technische universität dortmund

fakultät rehabilitationswissenschaften

Masterarbeit

Konstruktion eines Direct Behavior Ratings

Fakultät der Rehabilitationswissenschaften

Entwicklung und Erforschung inklusiver Bildungsprozesse

Erstgutachter: Prof. Dr. Markus Gebhardt

Zweitgutachter: Prof. Dr. Jörg-Tobias Kuhn

Vorgelegt von:

Anna Sauerland

[redacted]

Vorgelegt am:

18.07.2018

Kurzfassung

Die vorliegende Arbeit hat die Konstruktion eines Direct Behavior Ratings zur Anwendung in Schulen zum Ziel. Die Grundlage für die Entwicklung dieses verhaltensverlaufsdiagnostischen Instruments liefert das bereits umfänglich erprobte statusdiagnostische Instrument des Strengths and Difficulties Questionnaires. In mehreren Erprobungsschritten wird das im Vorhinein theoretisch begründete Direct Behavior Rating in der Praxis an einer Förderschule auf Praktikabilität, Ökonomie, Formulierung der Items und Interrater-Reliabilität hin untersucht und weiterentwickelt. Dabei zeigt sich, dass das überprüfte Instrument von den Lehrkräften generell positiv angenommen wird und sich als praktikabel erweist. In Bezug auf die Formulierung der Items ergibt sich insbesondere, dass eine spezifische Formulierung zu einer Senkung der Inferenz und einer damit verbundenen Steigerung der Ökonomie führt. Auch die Interrater-Reliabilität des Instruments wird in einer Erprobungsphase mit positivem Resultat getestet. Die finale Version des Direct Behavior Ratings wird in einer anschließenden Pilotierungsstudie eingesetzt und dessen Testgüte bestätigt.

Inhaltsverzeichnis

1. Einleitung

Das Verhalten von Kindern und Jugendlichen im Kontext Schule ist in jeglicher Hinsicht von zentraler Bedeutung. Ungeachtet der Schulform, der Klassenstufe, des Geschlechts und des Unterrichtsfachs – auffälliges Verhalten kann immer auftreten, wobei in der Regel Intensität, Art des Auftretens, Ursachen und Bereiche, in denen es vorkommt, variieren. Unterschieden wird häufig zwischen externalisierenden, also ausagierenden, direkt auffälligen Verhaltensweisen und internalisierenden Verhaltensauffälligkeiten, welche sich eher im Verborgenen zeigen und manchmal sogar unbemerkt bleiben. Aufgrund einer Heterogenität innerhalb der Schülerschaft, die im Rahmen der Inklusion noch verstärkt wird, wird die Diagnostik in den Bereichen Lernen und Verhalten auch an Regelschulen immer wichtiger. Lernen und Verhalten, welche in direkter Interaktion miteinander stehen und einander bedingen und beeinflussen können (DeVries, Rathmann & Gebhardt, 2018; Hartmann, 2017), stehen im Fokus schulischen Handelns. Ein ständiger Fortschritt beziehungsweise eine fortlaufende Verbesserung in beiden Bereichen wird von Lehrkräften angestrebt. Für eine Überprüfung und Evaluation der Effektivität angewendeter Fördermaßnahmen und Interventionen und damit eine evidenzbasierte Pädagogik reichen statusdiagnostische Instrumente, die den Ist-Stand darstellen und schon seit einiger Zeit aufgrund ihrer ausschließlich punktuell gültigen Diagnosen immer stärker in die Kritik geraten, nicht aus (u. a. Hartmann, 2017; Breitenbach, 2003). Im Bereich der Lernentwicklungsforschung und der entsprechenden Diagnostik sind mit dem curriculum-based measurement und dem Response-To-Intervention-Ansatz inzwischen einige Schritte in Richtung einer Lernverlaufsdiagnostik gegangen worden (Jungjohann & Gebhardt, 2018). Insbesondere in den Bereichen Lesen, Schreiben und mathematische Grundbildung gibt es bereits vielfältige Instrumente zur Lernverlaufsdiagnostik (Gebhardt, Diehl & Mühling, 2016a, 2016b; Jungjohann, Gebhardt, Diehl & Mühling, 2017; Jungjohann, DeVries, Gebhardt & Mühling, 2018), wobei sich die Diskussion um verlaufsdiagnostische Instrumente derzeit noch in vielen Fällen auf den kognitiven Bereich beschränkt (Huber & Rietz, 2015). Für die Verhaltensverlaufsdiagnostik gibt es im deutschsprachigen Raum gegenwärtig noch keine derartigen Instrumente und auch im englischsprachigen Raum beschränken sich diese im Rahmen von Direct Behavior Ratings (im Folgenden DBR) auf die Untersuchung einzelner Verhaltensweisen in Form von Single-Item-Skalen (Casale, Hennemann, Huber & Grosche, 2015; Casale, Hennemann & Grosche, 2015). Die mehrere Verhaltensbereiche umfassende Verlaufsdiagnostik mit Hilfe von Multi-Item-Skalen muss dementsprechend für den deutschsprachigen

Raum noch entwickelt und etabliert werden. Ein Schritt in diese Richtung soll mit der vorliegenden Arbeit geleistet werden, welche sich mit ebendieser Entwicklung eines Instruments zur Verhaltensverlaufsdiagnostik auseinandersetzt. Ziel des Vorhabens ist es, ein gut funktionierendes Direct-Behavior-Rating-Instrument aus einem bereits bestehenden statusdiagnostischen Diagnoseinstrument zu entwickeln, das Lehrkräfte im Unterrichtskontext erlaubt, die Wirkung von Fördermaßnahmen zu überprüfen und Verhaltensentwicklung zu messen. Die vorliegende Arbeit soll dementsprechend einen Beitrag zur Entwicklung verhaltensverlaufsdiagnostischer Instrumente leisten. Im Vordergrund stehen die Praktikabilität, die Ökonomie und die generelle Durchführbarkeit des Instruments, welche in mehreren Schritten überprüft und durch mehrfache Anpassung sichergestellt werden sollen.

Zu Anfang soll im Rahmen einer theoretischen Verortung auf die Bereiche Verhalten und Verhaltensauffälligkeiten in der Schule und Lern- und Verhaltensverlaufsdiagnostik sowie auf das für die vorliegende Arbeit relevante Diagnoseinstrument, den „Strengths and Difficulties Questionnaire“ (im Folgenden SDQ), eingegangen werden. Anschließend wird die Methode des DBRs erläutert, in den aktuellen Forschungsstand eingeordnet und in Bezug auf notwendige testdiagnostische Kriterien dargestellt. Auf Grundlage des theoretischen Hintergrundes werden im Folgenden drei Forschungsfragen erarbeitet. Diese werden im anschließenden Kapitel durch die Analyse zweier Erprobungen, verschiedene qualitative Überprüfungen sowie eine Überprüfung der Interrater-Reliabilität des Instruments und eine an diese Arbeit anknüpfende Studie beantwortet. Das dargelegte Vorgehen wird im letzten Kapitel zusammengefasst und diskutiert. Zudem werden die Ergebnisse genutzt, um einen Ausblick auf sich eventuell ergebende Forschungsdesiderate zu geben.

2. Theoretische Rahmung

Im folgenden Kapitel wird die vorliegende Arbeit in ihren theoretischen Kontext eingeordnet. Hierfür soll zunächst der Bereich des Verhaltens in der Schule umschrieben und auf die Thematik der emotionalen und sozialen Entwicklung von Kindern und Jugendlichen insgesamt eingegangen werden. In diesem Rahmen werden die für die vorliegende Arbeit bedeutsamen Verhaltensbereiche und das Gebiet der schulischen Verhaltensbeobachtung skizziert. Anschließend wird der Themenbereich der Verlaufsdiagnostik dargestellt, indem sowohl die Lernverlaufsdiagnostik als auch die Verhaltensverlaufsdiagnostik beschrieben werden. Danach werden der SDQ sowie das DBR als für die Arbeit bedeutsame Instrumente zur Erfassung

von Verhalten, Verhaltensproblemen und psychischen Störungen vorgestellt. Abschließend wird kurz auf die Testtheorie und Fragebogenkonstruktion nach Bühner eingegangen.

2.1 Verhalten in der Schule

Verhalten wird in der Pädagogik definiert als die Summe aller beobachtbaren oder messbaren Aktivitäten eines lebenden Organismus. Es ist ein prozesshafter Vorgang und tritt in der Regel als Reaktion auf einen oder mehrere Reize auf. Unterschieden wird hierbei zwischen offenem, direkt beobachtbarem Verhalten auf der einen und verdecktem Verhalten auf der anderen Seite. Zu ersterem zählen zum Beispiel Bewegungen, letzteres äußert sich in physiologischen Veränderungen auf einen Reiz hin und lässt sich nur mit Hilfe bestimmter Messverfahren erfassen. In den Begriff des Verhaltens können auch geistige Tätigkeiten wie das Denken einbezogen werden. Man unterscheidet außerdem verschiedene Kategorien von Verhaltensweisen wie angeborenes oder erworbenes Verhalten (Tenorth & Tippelt, 2007). Weicht das Verhalten von Kindern und Jugendlichen so von den gesellschaftlich geltenden Werten und Normen ab, dass die Interaktion beeinträchtigt ist und sonderpädagogisch-therapeutische Hilfe für eine Veränderung des Verhaltens benötigt wird, spricht man von einer Verhaltensstörung. Hierbei wird zwischen externalisierenden und internalisierenden Verhaltensstörungen sowie sozial unreifem und delinquentem Verhalten differenziert (Tenorth & Tippelt, 2007). Der Begriff der „Verhaltensstörung“ wird in der Literatur häufig synonym mit dem Begriff „Verhaltensauffälligkeit“ verwendet. Einige Autor_innen unterscheiden hinsichtlich des Schweregrades und sprechen bei leichten Abweichungen von Verhaltensauffälligkeiten, bei schwerwiegenderen Verstößen gegen Verhaltensnormen von Verhaltensstörungen. Außerdem können die Begriffe auf Basis der Betrachtungsperspektiven voneinander abgegrenzt werden. So wird ein Verhalten deshalb als auffällig wahrgenommen, weil es sich von bestimmten normativen Maßstäben absetzt. Im Rahmen dieser phänomenal-deskriptiven Definition tritt es unter ganz bestimmten Bedingungen auf, die zeitlich, räumlich oder auch situativ etwas zur Auslösung dieses Verhaltens und der damit verbundenen Auffälligkeit beitragen. Die Verhaltensstörung kann wiederum durch eine Störung des Regelkreises der Personen-Umwelt-Beziehung definiert werden, die durch ein Ungleichgewicht in der Interaktion einer Person mit ihrer Umwelt entsteht. Im Sinne dieser funktionalen Definition ist ein sogenanntes „störendes“ Verhalten als problemlösendes Verhalten zu betrachten, das von Differenzen zwischen der Person und ihrer Umwelt ausgelöst wird und dazu dient, die jeweilige Lebenslage zu bewältigen. Auch wenn dieses Verhalten Normen und Regeln missachtet, kann es für die Person sinnorientiert

sein (Seitz & Stein, 2010). Obgleich man abweichendes Verhalten nach der funktionalen oder der phänomenal-deskriptiven Definition beschreibt, steht es immer in Bezug zu den normativen Kriterien, denen es nicht entspricht. Mit diesen werden die Verhaltensweisen beziehungsweise die Interaktionsprozesse der Person mit ihrer Umwelt verglichen. Dazu zählen insbesondere Bezugssysteme, die sich auf statistische Normen, explizite Normen, sozio-kulturelle Normen oder persönliche normative Wertvorstellungen einzelner Personen beziehen (Seitz & Stein, 2010). Hinzukommen weitere subjektive normative Maßstäbe der Person selbst, die dazu führen können, dass sich die Person aufgrund ihres Verhaltens selbst als „auffällig“ oder „andersartig“ wahrnimmt, was mit einem hohen Leidensdruck verbunden sein kann. Aufgrund der verschiedenen Bedeutungsgehalte der normativen Kriterien kann Verhalten je nach Blickrichtung unterschiedlich eingestuft werden (Seitz & Stein, 2010). Verhaltensstörungen sind inhaltlich von anderen Auffälligkeiten abzugrenzen, die ebenfalls sonderpädagogischen Handlungsbedarf signalisieren. Verhaltensstörungen äußern sich auf der Persönlichkeitsebene und beziehen sich nicht ausschließlich auf das manifeste Verhalten einer Person, sondern zusätzlich auf das psychische Verhalten im weiteren Sinne und demnach auch auf das innere Erleben der Person. Deshalb kann auch von Verhaltens- und Erlebnisstörungen gesprochen werden (Seitz & Stein, 2010).

