. (2015). Information for researchers and professionals about the Strengths & Difficulties Questionnaires. Verfügbar unter http://www.sdqinfo.org/py/sdqinfo/b3.py?language=German [13.05.18]

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627 [13.05.18]

Volpe, R. J. & Briesch, A. M. (2012). Generalizability and Dependability of Single-Item and Multiple-Item Direct Behavior Rating Scales for Engagement and Disruptive Behavior. *School Psychology Review*, *41*(3), 246–261.

Volpe, R. J. & Fabiano, G. A. (2013). *Daily behavior report cards: an evidence-based system of assessment and intervention*. New York: The Guilford Press.

Volpe, R. J. & Gadow, K. D. (2010). Creating Abbreviated Rating Scales to Monitor Classroom Inattention-Overactivity, Aggression, and Peer Conflict: Reliability, Validity, and Treatment Sensitivity. *School Psychology Review*, *39*(3), 350–363.

Voß, S. (2014). *Curriculumbasierte Messverfahren im mathematischen Erstunterricht - Zur Güte und Anwendbarkeit einer Adaption US-amerikanischer Verfahren im deutschen Schulsystem*. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften. Verfügbar unter http://nbn-resolving.de/urn:nbn:de:101:1-201403124503 [17.04.18]

Voß, S. & Gebhardt, M. (2017a). Monitoring der sozial-emotionalen Situation von Grundschülerinnen und Grundschülern – Ist der SDQ ein geeignetes Verfahren? *Empirische Sonderpädagogik*, (1), 19–35.

Voß, S. & Gebhardt, M. (2017b). Verlaufsdiagnostik in der Schule. *Empirische Sonderpädagogik*, (2), 95–97.

Voß, S. & Hartke, B. (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 83–99). Göttingen Bern Wien Paris: Hogrefe.

Voß, S., Sikora, S. & Hartke, B. (2017). Lernverlaufsdiagnostik als zentrales Element der Prävention von Rechenschwierigkeiten. In A. Fritz, S. Schmidt & G. Ricken (Hrsg.), *Handbuch Rechenschwäche: Lernwege, Schwierigkeiten und Hilfen bei Dyskalkulie* (3. Aufl.) (S. 339–355). Weinheim Basel: Beltz.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. *Heilpädagogische Forschung*, *34*(2), 62–79.

Walter-Klose, C. (2015). Interview. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 280–288). Göttingen: Hogrefe.

Weiß, S., Kollmannsberger, M. & Kiel, E. (2013). Sind Förderschullehrkräfte anders? Eine vergleichende Einschätzung von Expertinnen und Experten aus Regel- und Förderschulen. *Empirische Sonderpädagogik*, (2), 167–186.

Wettstein, A., Bryjová, J., Faßnacht, G. & Jakob, M. (2011). Aggression in Umwelten frühadoleszenter Jungen und Mädchen. Vier Einzelfallstudien mit Kamerabrillen. *Psychologie in Erziehung und Unterricht*, *58*(4), 293–305. https://doi.org/10.2378/peu2011.art14d [13.05.18]

Wiedebusch, S. & Petermann, F. (2011). Förderung sozial-emotionaler Kompetenz in der frühen Kindheit. *Kindheit und Entwicklung*, *20*(4), 209–218. https://doi.org/10.1026/0942-5403/a000058 [17.04.18]

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 281–308). Göttingen Bern Wien Paris: Hogrefe.

Wilbert, J. (2015). Streuung, Standardabweichung und Varianz. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik: eine Einführung* (S. 129–136). Göttingen: Hogrefe.

Wilbert, J. & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, *3*, 225–242.

Woerner, W., Becker, A., Friedrich, C., Rothenberger, A., Klasen, H. & Goodman, R. (2002). Normierung und Evaluation der deutschen Elternversion des Strengths and Difficulties Questionnaire (SDQ): Ergebnisse einer repräsentativen Felderhebung. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *30*(2), 105–112. https://doi.org/10.1024//1422-4917.30.2.105 [17.04.18]

Woerner, W., Fleitlich-Bilyk, B., Martinussen, R., Fletcher, J., Cucchiaro, G., Dalgalarrondo, P., Lui, M., Tannock, R. (2004). The Strengths and Difficulties Questionnaire overseas: Evaluations and applications of the SDQ beyond Europe. *European Child & Adolescent Psychiatry*, *13*(S2). https://doi.org/10.1007/s00787-004-2008-0 [26.04.18]

VI. Tabellenverzeichnis

VII. Abbildungsverzeichnis

VIII. Anhang

Der Anhang ist in digitaler Form auf dem beigefügten Datenträger, der CD-ROM zu finden. Neben der SPSS-Datenmaske und der Syntax weist der Datenträger ein PDF-Dokument auf, das alle relevanten Anlagen zusammenführt. Nachfolgend wird die Strukturierung des Anhangs durch ein Anhangs-Verzeichnis visualisiert.