Im schulischen Kontext treten Verhaltensauffälligkeiten und Verhaltensstörungen durchaus nicht selten auf. Die Ursachen für das wahrgenommene Verhalten können verschiedenen Ursprungs sein. Neben Auslösern, die in der Person selbst begründet sind, zum Beispiel neurologischen Ursachen bei der Aufmerksamkeits-Defizit-Hyperaktivitäts-Störung, kann auch das schulische Umfeld mit der Klasse, den Lehrkräften oder der Schulorganisation zu auffälligem Verhalten führen (Schmischke, 2008). Im schulischen Kontext wird von auffälligem Verhalten gesprochen, wenn dieses nicht alters- und entwicklungsstandentsprechend ist, gegen soziale Normen und Werte verstößt und für das soziale Umfeld nicht akzeptabel ist, in verschiedenen Situationen auftritt (verschiedene Fächer, verschiedene Lehrpersonen, zu Hause, Pausensituationen), das Ausbilden oder Erhalten von sozialen Kontakten für das Kind erschwert oder verhindert, die Lernprozesse und Entwicklung des Schülers oder der Schülerin erschwert oder verhindert sowie das Umfeld des Kindes beeinträchtigt und somit auch andere Kindere behindert und das Unterrichten erschwert. Hinzukommt, dass die Verhaltensauffälligkeiten über einen längeren Zeitraum bestehen und nicht punktuell und kurzfristig auftreten (Schmischke, 2008).

Das Verhalten von Kindern und Jugendlichen im Kontext Schule spielt sowohl für den Lernerfolg als auch für das allgemeine Klassen- und Schulklima sowie für die Gesundheit der Lehrkraft eine entscheidende Rolle. Insbesondere das Emotionswissen und die Fähigkeit, Emotionen regulieren zu können, wirken sich auf die akademischen und schulischen Kompetenzen der Schüler_innen aus. Beim Emotionswissen handelt es sich um die Fähigkeit, sich der eigenen und fremden Emotionen bewusst zu sein, diese zu nutzen und emotionale Äußerungen zu verstehen, Emotionen von Gesichtern abzulesen und in den Kontext der auslösenden Situationen zu stellen sowie die kulturellen Normen zu kennen, die es in Bezug auf Emotionen und deren angemessenen Ausdruck im jeweiligen Bereich gibt. Die Emotionsregulation beschreibt die Fähigkeit, emotionale Erregung zu kontrollieren und den eigenen inneren und äußeren Zustand zu verwalten (Garner, 2010). Die genannten Bereiche haben unter anderem Einfluss auf die Beziehung zur Lehrkraft, was sich wiederum auf den Lernerfolg auswirken kann (Garner, 2010; Hattie, 2009).

In Prävalenzstudien aus den letzten Jahren und Jahrzehnten konnten enge Zusammenhänge zwischen Verhaltens- und Lernschwierigkeiten nachgewiesen werden. Das Risiko zeitgleichen Auftretens von Verhaltensauffälligkeiten ist bei vorliegenden Lernschwierigkeiten um das Eineinhalb- bis Dreifache erhöht (Hartmann, 2017). Es ist also davon auszugehen, dass Verhaltensstörungen und Lernschwierigkeiten einander bedingen. Die Auswirkungen von Lernschwierigkeiten auf das Verhalten können sich unter anderem durch ein ungünstiges Selbstkonzept in Folge häufiger Rückschläge oder den Vergleich mit leistungsstärkeren Schüler_innen zeigen. Andererseits können Verhaltensauffälligkeiten sich auch auf das Lernverhalten auswirken und dieses beeinträchtigen. Unkonzentriertheit, eine reduzierte Aufmerksamkeit oder die unzureichende soziale Integration in der Klasse können Gründe dafür sein, dass den betroffenen Schüler_innen das Lernen schwerfällt (Hartmann, 2017).

Verhaltensstörungen können sich neben den angesprochenen problematischen Verläufen im schulischen Bereich zudem auf die persönliche Entwicklung auswirken, indem sie sich manifestieren und im weiteren Verlauf in der Adoleszenz zu delinquentem Verhalten beziehungsweise psychiatrischen Störungen führen (Brezinka, 2003; Beelmann & Raabe, 2007; Ihle & Esser, 2008). Kindliche Verhaltensstörungen können in Form externalisierend-ausagierender Abweichungen als aggressives, impulsives, oppositionelles oder hyperaktives Verhalten hervortreten, sich aber auch in Form von Ängstlichkeit, Traurigkeit und depressiven Verstimmungen äußern, somit eher verdeckt ablaufen und als internalisierend-ängstliche Störungen

zu den emotionalen und entwicklungsspezifischen Störungen zählen. Auch sozial-delinquentes Verhalten wie Gewalttätigkeit, Zerstörung von Gegenständen oder Diebstahl kann im schulischen Kontext von Verhaltensauffälligkeiten auftreten (Schmischke, 2008). Außerdem können sozial-unreife Verhaltensweisen wie altersunangemessene Unaufmerksamkeit, Träumerei oder Spielen während des Unterrichts ebenfalls zu den Verhaltensstörungen zählen (Hensle, 1994). Verschiedene nationale und internationale Studien geben an, dass zwischen 10 % und 20 % der Kinder und Jugendlichen Verhaltensstörungen zeigen (Costello, Mustillo, Erkanli, Keeler & Angold, 2003; Ihle & Esser, 2008).

Die inklusive Beschulung von Schüler_innen mit Förderbedarf im Bereich der Emotionalen und Sozialen Entwicklung rückt die Notwendigkeit für Verhaltensdiagnostik im schulischen Kontext in den Fokus. Im Folgenden soll daher kurz auf die Inklusion allgemein und in Bezug auf Schüler_innen mit dem Förderschwerpunkt der Emotionalen und Sozialen Entwicklung eingegangen werden.

Entsprechend den Vereinbarungen der UN-Behindertenrechtskonvention werden Schüler_innen mit sonderpädagogischem Förderbedarf in Deutschland zunehmend inklusiv beschult (United Nations, 2006). Die jeweiligen Rahmenbedingungen zur Umsetzung der Inklusion werden auf Basis bundesweit einheitlicher Vorgaben der Ständigen Konferenz der Kultusminister von den einzelnen Bundesländern selbst organisiert (Gebhardt, Sälzer & Tretter, 2014). In Deutschland wurden im Jahr 2014 508.400 Schüler_innen mit sonderpädagogischem Förderbedarf unterrichtet (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2016). Beim größten Teil dieser Schüler_innen besteht Förderbedarf im Bereich Lernen, gefolgt von den Förderschwerpunkten Geistige Entwicklung und Emotionale und Soziale Entwicklung. Gebhardt (2015) konnte in einer empirischen Übersicht zum gemeinsamen Unterricht von Schüler_innen mit und ohne Förderbedarf nachweisen, dass inklusive Beschulung allen Schüler_innen erfolgreiches Lernen ermöglichen kann. Hierfür muss von der Erwartung abgerückt werden, alle Schüler_innen könnten im Gleichschritt lernen; stattdessen ist der Unterricht durch kooperative und offene Lernformen anzupassen (Gebhardt, 2015).

Im Jahr 2014 wurden insgesamt 81.675 Schüler_innen mit Bedarf in der Förderung der emotionalen und sozialen Entwicklung in Deutschland unterrichtet. Von ihnen besuchten im Rahmen von Integration und Inklusion mehr als 23.000 Regelschulen (ebd.). Laut der Ständigen

Konferenz der Kultusminister wird mittlerweile sogar die Hälfte der Schüler_innen im Bereich Emotionaler und Sozialer Entwicklung allgemeinbildend beschult (ebd.). Da diese Gruppe von Schüler_innen somit einen Großteil der inkludierten Kinder darstellt, erscheint auch an Regelschulen die Diagnose von Verhalten und Verhaltensverläufen/-entwicklungen im Rahmen der Überprüfung positiver oder negativer Veränderungen notwendig. Insbesondere vor dem Hintergrund des Zusammenhangs zwischen Verhaltensauffälligkeiten und Lernschwierigkeiten sollten solche Diagnosen durchgeführt und ausgewertet werden. Auf Grundlage der Ergebnisse sind Konsequenzen für eventuell bereits bestehende Förderangebote zu ziehen. Zudem sollte die Notwendigkeit einer Förderung im entsprechenden Verhaltensbereich diskutiert werden.

2.1.1 Emotionale und soziale Entwicklung von Kindern und Jugendlichen

Verhaltensstörungen und -auffälligkeiten treten in jeder Schulform und bei Schüler_innen aus jeder Altersstufe auf. Besonders häufig zu beobachten sind diese jedoch im Rahmen der Beschulung von Schüler_innen mit dem Förderschwerpunkt Emotionale und Soziale Entwicklung (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), 2000). Geprägt sind dieser Förderschwerpunkt und das Verhalten der nach den sonderpädagogischen Gesichtspunkten dieses Förderschwerpunktes unterrichteten Schüler_innen von den Wechselwirkungen zwischen Gesellschaft und Individuum sowie sozialem Umfeld und Persönlichkeitsentwicklung. Zusätzlich können Folgen von Entwicklungsstörungen, Krankheiten und Behinderungen die Probleme verstärken (KMK, 2000). In diesem Bereich stehen insbesondere soziale und emotionale Entwicklungsprozesse im Vordergrund, die von unterschiedlichen Faktoren beeinflusst werden können. Im Fokus stehen hierbei die persönliche, die familiäre, die schulische und die gesellschaftliche Ebene, welche eine signifikante Rolle bei der Ausbildung von Verhaltensstörungen und Verhaltensauffälligkeiten spielen können und dementsprechend den angesprochenen Förderschwerpunkt maßgeblich beeinflussen (KMK, 2000).

Schüler_innen mit Schwierigkeiten im sozialen und emotionalen Bereich erleben, wie die Bezeichnung des Förderschwerpunktes schon sagt, Probleme in der Ausbildung, Nutzung und Weiterentwicklung emotionaler und sozialer Kompetenzen. Dies geschieht unabhängig davon, ob die Schüler_innen einen attestierten Förderbedarf mit diesem Schwerpunkt haben, nicht

diagnostiziert sind oder lediglich durch verschiedene Verhaltensweisen aus dem Gebiet auffallen. Zu den Problemen zählen im emotionalen Bereich die emotionale Regulationsfähigkeit, die emotionale Bewusstheit, der adäquate emotionale Ausdruck eigener Gefühle, die emotionale Eindrucksfähigkeit für das Erleben anderer Personen, Probleme mit dem Selbstwertgefühl und dem Kontrollerleben. Die sozialen Kompetenzen der Schüler_innen können in Bezug auf Kommunikations-, Kooperations- und Verhandlungsfähigkeit, Konfliktbewältigung, soziale Sensibilität, Sachlichkeit, Fairness und Rücksicht, Toleranz und die Art und Weise der Selbstdarstellung eingeschränkt sein (Stein & Müller, 2015). Es kann zudem vorkommen, dass die Schüler_innen sich die zuletzt genannten Verhaltensbereiche und Kompetenzen zwar angeeignet haben, aber nicht wissen, wie sie diese zeigen, umsetzen und anwenden sollen. In diesem Fall kann von Performanzproblemen gesprochen werden (Stein & Müller, 2015).

Das Verhalten von Schüler_innen wird in der vorliegenden Arbeit in sechs Bereiche eingeteilt: Schulbezogenes Verhalten, Verhaltensprobleme, Hyperaktivität, Emotionale Probleme, Prosoziales Verhalten und Verhaltensprobleme mit Gleichaltrigen. Diese Verhaltensbereiche beinhalten sowohl externalisierende als auch internalisierende Arten von Verhalten. Sozial-delinquentes und sozial-unreifes Verhalten sind ebenfalls in diese Bereiche eingebunden. Im Folgenden sollen die einzelnen Verhaltensbereiche kurz vorgestellt werden.

Schulbezogenes Verhalten

Wie die Bezeichnung dieses Verhaltensbereichs bereits angibt, geht es beim schulbezogenen Verhalten um die Verhaltensweise eines Schülers oder einer Schülerin im Schul- und Klassenkontext. Störungen und Abweichungen in diesem Bereich treten vor allem dann auf, wenn grundlegende schulbezogene Regeln nicht beachtet werden. Hierzu zählen vor allem Regeln zum Umgang miteinander, zur Kommunikation und zur Konzentration im Unterricht. Außerdem enthalten sind die generelle Bereitschaft und Motivation, in adäquater Weise am Unterrichtsgeschehen teilzunehmen. Verstöße gegen die im Unterricht geltenden Regeln beziehungsweise Verhaltensauffälligkeiten in diesem Bereich führen zu Unterrichtsstörungen. Abweichungen vom schulbezogenen Verhalten und daraus resultierende Störungen sind in der Regel an die anderen Gruppenmitglieder, die Lehrkraft, die unterrichtlichen Rahmenbedingungen und weitere äußere Faktoren geknüpft und sind dementsprechend selten isoliert im störenden Schüler oder in der störenden Schülerin begründet (Schmischke, 2008a).

Verhaltensprobleme

Unter Verhaltensproblemen werden in der vorliegenden Arbeit insbesondere externalisierende Verhaltensauffälligkeiten gefasst. Hierzu zählen aggressives Verhalten, oppositionelles Verhalten sowie Norm- und Eigentumsverletzungen. Unterschieden wird hierbei in der Literatur zwischen destruktiven und non-destruktiven sowie zwischen verdeckten und offenen Formen (u.a. Hillenbrand & Melzer, 2017; Beelmann & Raabe, 2007). Die Verhaltensweisen werden außerdem eingeteilt in körperliche oder verbale, direkte oder indirekte, instrumentelle oder feindselige und proaktive oder reaktive Formen (Beelmann & Raabe, 2007). Im schulischen Kontext tritt solches Verhalten häufig in Form von Wut, geringer Frustrationstoleranz, körperlicher und verbaler Gewalt, Provokationen und Streit auf. Häufig ist ein solches Verhalten das Ergebnis verschiedener biologischer, psychischer und sozial-ökologischer Faktoren, welche sich bei ungünstigem Verlauf zu Risikofaktoren entwickeln und bereits im frühen Kindesalter das Verhalten der Schüler_innen prädisponieren können (Hillenbrand & Melzer 2017).

Hyperaktivität

In der Fachliteratur wird die Hyperaktivität als eine Störung des Sozialverhaltens aggressiver Art verstanden und tritt häufig in Verbindung mit dem Aufmerksamkeits-Defizit-Syndrom auf (Hellwig, 2010). Merkmale hyperaktiven Verhaltens sind unter anderem das Zappeln mit Händen und Füßen, übermäßige Lautstärke und ein starker Rededrang. Bezogen auf schulische Situationen kommen unangemessenes Aufstehen und Herumlaufen, Herumrutschen auf dem Stuhl, Schwierigkeiten ruhig zu spielen oder sich ruhig zu beschäftigen, Impulsivität, leichte Ablenkbarkeit und Unaufmerksamkeit hinzu. Diese Verhaltensweisen machen es den betroffenen Schüler_innen schwer, dem Unterrichtsgeschehen zu folgen sowie Aufgaben zu bearbeiten und fertigzustellen. Sie stellen eine erhebliche Beeinträchtigung des Unterrichtsablaufs dar (Hellwig, 2010). Die Gründe für das Auftreten hyperaktiver Verhaltensweisen sind bis heute nicht endgültig geklärt. Im Raum stehen Ursachen wie winzige Gehirnschäden, Allergien gegen bestimmte Lebensmittel, Alkoholkonsum und Rauchen in der Schwangerschaft. Auch ungesunde Ernährung oder das soziale und familiäre Umfeld können zur Entstehung von Hyperaktivität beitragen (Hellwig, 2010).

Emotionale Probleme

Zu den emotionalen Problemen werden in der vorliegenden Arbeit die internalisierenden Verhaltensstörungen gezählt. Hierbei handelt es sich um ängstliches, sozial unsicheres oder depressives Verhalten, wobei der Schweregrad und die Ausprägung des jeweiligen Verhaltens

stark variieren und von leichten, oberflächlichen Beeinträchtigungen bis hin zu manifesten Angststörungen, sozialen Phobien oder Depressionen reichen. Angstreaktionen lassen sich im schulischen Kontext in unterschiedlichen Bereichen erkennen. Auf der physiologischen Ebene kann Angst sich zum Beispiel in Form von Zittern, Schwitzen oder erhöhtem Pulsschlag zeigen. Dagegen können als motorische Reaktionen auf Angst beispielsweise Fluchtverhalten oder Hilfeschreie auftreten. Im schulischen Kontext können häufig im Ausdruck Formen der Angst beobachtet werden. So können eine leise Stimme, angespannte Mimik und Körpersprache hier auf Probleme im emotionalen Bereich hinweisen (Stein, 2017). Im Unterschied zu Phänomenen wie Aggressivität oder Hyperaktivität leidet die Umwelt der betroffenen Schüler_innen seltener unter ihren Schwierigkeiten. Dies kann dazu führen, dass diese emotionalen Probleme seltener auffallen beziehungsweise erkannt und erfasst werden (Stein, 2017).

Prosoziales Verhalten

Die wesentlichen Charakteristika prosozialen Verhaltens sind Freiwilligkeit, eine positive gesellschaftliche Bewertung des Verhaltens und die Ausrichtung des Verhaltens an den Bedürfnissen anderer. Zu prosozialem Verhalten zählen im schulischen Kontext die Bereiche Hilfeverhalten, kooperatives Verhalten und konstruktives Konfliktverhalten (Knopf & Gallschütz, 2006). Unter Hilfeverhalten kann der Kernbereich prosozialen Handelns verstanden werden. Definiert wird es durch die Intention, andere aus einer Notlage zu befreien. Diese Notlagen können sich im schulischen Kontext sehr unterschiedlich darstellen und von kleinen Schwierigkeiten wie einem vergessenen Radiergummi bis hin zu größeren Problemen wie schlechten Benotungen reichen. Das kooperative Verhalten wiederum zeichnet sich durch gemeinsames, partnerschaftliches oder gruppendynamisches Handeln aus, welches dem Zweck dient, der Gruppe einen maximalen gemeinsamen Gewinn zu bescheren. Dieser Gewinn muss nicht zwingend materiellen Ursprungs sein. Kooperatives Handeln, das häufig Hilfeverhalten inkludiert und ohne dies kaum möglich wäre, kann von Hilfeverhalten dennoch abgegrenzt werden, da ihm immer ein gemeinsam zu erreichendes Ziel zugrundeliegt. Trotzdem sind Überschneidungen in diesem Bereich häufig. Zum konstruktiven Konfliktverhalten, welches den dritten Bereich prosozialen Verhaltens beschreibt, gehört der konstruktive Umgang mit Konflikten, welche im schulischen Kontext des Öfteren insbesondere unter Schüler_innen mit Schwierigkeiten im emotionalen und sozialen Bereich auftreten. Dieser konstruktive Umgang mit Kon-

flikten setzt Kompetenzen wie Konfliktresistenz (das Aushalten von Konflikten), Konflikttoleranz (das Vermeiden von Konflikten durch Umbewertung) und Konfliktkompetenz (die schrittweise Lösung von Konflikten) voraus. (Knopf & Gallschütz, 2006).

Verhaltensprobleme mit Gleichaltrigen

In diesen Bereich können sowohl externalisierende als auch internalisierende Störungsbilder fallen, da sowohl durch extrovertiertes, etwa aggressives, als auch durch zurückgezogenes, introvertiertes Verhalten Probleme im Umgang mit Gleichaltrigen entstehen können. Zu den Verhaltensproblemen mit Gleichaltrigen werden Auffälligkeiten in der Kommunikation, in der Kooperation und im allgemeinen Umgang einer Schülerin oder eines Schülers mit seinen oder ihren Mitschülerinnen und Mitschülern gezählt. Der hierbei entscheidende Faktor ist, dass die Lernenden in etwa das gleiche Alter haben. Bevorzugt eine Schülerin oder ein Schüler beispielsweise die Arbeit mit Erwachsenen, zieht sich aus gemeinsamen Spielen mit den Klassenkamerad_innen zurück oder wird von ihnen gehänselt, schikaniert oder provoziert, kann dies für die genannten Verhaltensprobleme sprechen. Diese können durch sozial unreifes Verhalten oder fehlende soziale Kompetenzen hervorgerufen werden, aber auch durch andere, oben bereits aufgeführte Verhaltensstörungen und -auffälligkeiten, die ein Miteinander erschweren. Zu nennen sind hier insbesondere aggressives und ängstliches Verhalten (u.a. Hellwig, 2010; Stein, 2017).

2.1.2 Verhaltensbeobachtung in der Schule

Allgemein gesprochen dienen Beobachtungen der Datenerhebung und Diagnostik. Durch das gezielte Wahrnehmen von Situationen, Vorgängen und Ergebnissen und die Hinterfragung ihrer Bedeutung sind diese zu verstehen. In der Schule erfüllt die Beobachtung mehrere Funktionen, zu denen unter anderem die Beurteilung von Leistungen der Schüler_innen, die Analyse von Lehr- und Lernprozessen und die Nutzung von Beobachtungen als Instrument der Förderdiagnostik zu zählen sind. Es wird zwischen verschiedenen Formen von Beobachtungen differenziert. Die teilnehmende Beobachtung ist im schulischen Kontext weit verbreitet. Hierbei nimmt die beobachtende Person am Unterrichtsgeschehen oder Schulalltag teil, während sie ihrem Beobachtungsauftrag nachkommt. Eine weitere gängige Beobachtungsform ist die Selbstbeobachtung. Sie kann von der Lehrkraft unter anderem der Evaluation des eigenen Unterrichts dienen, aber auch von Schüler_innen beispielsweise im Kontext einer Selbstbeurtei-

lung oder der Einschätzung von Gruppenarbeitsprozessen eingesetzt werden. Auch die Fremdbeobachtung durch eine dritte, möglichst unabhängige Person wird häufig in der Schule durchgeführt (Schmischke, 2008).

Insbesondere in Bezug auf das Verhalten von Schüler_innen im schulischen Kontext und die unterschiedlichen Ausprägungen sowie verschiedenen Ursachen und vielfältigen Konsequenzen, die mit den verbundenen Verhaltensauffälligkeiten und -störungen einhergehen, ist eine differenzierte Vorgehensweise erforderlich. Zur Eingrenzung der Bereiche, in denen die Störungen auftreten können, und aufgrund der oben bereits erwähnten engen Verknüpfung von Lern- und Verhaltensschwierigkeiten bieten sich im schulischen Kontext Verhaltensbeobachtungen als ein erster Schritt in Richtung Diagnostik an (Schmischke, 2008; Hartmann, 2017).

Im Allgemeinen sind Beobachtungen nur dann gewinnbringend, zielführend und effektiv, wenn sie geplant sind und mit einer bestimmten Absicht durchgeführt werden. Hierzu müssen die Kriterien zielgerichtet und sachlich erfüllt werden. Zudem ist im Vorhinein eine geeignete Methode auszuwählen. Das Kriterium der Zielgerichtetheit erfordert vor der eigentlichen Beobachtung, die Klärung einiger Fragen in Bezug auf die Beobachtungsperson, die Intention und das eigentliche Ziel wie zum Beispiel eine anschließende Beratung, die Anpassung des Förderplans oder die Umstrukturierung von Lernsettings. Um den Faktor der Sachlichkeit zu erfüllen, muss die beobachtende Person sich ihrer eigenen Erwartungen, Einstellungen und Interessen bewusst sein, da diese möglichst wenig Einfluss auf die eigentliche Beobachtung haben sollen. Insbesondere Beobachtungen, die von der Lehrkraft selbst durchgeführt werden, müssen immer im Kontext der von der Lehrkraft eingenommenen Rolle gesehen werden. Diese ist in der Regel zweigeteilt, da die beobachtende Person gleichzeitig den strukturierenden und unterrichtenden Teil des Unterrichtsgeschehens vorgibt und deshalb nicht vollkommen neutral der zu beobachtenden Situation gegenüberstehen kann. Vor der Beobachtung sollten außerdem verschiedene Methoden gesichtet und auf ihre Eignung in der jeweiligen Situation geprüft werden. Screenings, Testverfahren, Beobachtungsbögen und ähnliche Methoden können je nach Intention gewählt werden. Die beobachtende Person sollte sich im Vorhinein mit dem jeweiligen Instrument vertraut machen, um die Beobachtungssituation zu entlasten und den Fokus ganz auf das zu Beobachtende legen zu können. Außerdem ist zu überlegen, in welcher Form die Beobachtungen festgehalten werden. Denkbar sind dazu etwa eine direkte

Dokumentation im gewählten Beobachtungsmaterial, eine Mitschrift oder ein Protokoll (Schmischke, 2008).