Anhangs-Verzeichnis

1. SPSS-Datenmaske
2. SPSS-Syntax
3. Anlagen
   - 3.1. Anschreiben zur Studie
   - 3.2. Kurzinformation zum Forschungsprojekt LEVUMI
   - 3.3. Poster
   - 3.4. Erhebungsinstrument: Direct Behavior Ratingskala
   - 3.5. Erhebungsinstrument – Interviewleitfaden
   - 3.6. Transkriptionsregeln nach Dresing & Pehl (2011)
   - 3.7. Tanskript – Regelschullehrerin der Grundschule
   - 3.8. Transkript – Sonderpädagogin der Grundschule
   - 3.9. Transkript – Regelschullehrerin der Gesamtschule
   - 3.10. Schriftliche Dokumentation des Interviews – Sonderpädagoge der Gesamtschule
   - 3.11. Qualitative Inhaltsanalyse nach Mayring (2015)
   - 3.12. Auswertung der Kommentare und Anmerkungen der Lehrkräfte an den Items der Ratingskala

IX. Danksagung

An dieser Stelle möchte ich all denjenigen Personen danken, die mich während der Erstellung der vorliegenden Arbeit unterstützt, beraten und ermutigt und dadurch zum Gelingen dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt Herrn Prof. Dr. Gebhardt und Herrn Prof. Dr. Kuhn, die meine Masterarbeit betreut haben. Ihre Anregungen und konstruktive Kritik ermöglichte mir eine vertiefende Bearbeitung der Fragestellungen dieser Arbeit. Im Besonderen möchte ich mich bei Herrn Prof. Dr. Gebhardt für die wertvollen Hinweise zur Entwicklung geeigneter Forschungsfragen und die engmaschige und richtungsweisende Betreuung bedanken.

Eine besondere Herausforderung stellte die inferenzstatistische Auswertung der erhobenen Daten dar. Denn obwohl die Bedeutung statistischen Grundwissens für den Lehrberuf in den letzten Jahren stetig an Bedeutung gewinnt, stand die Vermittlung statistischen Grundwissens bisher nicht im Zentrum des Studiengangs „Lehramt für sonderpädagogische Förderung“ an der TU Dortmund, Fakultät Rehabilitationswissenschaften. Einen besonderen Dank möchte ich deshalb an Herrn Prof. Dr. Kuhn für Beratungen zu inferenzstatistischen Fragen richten. Des Weiteren möchte ich mich bei Herrn DeVries und den MitarbeiterInnen des Statistischen Beratungs- und Analyse Zentrum (SBAZ) für Beratungen zur Arbeit mit der statistischen Analysesoftware „SPSS“ bedanken.

Die vorliegende Masterarbeit entstand im Rahmen einer Pilotierungsstudie des Forschungsprojektes „LEVUMI“, welches an der Technischen Universität Dortmund (Prof. Dr. Markus Gebhardt), der Europa-Universität Flensburg (Prof. Dr. Kirsten Diehl) und der Universität Kiel (Prof. Dr. Andreas Mühling) durchgeführt wird. Ich möchte mich insbesondere bei Anna Sauerland für die Unterstützung bei der Durchführung der Pilotierungsstudie an zahlreichen Grund- und Gesamtschulen bedanken. In Zusammenarbeit mit ihr konnte ein umfangreicher Datensatz generiert werden, der eine fundierte Grundlage für die vorliegende wissenschaftliche Studie darstellt. Ein weiterer Dank geht an Frau Jungjohann, die uns hinsichtlich der organisatorischen Vorbereitung der Pilotierungsstudie beratend zur Seite stand. Mein Dank gilt an dieser

Stelle ebenso den vielen Lehrkräften, die an unserer Studie teilgenommen und diesbezüglich eine weitere Aufgabe im anspruchsvollen Schulalltag auf sich genommen haben, um die Realisierung dieser Pilotierungsstudie und damit meiner Masterarbeit zu ermöglichen.
Ebenfalls möchte ich mich bei Claudia Hisker und Pia Trösken bedanken, die meine Arbeit Korrektur gelesen haben und so zur Qualität der vorliegenden Arbeit beigetragen haben.
Abschließend möchte ich mich bei meinen Eltern und meinem Freund für die bedingungslose Unterstützung sowohl im Zuge dieser Masterarbeit als auch im Zuge meines gesamten Studiums danken.
Sarah Hisker
Münster, 24.07.2018

X. Eidesstattliche Versicherung

Eidesstattliche Versicherung
(Affidavit)

Hisker, Sarah
Name, Vorname
(Last name, first name)

[redacted]
Matrikelnr.
(Enrollment number)

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/Masterarbeit* mit dem folgenden Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present Bachelor's/Master's* thesis with the following title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution.

Titel der ~~Bachelor-~~/Masterarbeit*:
(Title of the Bachelor's/ Master's* thesis):

Veränderungen im Direct Behavior Rating (DBR) über die Zeit - Eine Pilotierung mit fünf Messzeitpunkten in Grund- und Gesamtschulen

*Nichtzutreffendes bitte streichen
(Please choose the appropriate)

Münster, 24.07.2018
Ort, Datum
(Place, date)

[redacted]
Unterschrift
(Signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to €50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, section 63, subsection 5 of the North Rhine-Westphalia Higher Education Act (*Hochschulgesetz*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:**

Münster, 24.07.2018
Ort, Datum
(Place, date)

[redacted]
Unterschrift
(Signature)

**Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.