Verhaltensbeobachtungen in der Schule können entweder systematisch oder unsystematisch durchgeführt werden. Eine unsystematische Verhaltensbeobachtung lässt dem Beobachter einen relativ großen Ermessensspielraum, da die Auswahl von Ereignissen, die der beobachtenden Person relevant erscheinen, nach subjektiven Kriterien erfolgt. In Bezug auf die Erfüllung der Gütekriterien, die für die Verhaltensverlaufsdiagnostik relevant sind und auf die im Verlauf der Arbeit und insbesondere in den Kapiteln 2.2.2 und 2.4 eingegangen wird, liefert die unsystematische Verhaltensbeobachtung wenig überzeugende Ergebnisse. So erfüllt sie laut Casale, Hennemann, Huber und Grosche (2015) nur eines der von den Autoren geforderten elf Gütekriterien (Direktheit, siehe Kapitel 2.2.2) zur Verhaltensverlaufsdiagnostik. Ein deutlich besseres Ergebnis erzielt hier die systematische Verhaltensbeobachtung, welche zehn von elf möglichen Gütekriterien erfüllt. Unter einer systematischen Verhaltensbeobachtung wird eine Art der Observierung verstanden, die der beobachtenden Person genau vorgibt, worauf geachtet werden soll und in welcher Form die Ergebnisse zu protokollieren sind. Für die Dokumentation der Verhaltensweisen und deren Auftreten werden in der Regel Strichlisten in Protokollbögen und Time-Sampling-Methoden genutzt.

Auch die Verhaltensbeurteilung mittels Ratingskalen schneidet bei der Erfüllung von Gütekriterien zur Verhaltensverlaufsdiagnostik gut ab. Sie erreicht acht von elf Gütekriterien und damit Platz drei der geeignetsten Verfahren zur diagnostischen Erfassung von Schülerverhalten im Bereich der Verlaufsdiagnostik. Die Verhaltensbeurteilung mit Hilfe von Ratingskalen erfolgt durch das retrospektive Einschätzen eines beobachteten Verhaltens über einen längeren Zeitraum. Dies geschieht in der Regel durch reflektive Items, die zur Abbildung eines bestimmten Konstrukts in Ratingskalen abgebildet werden (Casale et al., 2015).

Ein weiteres im Rahmen von Verhaltensbeobachtungen im verhaltensverlaufsdiagnostischen Kontext zu nennendes Verfahren ist die Alltagsbeobachtung. Diese schneidet bei der Überprüfung auf die Eignung für die Verhaltensverlaufsdiagnostik jedoch so schlecht ab, dass eine weitere Verwendung in diesem Kontext wenig Sinn macht. Die Alltagsbeobachtung erfüllt nur eines der elf Gütekriterien zur Verlaufsdiagnostik im Bereich des Schülerverhaltens, nämlich das der Ökonomie, und wird deshalb in der vorliegenden Arbeit nicht weiter berücksichtigt (Casale et al., 2015).

2.2 Verlaufsdiagnostik

In diesem Kapitel wird das Konzept der Verlaufsdiagnostik skizziert und von dem der Statusdiagnostik abgegrenzt. Zur weiteren Spezifizierung sollen im Anschluss an diese eher allgemeine Einführung in das Thema Verlaufsdiagnostik die beiden Bereiche Lern- und Verhaltensverlaufsdiagnostik beschrieben und Gemeinsamkeiten sowie Unterschiede herausgearbeitet werden. Aktuelle Diskussionen zum Thema sollen zudem summarisch einbezogen werden.

Die im Kontext Schule immer noch gängige Statusdiagnostik (auch Platzierungsdiagnostik) steht seit einigen Jahren vermehrt unter Kritik. Dies liegt unter anderem daran, dass diese Art der Diagnostik versucht, zu einem bestimmten Zeitpunkt mit Hilfe eines Testverfahrens eine prognostische Aussage über weitere Entwicklungsverläufe vorzunehmen. Diese zeitlich punktuell durchgeführten Tests und Überprüfungen, die zu einem bestimmten Zeitpunkt der Entwicklung einer Schülerin oder eines Schülers angewandt werden, bilden lediglich den tagesaktuellen Entwicklungsstand zur Datenerhebung ab und geben nicht, wie manchmal erwartet, erhofft oder angenommen, Auskunft über sinnvolle Förderung, die angeschlossen werden müsste, um eventuelle Entwicklungsrückstände aufzuholen (Hartmann, 2017). Zudem wird der Umgang mit den durch statusdiagnostische Verfahren erhobenen Daten diskutiert. Dieser muss sensibel gestaltet werden, da den Schüler_innen auf Grundlage der erfassten Daten eine Weiterentwicklung ermöglicht werden soll. Dies setzt allerdings voraus, dass die Diagnostik benutzt wird, um Hinweise für die Unterrichts- und Förderplanung sowie pädagogisch-therapeutische Maßnahmen zu erhalten (Hartmann, 2017). Die Statusdiagnostik wirkt in Bezug auf die Schlüsse, Konsequenzen und Aussagen, die auf Grundlage der Ergebnisse getroffen werden sollen, wenig überzeugend und nicht ausreichend. Eine Prognose auf Grundlage einer Diagnostik zu treffen, die lediglich eine Aussage über den Lernstand oder das Verhalten zu einem einzigen Zeitpunkt trifft, sich aber eventuell auf signifikante Entscheidungen wie beispielsweise die Zuweisung zu einer Schulform auswirkt und damit auf die gesamte Schullaufbahn Einfluss nehmen kann, erscheint wenig angemessen (Hartmann, 2017).

Aus den angeführten Gründen werden statusdiagnostische Wege der Datenerhebung derzeit vermehrt von Maßnahmen zur Verlaufsdiagnostik abgelöst. Dies geschieht sowohl im Bereich des Lernens als auch im emotionalen und sozialen Bereich des Verhaltens, der für diese Arbeit von besonderer Bedeutung ist. Verlaufsdiagnostik wird charakterisiert durch den regelmäßigen Einsatz fundierter Erhebungsmethoden, die ökonomisch, das heißt ohne Überforderung

der beteiligten Personen (Lehrkräfte, Lernende), durchgeführt werden und nach der Auswertung einen Verlauf zeigen sollen. Der in der Literatur bereits ausführlich evaluierte Response-to-Intervention-Ansatz (im Folgenden RTI) gibt beispielsweise vor, dreimal jährlich die Daten zu erheben und für Bereiche mit auffälligen Werten häufigere Messungen vorzunehmen. Je nach Bereich und Zielstellung kann dies einmal wöchentlich oder in manchen Fällen auch täglich der Fall sein (Grosche & Volpe, 2013).

Unter dem Response-to-Intervention-Ansatz wird eine Art konzeptueller Rahmen verstanden, in dem drei grundlegende Kernbereiche zu einem präventiv und inklusiv ausgerichteten Konzept kombiniert werden. Im Zentrum dieses auf die inklusive Beschulung von Schüler_innen ausgerichteten Konzeptes stehen nach Intensität gestufte Förderebenen, die als sogenannte Mehrebenenprävention die Prävention von Lern- und Verhaltensschwierigkeiten verfolgen. Den Förderebenen liegen datengeleitete, auf die Ergebnisse von Screenings und Lernverlaufsdokumentationen zurückzuführende Förderentscheidungen sowie evidenzbasierte Lehr- und Fördermethoden zugrunde (Blumenthal, 2017). Das Hauptmerkmal des RTI-Ansatzes ist die kontinuierliche Überprüfung der Frage, ob alle Schüler_innen den Fördermethoden entsprechende Lern- und Entwicklungsfortschritte machen. Anhand dieses Hauptmerkmals kann auch die Bezeichnung „Response to Intervention“ aufgeschlüsselt werden. So sind die von den die Schüler_innen erzielten Fortschritte die „response“, die durch die Förderung, also die „intervention“, hervorgerufen wird. Durch diagnostische Verfahren werden die Erfolge und damit die Passung zwischen Lernenden und Fördermaßnahmen beziehungsweise Unterricht überprüft. Stellt sich der erwartete Fortschritt bei einem Kind ein, wird es als „responder“ gesehen. Gibt es keinen nennenswerten Fortschritt zu verzeichnen, ist das Kind ein „non-responder“. In diesem Fall muss von pädagogischer Seite her eine Verbesserung der Fördermethode geschaffen werden (Blumenthal, 2017). Die regelmäßige Messung des Lernfortschritts, auch Lernverlaufsdiagnostik genannt, wird in Kapitel 2.2.1 näher erklärt.

Wie oben erwähnt, handelt es sich beim RTI-Ansatz um eine Art der Mehrebenenprävention. Er ist in drei pyramidenartig angeordnete Förderebenen aufgeteilt, beginnend mit Förderebene I, dem hochwertigen inklusionsorientierten Unterricht. Wenn dieser Unterricht allein nicht ausreicht, um einen Lernzuwachs zu ermöglichen, muss die Lehrkraft zu Förderebene II übergehen und im Rahmen qualifizierten Förderunterrichts evidenzbasiert in Kleingruppen för-

dern. Sollte auch diese Ebene keine zufriedenstellenden Fortschritte beim Kind erzielen können, wird Förderebene III erreicht. Die Regelschullehrkraft gibt dann an eine sonderpädagogische Lehrkraft ab, welche durch spezifische, intensive Förderung und/oder zieldifferente Förderung den Unterricht, zumindest teilweise, übernimmt. Bereits auf Ebene II kann die Regelschullehrkraft von einer Sonderpädagogin oder einem Sonderpädagogen unterstützt werden. Die erste Ebene liegt gewöhnlich in der Verantwortung der Regelschullehrkraft (Blumenthal, 2017). Der Response-to-Intervention-Ansatz wurde für verschiedene Bereiche aus dem schulischen Kontext adaptiert, zum Beispiel für Fächer wie Deutsch oder Mathematik, aber auch für Förderbereiche wie Sprachliche Entwicklung oder Emotionale und Soziale Entwicklung. Wie der RTI-Ansatz für den Bereich der Emotionalen und Sozialen Entwicklung aufgebaut ist, soll in Kapitel 2.2.2 aufgegriffen werden.

Im Vergleich zu den USA begann die Entwicklung der Lernverlaufsdiagnostik im deutschsprachigen Raum erst relativ spät. In den USA wurde bereits in den 1960er Jahren über die Unterscheidung von formativer und summativer Evaluation diskutiert (Klauer, 2014). Der Begriff der formativen Evaluation bezeichnet die in Abständen wiederholte Evaluation eines gerade stattfindenden Lernprozesses. Formen der summativen Evaluation hingegen evaluieren die Ergebnisse am Ende eines Lernvorgangs. In Deutschland fanden lange Zeit keine derartigen Diskussionen über die genannten Evaluationsformen statt. Auch das amerikanische „Curriculum Based Measurement“ wurde am Anfang weder von der allgemeinen Pädagogik noch von der Sonderpädagogik wahrgenommen. Hierunter wird nicht der Einsatz üblicher standardisierter schulischer Leistungstests verstanden, sondern eine Art von Leistungstest, die überprüft, was im aktuellen Unterricht durchgenommen wurde und dementsprechend ermittelt, wie gut das im aktuellen Zeitraum (z. B. eine Woche) bearbeitete Teilziel des Unterrichts von den einzelnen Schüler_innen erreicht wurde (Klauer, 2014). Erst in Folge der Veröffentlichung eines Artikels von Klauer im Jahr 2006, welcher die Entwicklung des Curriculum Based Measurements in den USA vorstellte und positiv kommentierte, fand diese auch in Deutschland Beachtung. Anknüpfend an Klauers Artikel, widmeten sich weitere Autorinnen und Autoren dem Thema und veröffentlichten in den Folgejahren erste Studien. Die Befunde fielen in den meisten Fällen sehr positiv aus, wobei sich im Bereich der Rechtschreibung aus testtheoretischer Sicht weniger erfreuliche Ergebnisse ergaben (Klauer, 2014).

Auch die Begrifflichkeiten wurden intensiv diskutiert und haben sich mehrfach gewandelt. An die Stelle der Bezeichnung „Curriculum Based Measurement“ beziehungsweise die Abkürzung „CBM“ trat in Klauers Artikel von 2006 der Begriff „Lernfortschrittsmessung“, der auch von anderen Autor_innen aufgenommen wurde. Da die Forschung jedoch ergab, dass sich keinesfalls immer ein tatsächlicher Fortschritt einstellen muss, sondern auch Stagnation oder sogar Rückschritte im Lernprozess auftreten können, wurde letztlich der ebenfalls in der vorliegenden Arbeit verwendete Begriff „Lernverlaufsdiagnostik“ als angemessener angesehen (Klauer, 2014).

Differenzen zwischen Status- und Verlaufsdiagnostik finden sich in verschiedenen Bereichen. Wie bereits kurz erwähnt, divergieren die beiden Verfahren unter anderem in ihrer jeweiligen Zielsetzung. Die Statusdiagnostik verfolgt in der Regel das Ziel einer normorientierten Einordnung der Testleistung, während in der Verlaufsdiagnostik der Fokus auf der Analyse individueller Veränderungen von Merkmalen liegt. Auch im Umfang weichen die Verfahren voneinander ab. Statusdiagnostische Instrumente werden eher selten und dafür sehr intensiv angewendet. Im Bereich der Verlaufsdiagnostik finden die Erhebungen, Beobachtungen oder Beurteilungen häufiger statt und benötigen weniger Zeit für die einzelne Durchführung. Das zu messende Konstrukt ist bei der Statusdiagnostik eher breit gefasst, wohingegen ein enger gesteckter Bereich bei der verhaltensverlaufsdiagnostischen Überprüfung eine Rolle spielt. Ein weiterer wichtiger Faktor ist die Art der Testkonstruktion. Während in der Statusdiagnostik mit änderungsresistenten Verfahren gearbeitet wird, nutzt die Verlaufsdiagnostik änderungssensible beziehungsweise -sensitive Instrumente, um die Merkmalsänderungen präzise erfassen und abbilden zu können (Casale et al., 2015b).

2.2.1 Lernverlaufsdiagnostik

Auf den Bereich der Lernverlaufsdiagnostik soll hier nur vergleichsweise kurz eingegangen werden, da der Hauptfokus der vorliegenden Arbeit auf der Verhaltensverlaufsdiagnostik liegt.

Bei der Lernverlaufsdiagnostik handelt es sich, wie oben angesprochen, um eine Variante der formativen Evaluation, welche sich mit der wiederholten Messung ein und derselben Kompetenz befasst. Im Unterschied zur summativen Diagnostik verfolgt die formative Lernverlaufsdiagnostik das Ziel der Anpassung beziehungsweise der Modifikation laufender Förderungen. Die im Laufe der Messungen erhaltenen Rückmeldungen können Rückschlüsse zum Lehr-

Lernprozess und für weitere notwendige Fördermaßnahmen geben. Es geht bei der Lernverlaufsdiagnostik also nicht ausschließlich um das Erfassen der Daten von Schüler_innen, sondern darüber hinaus auch um die Optimierung von Unterricht und Förderung (Hartke, Sikora & Voß, 2017).

Hartke et al. (2017) fassen einige Aspekte zusammen, die von im Rahmen der Lernverlaufsdiagnostik benutzten Instrumenten erfüllt werden müssen. Hierzu zählen die Eignung eines verlaufsdiagnostischen Verfahrens zur Erfassung des Leistungsstandes von Schüler_innen und die hiermit verbundenen Gütekriterien, welche gegeben sein müssen. Zu diesen gehören die Hauptgütekriterien der klassischen Testtheorie, die auch bei statusdiagnostischen Instrumenten gewährleistet sein müssen: Objektivität, Reliabilität und Validität. Diese sollen Wahrnehmungs- und Beurteilungsfehler zu einem größtmöglichen Grad reduzieren (Hartke et al., 2017). Außerdem müssen die Messungen wiederholt, das heißt an mindestens zwei Zeitpunkten, durchgeführt werden. Der Aufforderungscharakter muss hierbei gleichbleiben. Als letzten Aspekt führen Hartke et al. (2017) die Praktikabilität und Anwendbarkeit eines Dokuments zur Lernverlaufsdiagnostik auf. Aufgrund des wiederholten Einsatzes der Tests im Verlauf des Schuljahres sollten diese ökonomisch in Durchführung und Auswertung sein und trotzdem einen möglichst hohen Erkenntnisgewinn liefern.

Wie bereits erwähnt, bietet es sich an, zur Messung des Lernverlaufs der Schüler_innen und zur Überwachung der damit verbundenen Entwicklung von Kompetenzen über längere Zeiträume hinweg ebendiese Kompetenzen immer wieder zu überprüfen und die Ergebnisse zu dokumentieren. Für eine solche Testung werden im Normalfall Paralleltests eingesetzt. Dieses Vorgehen wäre aber im Rahmen einer Lernverlaufsdiagnostik zu aufwendig und würde dem Aspekt der Ökonomie nicht gerecht, da selbst bei einer Überprüfung im 14-tägigen Rhythmus pro Schuljahr ungefähr 20 im Vorhinein zu entwickelnde Paralleltests anstünden. Diese Entwicklung wird in der Regel nicht geleistet; stattdessen wird entweder auf bereits bestehende Paralleltests zurückgegriffen oder aber die Möglichkeit herangezogen, das Lehrziel und die damit verbundene Kompetenz über die Aufgabenmenge zu definieren, zu der die entsprechende Kompetenz befähigt. Nachdem diese definiert ist, können hieraus repräsentative Stichproben gezogen werden, zum Beispiel durch die zufallsgesteuerte Generierung von Aufgabenstichproben (Klauer, 2014). Dieses Vorgehen erscheint in der Praxis auf den ersten Blick jedoch etwas ungewöhnlich, da über den gesamten Zeitraum, in dem die Lernverlaufsdiagnostik

durchgeführt werden soll, auch auf Kompetenzen hin überprüft wird, die erst am Ende dieses Zeitraums beherrscht werden sollen. Das bedeutet, dass anfänglich Aufgaben enthalten sind, die von den Schüler_innen noch nicht gelöst werden können. Dies muss im Vorhinein kommuniziert werden, um Verwirrung und Frustration möglichst gering zu halten (Klauer, 2014).

Bei der Verwendung unterschiedlich schwerer Tests zur Lernverlaufsdiagnostik könnte man automatisch Leistungsverbesserungen oder -verschlechterungen der Schüler_innen verzeichnen, obwohl diese sich eventuell gar nicht tatsächlich verändert haben. Deshalb müssen zur Lernverlaufsdiagnostik eingesetzte Instrumente gewährleisten, dass die einzelnen Tests das Gleiche erfassen und außerdem den gleichen Schwierigkeitsgrad besitzen. Trotzdem darf ein Test nicht mehrfach verwendet werden. Dies wäre zwar ökonomisch und stellte auch in Bezug auf die Testschwierigkeit auf den ersten Blick eine Lösung da, allerdings wäre der Test aufgrund der Wiederholung automatisch bei der zweiten Durchführung leichter als bei der ersten. Zudem stellt sich die Frage, ob der Test beim zweiten Mal immer noch das Gleiche mäße wie im ersten Durchgang. Es müssen also immer neue, gleich schwierige Tests entwickelt werden, welche die gleichen Kompetenzen abfragen. Eine Homogenität in Bezug auf die Testschwierigkeit ist dementsprechend von zentraler Bedeutung für die Lernverlaufsdiagnostik (Klauer, 2014).

Der Nachweis der homogenen Schwierigkeit aufeinanderfolgender Tests erscheint problematisch, da man bei den Schüler_innen von einem sukzessiven Lernzuwachs ausgeht und Tests mit objektiv gleichem Schwierigkeitsgrad dementsprechend mit der Zeit einfacher würden. Je stärker sich die Kompetenzen der Schüler_innen also verbesserten, desto einfacher würde der Test. Einige Autorinnen und Autoren nutzen deshalb immer nur zwei aufeinanderfolgende Tests, um diese auf homogene Schwierigkeit zu überprüfen. Da der Lernzuwachs in diesem in der Regel kurzen Zeitfenster meist relativ gering ist, hat sich dieses Vorgehen bewährt (Klauer, 2014).

In Bezug auf die Itemschwierigkeit schlagen Wilbert und Linnemann (2011) vor, die exakte Schwierigkeit eines Tests im Vorhinein zu bestimmen und bei Tests mit unterschiedlichen Anforderungsniveaus einen Korrekturparameter zu definieren. Das würde bedeuten, dass bei unterschiedlichen Items eine unterschiedliche Gewichtung vorgenommen würde. Für ein solches Vorgehen wäre es allerdings notwendig, die Schwierigkeit des Tests oder eines einzelnen

Items im Vorhinein unabhängig zu bestimmen. Hierzu müsste auf die Probabilistische Testtheorie zurückgegriffen werden, da diese, anders als die Klassische Testtheorie, das Ermitteln von Itemparametern unabhängig von den Personenparametern zulässt. Begründet wird dies durch das Beruhen auf einer geschätzten Wahrscheinlichkeit zur Erreichung eines bestimmten Punktwertes in einem Test.

Um die Lernverlaufsdiagnostik möglichst effektiv zu gestalten, müssen die Tests sowohl Lernfortschritte beziehungsweise Kompetenzzuwachs als auch Lernrückschritte beziehungsweise Kompetenzverlust sensibel abbilden können. Zur Überprüfung der Änderungssensibilität eines Verfahrens bietet sich ein Zwei-Gruppen-Versuchsplan an. Hierbei sollen eine Fördergruppe und eine Kontrollgruppe ohne Förderung mit Prä- und Posttests überprüft werden. Bei der Auswertung sollte ein Lernzuwachs bei der Fördergruppe festgestellt werden. Dies spricht zum einen für die Art und Weise, in der die Interventionen und dementsprechend die Förderung stattgefunden haben, zum anderen für die in diesem Kontext deutlich relevantere Möglichkeit, mit Hilfe des Tests Änderungen in diesem Bereich zu diagnostizieren und darzustellen (Klauer, 2014).

Klauer (2011) unterscheidet zwischen zwei Formen der Lernverlaufsdiagnostik. In der Schule geht es in der Regel entweder darum, Gelerntes zu verbessern, es schneller anwenden zu können beziehungsweise bei der Bearbeitung bereits bekannter Aufgabenmodelle weniger Fehler zu machen oder aber darum, sein Wissen zu erweitern, Neues dazu zu lernen und sich neue Bereiche anzueignen. Ersteres lässt sich durch Tests überprüfen, die erfassen sollen, ob die Übung einer bereits beherrschten Kompetenz dazu führt, dass die Aufgaben schneller oder fehlerärmer bewältigt werden können. Ob die Erweiterung von Wissen tatsächlich stattfindet, kann von Tests gemessen werden, die auf eben diesen Wissens- und Kompetenzzuwachs ausgelegt sind. Für letzteres werden Verfahren benötigt, die bereits am Anfang alle Kompetenzen enthalten, die am Ende von den Schüler_innen beherrscht werden sollen. Wie oben bereits angesprochen, muss dies transparent kommuniziert werden, um Frustration auf Seiten der Schülerschaft zu vermeiden. Tests zur Verbesserung von Geschwindigkeit und/oder Genauigkeit werden häufig im sprachlichen Kontext oder aber bei Aufmerksamkeitstrainings angewendet. Im Gegensatz dazu können Verfahren, welche die Erweiterung einer Kompetenz um neue Bereiche erfassen, häufig in der Mathematik gefunden werden (Klauer, 2011)

2.2.2 Verhaltensverlaufsdiagnostik

Im Bereich der Verhaltensverlaufsdiagnostik sind die Entwicklungen noch nicht im gleichen Maß vorangeschritten wie in der Lernverlaufsdiagnostik. Auch die Verhaltensverlaufsdiagnostik kommt, wie die Lernverlaufsdiagnostik, aus dem englischsprachigen Raum und entwickelte sich um das Jahr 2000 (Grosche & Volpe, 2013). Sie wird häufig unter dem Namen Direct Behavior Rating (DBR) oder der deutschen Übersetzung Direkte Verhaltensbeurteilung geführt, auf die in Kapitel 2.4 näher eingegangen wird.

Die Notwendigkeit verhaltensverlaufsdiagnostischer Erfassungen im Schulalltag ergibt sich insbesondere aus der immer größer werdenden Heterogenität der Lernenden, die unter anderem durch Inklusion hervorgerufen wird. Eine solche Heterogenität bedeutet eine erhebliche Herausforderung für bereits bestehende standardisierte forschungsmethodische Designs zur Überprüfung der Effektivität von Fördermethoden im Bereich des Verhaltens. Traditionelle Verfahren im Bereich der evidenzbasierten Forschung sind in ihrer Durchführung sehr zeitintensiv und dauern in der Regel mehrere Monate. Außerdem sagen diese durchaus qualitativ hochwertigen Studien nichts über die Wirkung der untersuchten Fördermethode im Einzelfall aus. Aus diesem Grund müssen neue, angepasste wissenschaftliche Standards entwickelt werden, die im Bereich der evidenzbasierten (Sonder-)Pädagogik den Bedarf an Evidenzbasierung im Einzelfall abdecken (Casale, Grosche & Hennemann, 2015a). Auch der in Kapitel 2.1 erwähnte Zusammenhang zwischen Lernschwierigkeiten und Verhaltensauffälligkeiten unterstreicht die Notwendigkeit von Verhaltensverlaufsdiagnostik in schulischen Settings. Da Lernschwierigkeiten und Verhaltensauffälligkeiten häufig gemeinsam auftreten, einander bedingen und miteinander interagieren, ist die Entwicklung diagnostischer Verfahren zur Überprüfung dieser Bereiche unerlässlich (Hartmann, 2017).

Wie oben bereits ausgeführt (siehe Kapitel 2.2), geht der Response-to-Intervention-Ansatz exakt in diese Richtung und zielt auf eine engmaschige Überprüfung der Lern- und Entwicklungsfortschritte der Schüler_innen sowie die dementsprechende Anpassung der Fördermethoden ab. Auch für den Bereich der Emotionalen und Sozialen Entwicklung beziehungsweise für den Bereich des Verhaltens gibt es bereits sogenannte Mehrebenenkonzepte zur Förderung der emotionalen und sozialen Entwicklung im Rahmen von Response-to-Intervention-Ansätzen. Auch diese verhaltensbezogenen Modelle bestehen, ähnlich wie RTI-Ansätze im allgemeinen Kontext, aus drei Ebenen, welche durch regelmäßige Diagnostik in ihrer Effektivität

überprüft und evaluiert sowie im Anschluss adäquat modifiziert werden müssen (Blumenthal & Marten, 2017). Die erste Ebene sieht die Beschulung im allgemeinen Klassenunterricht vor. Ergibt die kontinuierliche Überprüfung der Fortschritte, dass diese Form der Förderung nicht effektiv ist, greift die für die zweite Ebene vorgesehene Förderung. Diese ist im Bereich des Verhaltens als unterrichtsintegrierte Verhaltensförderung definiert und umfasst 49 Handlungsmöglichkeiten (siehe hierzu Blumenthal & Marten, 2017), die von der Regelschullehrkraft gegebenenfalls in Kooperation mit einer sonderpädagogischen Lehrkraft angewendet werden. Weisen erneute Überprüfungen weiterhin auf nicht ausreichende Fortschritte oder sogar Rückschritte hin, wird der Schüler oder die Schülerin Ebene drei zugeordnet, die eine Intensivförderung mit Trainingsprogramm vorsieht. Hierfür ist ein Sonderpädagoge oder eine Sonderpädagogin verantwortlich (Blumenthal & Marten, 2017). Der Response-to-Intervention-Ansatz bietet folglich eine Möglichkeit, auf Basis kontinuierlicher Überprüfungen der erzielten Fortschritte die Verhaltensentwicklung der Schüler_innen zu fördern. Ein Hauptaugenmerk liegt dabei auf der Verhaltensverlaufsdiagnostik, die in diesem Kapitel erläutert wird.

Bei der Verhaltensverlaufsdiagnostik erfolgt mit Hilfe veränderungssensitiver Instrumente eine regelmäßige Beobachtung und Beurteilung spezifischer Verhaltensweisen. Auf Basis dieser Maßnahmen können Entwicklungsveränderungen im Bereich des untersuchten Verhaltens erkannt und Fördermaßnahmen zur Änderung eventuell problematischer Verhaltensweisen geplant werden (Hartmann, 2017). Verhaltensverlaufsdiagnostik soll also ein Evaluationsverfahren bereitstellen, das eine Überprüfung des Fördererfolgs angewendeter Maßnahmen in einem individuellen und situationsspezifischen Rahmen im Einzelfall zulässt. Das gewählte Prozedere muss alltagstauglich sein, da es, wie oben erwähnt und auch in den Ausführungen zur Lernverlaufsdiagnostik bereits thematisiert, regelmäßig und mit kurzem zeitlichem Abstand zwischen den Erfassungen durchgeführt werden muss. Insbesondere für wissenschaftlich noch nicht ausreichend evaluierte Maßnahmen, Methoden und Verfahren zur Förderung oder Änderung von Verhaltensweisen ist es von enormer Relevanz, zeitnah deren Vereinbarkeit mit den Lernbedürfnissen des jeweiligen Kindes oder des jeweiligen Jugendlichen zu überprüfen (Hartmann, 2017; Casale et al., 2015a).

Huber und Casale (2015) empfehlen auf Grundlage dieser Forderungen für die Durchführung verhaltensverlaufsdiagnostischer Erfassungen deshalb die Nutzung veränderungssensitiver Ratingskalen mit einer sechs- bis elfstufigen Einteilung. Die Orientierung sollte dabei anhand

von Zahlen oder qualitativen Bewertungen stattfinden. Außerdem sollte das Instrument regelmäßig eingesetzt und stets von der gleichen Lehrkraft durchgeführt werden. Die Regelmäßigkeit sollte durch eine im Vorfeld festgelegte Anzahl von Beobachtungen pro Woche sichergestellt werden. Bei der Erfassung mehrerer Werte an einem Tag sollte zur Auswertung das arithmetische Mittel aus den vorhandenen Werten gebildet werden. Empfehlenswert ist es, zur Visualisierung der Veränderung in der Entwicklung des jeweiligen Schülers oder der jeweiligen Schülerin eine Verhaltensverlaufskurve zu erstellen.

Die Forschung im deutschsprachigen Raum richtet sich im Rahmen evidenzbasierter (Sonder-)Pädagogik bei der Verlaufsdiagnostik des Verhaltens von Schüler_innen derzeit noch vorwiegend auf sonderpädagogische Bereiche wie den Förderschwerpunkt Emotionale und Soziale Entwicklung (Casale et al., 2015a). Die Verhaltensverlaufsdiagnostik zielt auch hier darauf ab, mit Hilfe der konsequenten Erfassung der Lern- und Entwicklungsausgangslage und auf Grundlage der Beobachtungsergebnisse Förderangebote abzuleiten. Darüber hinaus sollen Möglichkeiten ermittelt werden, ungünstigen Verhaltensweisen vorzubeugen, die nach der Identifikation durch die Diagnostik analysiert werden können. Übergreifende Ziele sind dementsprechend die Evaluation pädagogischer Handlungsmöglichkeiten und die Änderung von Verhalten durch individuell angepasste Förderangebote (Casale et al., 2015a).

Aufgrund der wiederholten, regelmäßigen Erfassung von Daten über einen verhältnismäßig kurzen Zeitraum hinweg und des hieraus resultierenden Potentials, Aussagen über Förderungen zu treffen, erfüllt die Verhaltensverlaufsdiagnostik einen wichtigen Teil der Forderungen evidenzbasierter Praxis (Casale et al., 2015a). Ein weiterer Vorteil besteht in der Ökonomie verhaltensverlaufsdiagnostischer Verfahren. Da die Verhaltensbeobachtungen und -beurteilungen von Personen aus dem natürlichen schulischen Umfeld der Schüler_innen durchgeführt werden und, sofern nach verhaltensverlaufsdiagnostischem Prinzip konzipiert, kaum Zeit beanspruchen, können sie leicht in den Schulalltag integriert werden. Außerdem ist die auf Basis der Auswertung verhaltensverlaufsdiagnostischer Instrumente mögliche Erstellung einer Verhaltensverlaufskurve positiv festzuhalten. Diese kann für diverse Bereiche, etwa für die erwähnte Erstellung von Förderplänen, für die Elternarbeit oder für den Austausch mit Kolleg_innen, im pädagogischen Alltag genutzt werden. Die Aussagen, welche die Verhaltensverlaufsdiagnostik über die Effektivität der einer Verhaltensentwicklung zugrundeliegenden

Förderungen treffen kann, wurden oben bereits thematisiert und zählen ebenfalls zu den Vorteilen der Verfahren (Hartmann, 2017; Casale et al., 2015a).

Im folgenden Abschnitt sollen einige von Verfahren zur Verhaltensverlaufsdiagnostik zu erfüllende, bereits erwähnte Testgütekriterien wie Ökonomie oder Veränderungssensibilität beziehungsweise -sensitivität zusammen mit weiteren genauer beschrieben werden. Hierbei soll eine Orientierung an dem Artikel von Casale, Hennemann, Huber und Grosche aus dem Jahr 2015 erfolgen, in dem die Autoren über diese Testgütekriterien und ihre Rolle in der Verhaltensverlaufsdiagnostik sprechen.

Einleitend ist zu erwähnen, dass die genannten Autoren die bisherigen, in der Regel für Statusdiagnostik eingesetzten Verfahren als für die Verlaufsdiagnostik von Schülerverhalten nicht geeignet einstufen. Auf die Problematik statusdiagnostischer Verfahren wurde oben bereits hingewiesen. Um Instrumente zur Verhaltensverlaufsdiagnostik zu entwickeln, müssen deshalb weitere Testgütekriterien beachtet werden, die nur zum Teil von statusdiagnostischen Verfahren erfüllt werden. Hierzu zählen obligatorische Bereiche wie Objektivität, Reliabilität und Validität, aber auch die Skalierung, die Anwendung eines gültigen Messmodells, Eindimensionalität, Inferenz, Direktheit und die Orientierung an der individuellen Bezugsnorm sowie die oben bereits genannte Ökonomie und die Veränderungssensibilität. Laut Casale et al. (2015b) lassen sich diese Gütekriterien auf Basis unterschiedlicher Anforderungen definieren, wobei hier zwischen zwei grundlegenden Fällen differenziert wird. Im ersten Fall sollen die Beobachtungsinstrumente zu zwei intersubjektiv vergleichbaren Ergebnissen führen. Das bedeutet, dass mehrere Beobachter bei der Beobachtung des gleichen Verhaltens am Ende zu gleichen numerischen Messwerten gelangen müssen. Dieses Verfahren wird unter dem Begriff der numerischen Invarianz geführt. Eine solche Form der Messung kann sowohl inter- als auch intraindividuelle Entwicklungen von Verhaltensweisen abbilden. Der zweite Fall beschreibt die numerische Invarianz als relative Übereinstimmung von Messwertprofilen über die Zeit. Hierbei müssen mehrere Beobachter_innen, auch Rater_innen genannt, Veränderungen von einer Beobachtungssituation zur anderen vergleichbar wahrnehmen. Beobachtet ein Rater oder eine Raterin eine Verbesserung einer Verhaltensweise um drei Skalenpunkte, so sollte der andere Rater oder die andere Raterin eine ähnliche Veränderungsintensität feststellen. Man

bezeichnet diese Form der Varianz als strukturelle Invarianz. Eine statusdiagnostische Aussage ist in diesem Fall nicht mehr möglich, wohl aber eine Aussage über die intraindividuelle Verhaltensentwicklung.

Casale et al. (2015b) analysieren die Testgütekriterien in ihrem Artikel auf Grundlage des ersten Falles, da die Verfahren der Verhaltensverlaufsdiagnostik in der Regel eine numerische Invarianz anstreben, um Aussagen über inter- und intraindividuelle Verhaltensveränderungen zulassen zu können. Im Folgenden werden ebendiese Testgütekriterien anhand des Artikels von Casale et al. (2015b) zusammengefasst dargestellt.

Objektivität

Unter Objektivität wird allgemein die Unabhängigkeit der Testergebnisse von der Person des Testleiters oder der Testleiterin und dem Zeitpunkt der Messung verstanden. Um die Objektivität zu wahren, sollten keine Variationen in der Durchführung, Auswertung oder Interpretation der Tests zwischen den jeweiligen Testleiter_innen und den Messzeitpunkten auftreten. Für Verlaufsmessungen stellt die Objektivität eine große Herausforderung dar, da insbesondere die Verhaltensverlaufsmessung in der Regel auf der subjektiven Wahrnehmung der durchführenden, das heißt beobachtenden und beurteilenden Person beruht. Eine Entwicklung im Verhalten des beurteilten Schülers oder der beurteilten Schülerin kann nur dann individuell nachgewiesen werden, wenn die leitende Person sich um Objektivität bemüht. Sonst könnten diagnostizierte Veränderungen auf Unterschiede in der Durchführung, der Auswertung oder der Interpretation zurückgeführt werden, nicht aber auf eine tatsächliche Änderung (Casale et al., 2015b).

Reliabilität

Die Reliabilität beschreibt den Grad der Messgenauigkeit bei der Erfassung eines bestimmten Merkmals. Unterschieden wird hier zwischen der internen Konsistenz, die aus der Korrelation der einzelnen Items untereinander berechnet wird, der Retest-Reliabilität, die angibt, wie stabil ein Merkmal ist und sich aus der Korrelation der Messwerte eines Tests zu zwei unterschiedlichen, zeitlich auseinanderliegenden Zeitpunkten berechnet, und der Paralleltestreliabilität, welche die Messergebnisse zweier gleich konstruierter Tests mit anderen, aber parallelen, das

heißt das gleiche Merkmal in anderer Form abfragenden Items vergleicht. Da die Verhaltensverlaufsmessung darauf ausgerichtet ist, mit Hilfe von häufigen Messungen eine genaue Veränderungsmessung durchzuführen, darf bei der Erfassung die Retest-Reliabilität nicht zu hoch ausfallen, da sie die Änderungssensibilität nicht beeinflussen soll. Das bedeutet, dass ausschließlich tatsächliche Verhaltensänderungen gemessen werden und keine inhärenten Schwankungen innerhalb eines Merkmals auftreten dürfen. Generell ist die Reliabilität von Instrumenten zur Verhaltensverlaufsdiagnostik wichtig um sicherzustellen, dass Veränderungen nicht zufällig zustande kommen. Zudem ist sie relevant, um Entwicklungen über die Zeit klar erkennbar abzubilden und so eine Grundlage für die Konzeption von Fördermaßnahmen zu gewinnen (Casale et al., 2015b).

Validität

Ob ein Test wirklich das misst, was er zu messen beansprucht, gibt die Validität an. Man unterscheidet hierbei zwischen Inhalts- und Konstruktvalidität. Erstere wird durch theoretische Ableitungen oder Expertenbefragungen erfasst. Die Konstruktvalidität, welche den Präzisionsgrad der Messung angibt, kann unterteilt werden in die konvergente Validität, bei der zwei das gleiche Konstrukt messende Tests stark miteinander korrelieren, und in die diskriminante Validität, die vorliegt, wenn zwei unterschiedliche Konstrukte messende Tests sehr schwach oder gar nicht miteinander korrelieren. Die Validität eines Verfahrens zur Verhaltensverlaufsdiagnostik im schulischen Kontext ist von hoher Relevanz und steht immer im direkten Zusammenhang mit den jeweiligen ausgewählten Konstrukten. Da viele Instrumente eher klinisch bedeutsame Verhaltensweisen messen, die im Schulalltag von geringerer Bedeutung oder nicht beobachtbar sind, sollte die soziale Validität berücksichtigt werden. Hierbei kommt es auf die Auswahl von Verhaltensweisen an, die unmittelbare Relevanz für das Verhalten von Schüler_innen in der Schule haben. Dies ist besonders wichtig, wenn Förderentscheidungen auf Basis der Verhaltensverlaufsdiagnostik getroffen werden sollen. Da diese kaum auf Basis unbedeutender Verhaltensweisen begründet werden können, sollten Bereiche gewählt werden, die für die tägliche pädagogische Arbeit der Lehrkräfte einen hohen Stellenwert aufweisen (Casale et al., 2015b; Pelham, Fabiano, & Massetti, 2005).

Skalierung

Unter der Skalierung wird die Verrechnungsvorschrift verstanden, mit der ein Testwert gebildet wird. Sie legt die Repräsentation von Merkmalsausprägungen durch einen Testwert fest und dient der Interpretation einer gemessenen Mermalsausprägung. Außerdem gibt die Skalierung an, ob die Vorgehensweise zur Bildung des Testwertes empirisch belegbare Unterschiede zwischen verschiedenen Testpersonen widerspiegelt. In Bezug auf die Erfassung von Verläufen im Schülerverhalten ist es wichtig, die Skalierung so zu wählen, dass sie die subjektive Unter- oder Überschätzung in der Testauswertung begrenzen kann. Bei unzureichend skalierbaren Tests, in denen beispielsweise von der Relevanz oder der Intensität des dahinterstehenden Verhaltens zu unterschiedliche Items miteinander verglichen werden, kann eine Interpretation der Testwerte nicht sinnvoll durchgeführt werden. Ein Summenscore, das ist die Summe der Ergebnisse einzelner Items, wäre dann inhaltlich wertlos (Casale et al., 2015b).

Ökonomie

Das Testgütekriterium der Ökonomie wurde bereits angesprochen und stellt eines der wichtigsten Kriterien für die Durchführung eines Verfahrens in der Praxis dar. Gute Tests und Diagnosematerialien sind in der Handhabung und Auswertung leicht anwendbar. Bei ihnen stehen Testzeit und Kosten zudem in einem angemessenen Verhältnis zum zu erwartenden Testertrag. Es kommt hierbei auf eine kurze und schnelle Durchführbarkeit, wenig benötigtes Personal und eine möglichst enge Kopplung der Messergebnisse an geeignete pädagogische Handlungsmaßnahmen an. Außerdem sollte prinzipiell möglichst wenig Material benötigt werden. Da Instrumente zur Verhaltensverlaufsdiagnostik verhältnismäßig häufig und regelmäßig angewendet werden, ist ein nicht ökonomisches Verfahren kaum durchführbar (Casale et al., 2015b).

Anwendung eines gültigen Messmodells

In Messmodellen werden latente, nicht beobachtbare Variablen mit manifesten, beobachtbaren Variablen in Beziehung gesetzt. Generell wird zwischen reflexiven und formativen Messmodellen unterschieden. Reflexiv bedeutet in diesem Fall, dass die Ausprägung der latenten Variablen die Merkmalsausprägung der manifesten Indikatoren beeinflusst. Unter formativ versteht man das Gegenteil, also die Beeinflussung der Ausprägung der latenten Variablen durch die manifesten Indikatoren. In der klassischen und in der probabilistischen Testtheorie wird

gewöhnlich entweder aus den manifesten Itemwerten und einem Messfehler die latente Variablenausprägung postuliert oder zur Berechnung der Lösungswahrscheinlichkeit eines Items beziehungsweise der Wahrscheinlichkeit der Wahl eines Items die Personenfähigkeit in Beziehung zur Itemschwierigkeit gesetzt. Der Begriff der Itemschwierigkeit muss für die Verhaltensverlaufsdiagnostik neu definiert werden, da sich die Frage stellt, ob die Verhaltensweisen die Items im eigentlichen Sinne darstellen. Hier besteht ein erheblicher Unterschied zur Lernverlaufsdiagnostik, bei welcher in der Regel Daten aus objektiven Indikatoren konzipiert werden. Im Bereich des Verhaltens gestaltet sich dies schwieriger, da es im Normalfall kein prinzipiell richtiges oder falsches Verhalten gibt, sondern sich eher die Frage nach der Angemessenheit oder Unangemessenheit des Verhaltens stellt. Wie bereits erwähnt, ist diese Einschätzung immer subjektiv geprägt, woraus sich ergibt, dass es sich hierbei um subjektive Indikatoren handelt. Diese sind allerdings messtheoretisch als stark fehlerbehaftet einzustufen (Bühner, 2011). Die Anwendung der Generalisierbarkeitstheorie macht es möglich, dass verschiedene Fehlerquellen betrachtet werden können, welche die Verhaltensbeurteilung beeinflussen und deren simultane Interaktion in der Analyse der Testgüte ausgewertet wird. Zu solchen Fehlerquellen können die Beurteilung durch verschiedene Rater_innen, das Beurteilen verschiedener Kinder und die Beurteilung in verschiedenen Situationen zählen. In der Klassischen Testtheorie ist eine solche Differenzierung nicht möglich. Darüber hinaus muss die Verlaufsdiagnostik von Schüler_innenverhalten auf einem gültigen Messmodell basieren, damit die Veränderungen in den Messwerten korrekt interpretiert werden können. Fehlt ein gültiges zugrundeliegendes Messmodell, könnten andere Fehlerquellen die Veränderungen der Messwerte beeinflusst haben (Casale et al. 2015b).

Eindimensionalität

Eine Eindimensionalität liegt dann vor, wenn das Antwortverhalten einer Person auf nur eine bestimmte Kompetenz zurückgeführt werden kann (Bühner, 2011). Die Eindimensionalität des Konstrukts ist Grundlage und notwendige Bedingung eines Tests im Bereich der Verlaufsdiagnostik. Ohne Eindimensionalität kann das Testergebnis nicht zweifelsfrei auf das entsprechende Merkmal zurückgeführt werden. Dies ist insbesondere für die Verhaltensverlaufsdiagnostik wichtig, da ihr immer die Messung der Veränderung eines Merkmals zugrundeliegt. Die lokale stochastische Unabhängigkeit der Items, die durch Eindimensionalität geschaffen werden kann und deren Beantwortung unabhängig erfolgen muss, ist deshalb essentiell in der

Verlaufsdiagnostik von Schülerverhalten. In der klassischen Testtheorie kann die Eindimensionalität mittels Faktorenanalyse überprüft werden (Bühner, 2011). Eine solche Überprüfung ist unerlässlich, damit in den Messwerten abgebildete Entwicklungsverläufe nicht durch das Zugrundeliegen unterschiedlicher Konstrukte entstehen. Wenn zum Beispiel verschiedene voneinander unabhängige Verhaltensdimensionen in einem Instrument erfasst werden, dürfen diese nicht zu einem gemeinsamen Rohwert addiert werden. Es ist beispielsweise ungünstig, bei der Förderung von Sozialkontakt auch aggressives Verhalten zu erfassen, da die Auswirkungen der Förderungen dann nicht zweifelsfrei nachgewiesen werden können (Casale et al. 2015b).

Änderungssensitivität

Die Items eines Tests müssen auch bei Anwendung innerhalb eines kurzen Zeitraums bereits kleine Veränderungen in den gemessenen latenten Merkmalen der getesteten Person abbilden können. Ein solches änderungssensitives Testdesign zielt, wie eingangs erwähnt, nicht auf das Abbilden zeit- oder situationsübergreifender Persönlichkeitseigenschaften ab, sondern auf das Erfassen von Änderungen der zu erfassenden Merkmale. Verfahren aus der Statusdiagnostik sollen keine Veränderungen abbilden und sind dazu konzipiert, veränderungsresistente Merkmale zu erfassen. Wie bereits im Abschnitt „Reliabilität“ angesprochen, sind Tests mit einer sehr hohen Retest-Reliabilität nicht änderungssensibel und daher für die Verhaltensdiagnostik eher ungeeignet, da sie in der Regel Items beinhalten, deren Ausprägung sich über die Zeit nicht verändert. Instrumente zur Verhaltensverlaufsdiagnostik sollten folglich nicht zur Messung von Dispositionen konstruiert sein, sondern konkrete Verhaltensweisen beinhalten, die veränderungssensitiv nachgewiesen werden können. Eine grobe Einteilung wie zum Beispiel „aggressives Verhalten“ wäre wenig gewinnbringend, stattdessen sollte der Verhaltensbereich kleinschrittig zergliedert werden. Nur eine solchermaßen detaillierte Erfassung ermöglicht es den Lehrkräften, adäquate Aussagen über individuell adaptierte Fördermaßnahmen zu treffen (Casale et al., 2015b).

Inferenz

Die Inferenz beschreibt den Aufwand beziehungsweise die schlussfolgernde Kognition, die zum Ausfüllen des Tests oder zur Lösung eines Items benötigt wird. Bei akademischen Tests ist der Grad an Inferenz in der Regel niedrig, da ein Item gelöst wird, sobald die hierfür erfor-

derliche Kompetenz vorhanden ist. Die Inferenz von Beobachtungsbögen im Bereich der Verhaltensbeobachtung und -beurteilung ist immer vom Grad der Operationalisierung, der Häufigkeit des Auftretens und der Zeitdauer eines Merkmals abhängig. Ist ein Merkmal eher global formuliert, fällt die Inferenz höher aus, als wenn es sehr eng definiert ist. Eine solche allgemeine Formulierung ist dementsprechend ungünstig für die Verhaltensverlaufsdiagnostik, da die Einschätzung weitgefasster, global formulierter Items sehr aufwendig sein kann und daher dem Prinzip der Ökonomie entgegensteht. Eine niedrig inferente Herangehensweise bietet sich deshalb an. Auch in Bezug auf die Objektivität können sich inhaltlich eng umgrenzte Items positiv auswirken. Empirisch muss die Inferenz in Bezug auf die Verhaltensverlaufsdiagnostik noch überprüft werden. Es ist noch nicht abschließend untersucht worden, welcher Grad an Inferenz sich am besten für die Verhaltensverlaufsdiagnostik eignet (Casale et al., 2015b).

Direktheit

Verhalten ist am genauesten zu erfassen, wenn es in der Situation selbst oder unmittelbar im Anschluss an diese beurteilt wird. Die Direktheit beschreibt dementsprechend die zeitliche Nähe der Messung zum Verhalten, das erfasst werden soll. Die Latenz zwischen diesen beiden Punkten beeinflusst die Qualität der Messung und der entstehenden Daten maßgeblich. Deshalb sollten insbesondere in der Verhaltensverlaufsdiagnostik, bei der es auf kleinste Änderungen in den zu messenden Verhaltensweisen ankommt, die Erfassungen der jeweiligen Merkmale direkt im Anschluss an die Situation erfolgen. Um dies zu ermöglichen, sollten die Beurteilungszeiträume möglichst kurz gehalten werden (Casale et al., 2015b).

Orientierung an individueller Bezugsnorm

Im Bereich der Statusdiagnostik werden in der Regel eher verschiedene Individuen miteinander verglichen und zueinander in Beziehung gesetzt. Dementsprechend findet die Beurteilung im Rahmen der sozialen Bezugsnorm statt. Bei der Lern- und der Verhaltensverlaufsdiagnostik werden die Schüler_innen mit sich selbst verglichen, also im Sinne der individuellen Bezugsnorm bewertet. Hierbei interessiert nicht die absolute Ausprägung des Verhaltens, sondern die individuelle Veränderung des Verhaltens innerhalb des Testzeitraums. Deshalb ist eine Normierung der Verfahren in der Verlaufsdiagnostik nicht von gleicher Wichtigkeit wie in der Statusdiagnostik (Casale et al., 2015b).

Wie oben erwähnt, werden laut Casale et al. (2015b) die bereits existierenden statusdiagnostischen Konzepte aus dem Bereich der emotionalen und sozialen Entwicklung den eben aufgeführten Testgütekriterien für die Verhaltensverlaufsdiagnostik nicht gerecht. Instrumente zur systematischen Verhaltensbeobachtung und Ratingskalen zur Verhaltensbeurteilung schnitten in der Analyse am besten ab. Beide erfüllen die Kriterien von Reliabilität, Validität, Skalierung, Eindimensionalität, Inferenz und individueller Bezugsnorm. Außerdem wenden beide Methoden ein gültiges Messmodell an. Die systematische Verhaltensbeobachtung kann zusätzlich den Kriterien der Objektivität und der Veränderungssensitivität sowie der Direktheit gerecht werden, wohingegen die Verhaltensbeurteilung dem Anspruch der Ökonomie nachkommt (Casale et al., 2015b).

Abschließend kann festgestellt werden, dass es derzeit an Methoden und Verfahren zur verlaufsdiagnostischen Erfassung von Schülerverhalten mangelt. Die Neuentwicklung von Verfahren, die sich für Verhaltensverlaufsdiagnostik eignen, stellt daher ein bedeutsames Desiderat dar. Einen Schritt in diese Richtung leistet das aus dem US-amerikanischen Raum kommende und aus dem Strengths and Difficulties Questionnaire hervorgegangene Direct Behavior Rating (Direkte Verhaltensbeurteilung) (siehe Kapitel 2.4).

2.3 Strengths and Difficulties Questionnaire

Der Strengths and Difficulties Questionnaire (SDQ) ist ein von Goodman 1997 entwickeltes Instrument zur statusdiagnostischen Erfassung verschiedener Verhaltensweisen. Er kann sowohl die Stärken als auch die Verhaltensauffälligkeiten der Kinder und Jugendlichen erfassen. Der SDQ hat im vergangenen Jahrzehnt international an Bedeutung gewonnen. Diese Entwicklung kann darauf zurückgeführt werden, dass das Instrument mit verhältnismäßig wenig Items eine Aussage über das Verhalten von Kindern und Jugendlichen zwischen zwei und 17 Jahren zu treffen ermöglicht (Gebhardt & Voß, 2017; Lohbeck, Schultheiß, Petermann & Petermann, 2015). Entwickelt wurde der SDQ als Screeninginstrument und wird häufig in der klinischen Diagnostik eingesetzt. Er besteht aus den in Kapitel 2.1.1 bereits beschriebenen fünf Dimensionen (Emotionale Probleme, Verhaltensprobleme, Hyperaktivität, Verhaltensprobleme mit Gleichaltrigen und Prosoziales Verhalten), die jeweils einen bestimmten Verhaltensbereich abdecken. Diese Dimensionen sind inhaltlich abgeleitet worden aus dem Bereich der psychischen Störungen bei Kindern im Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000). Jede Dimension setzt sich aus fünf Items

zusammen, die retrospektiv von der den SDQ anwendenden Person für die vergangenen sechs Monate bewertet werden. Es kann zwischen „zutreffend“, „nicht zutreffend“ und „teilweise zutreffend“ gewählt werden (Gebhardt & Voß, 2017). Neben Versionen für Lehrkräfte und Eltern existieren auch Selbstbeurteilungsversionen für Kinder und Jugendliche. Die drei Versionen (Version für die Lehrkräfte, für die Erziehenden und für die Kinder/Jugendlichen) sind inhaltlich vergleichbar. Die Selbstbeurteilungsbögen wurden im Vergleich zu den zur Fremdbeurteilung konzipierten geringfügig verändert (Gebhardt & Voß, 2017; Lohbeck et al., 2015). Der SDQ ist im Internet in circa 50 verschiedenen Sprachen kostenfrei verfügbar (Strengths and Difficulties Questionnaire, 2015).

Das Funktionieren des SDQs und seiner fünf Dimensionen ist bereits in verschiedenen Ländernormierungen durch explorative und konfirmatorische Faktorenanalysen bestätigt worden (z. B. Lohbeck et al. 2015). Die Aufteilung in eine sogenannte Fünf-Faktoren-Struktur konnte jedoch nicht von allen Studien als sinnvoll nachgewiesen werden. Goodman, Lamping und Ploubidis (2010) und Dickey und Blumberg (2004) sprechen sich beispielsweise für eine dreigliedrige Strukturierung in die Bereiche Internalisierende Verhaltensprobleme, Externalisierende Verhaltensprobleme und Prosoziales Verhalten aus. In den Bereich der Internalisierenden Verhaltensprobleme würden die SDQ-Skalen Emotionale Probleme und Verhaltensprobleme mit Gleichaltrigen, in den Bereich der Externalisierenden Verhaltensprobleme die SDQ-Skalen Verhaltensprobleme und Hyperaktivität eingeordnet. Die dritte Dimension, das Prosoziale Verhalten, bliebe unverändert. Diese Drei-Faktoren-Struktur hat sich in den genannten Studien als angemessener erwiesen als die Fünf-Faktoren-Struktur (Lohbeck et al., 2015). Es liegen allerdings noch keine abschließenden Studien vor, die eine der beiden Strukturen final befürworten oder ablehnen würden (DeVries, Gebhardt & Voß, 2017).

Der wohl wichtigste Wert bei der Verwendung und Auswertung des SDQs ist der Gesamtproblemwert, welcher sich als Summe aus den Dimensionen Verhaltensprobleme, Verhaltensprobleme mit Gleichaltrigen, Emotionale Probleme und Hyperaktivität ergibt. Für die Ergebnisse und errechneten Werte des SDQs gibt es internationale Normen, die eine Klassifikation des Gesamtwerts und der einzelnen Subskalen in „normal“, „grenzwertig“ oder „auffällig“ ermöglichen (Gebhardt & Voß, 2017; Lohbeck et al., 2015). Diese Werte wurden auf Basis der jeweiligen Verteilung der Rohwerte aus britischen Normierungsstudien bestimmt. Dabei

wurden 80% der Werte als unauffällig eingestuft. Jeweils 10% der Werte wurden als grenzwertig und auffällig eingeordnet (Lohbeck et al., 2015). Die Skala Prosoziales Verhalten wird nicht mit in den Gesamtproblemwert eingerechnet.

2.4 Direct Behavior Rating

In Kapitel 2.2.2 wurde bereits auf das Verfahren des Direct Behavior Ratings (DBR) hingewiesen, welches um das Jahr 2000 im englischsprachigen Raum entwickelt wurde. Im deutschsprachigen Raum ist das DBR erst seit einigen Jahren im Gespräch und wird in der Regel unter dem Begriff der direkten Verhaltensbeobachtung oder allgemein unter Verhaltensverlaufsdiagnostik diskutiert. Aufgrund der fehlenden Eignung aktueller Verfahren aus dem Bereich der Statusdiagnostik zur Messung von Verhaltensverläufen (siehe Kapitel 2.2.2 oder Casale et al. 2015b) bedarf es der (Weiter-)Entwicklung ebendieser neueren diagnostischen Methoden. Instrumente zur Erfassung von Entwicklungen im Bereich von Schülerverhalten müssen regelmäßig direkt nach der Beobachtung des Verhaltens eingesetzt werden und Veränderungen sensitiv abbilden. Im Anschluss an die Beurteilung des Verhaltens können die Ergebnisse dazu genutzt werden, günstige oder ungünstige Entwicklungen oder Stagnation zu erkennen und mit entsprechenden Förderungen darauf zu reagieren (Hartmann, 2017).

Das DBR ist aus einer Kombination aus der direkten systematischen Verhaltensbeobachtung und der Verhaltensbeurteilung mittels Ratingskalen entstanden. Diese beiden Verfahren wurden in Kapitel 2.1.2 bereits kurz angesprochen und sollen hier noch einmal detaillierter dargestellt werden. Die direkte systematische Verhaltensbeobachtung gehört zu den exaktesten und differenziertesten Formen der Beobachtung von Verhaltensweisen. Hierbei wird ein bestimmtes Verhalten fokussiert und mit Hilfe von Zeit- und Ereignisstichproben, die zum Beispiel die Häufigkeit des Auftretens zählen, quantifizierbar gemacht (Hussy, Schreier & Echterhoff, 2013). Um dies leisten zu können, benötigt die direkte systematische Beobachtung eine genaue und konkrete Operationalisierung des zu beobachtenden Verhaltens. Die für die Durchführung der Beobachtung vorgesehenen Personen müssen dementsprechend geschult sein, um den Anforderungen dieser Art von Beobachtungen gerecht werden zu können. Bei der direkten systematischen Verhaltensbeobachtung kann zwischen der Erfassung von Zeichensystemen und der Erfassung von Kategoriensystemen unterschieden werden (Schmidt-Atzert & Amelang, 2012). Werden Zeichensysteme gewählt, erfasst die beobachtende Person ausschließlich gezielt einzelne, für eine vorher definierte Fragestellung relevante Verhaltensweisen. Während

im Rahmen von Zeichensystemen eher ein kleiner, fest abgesteckter Abschnitt des Verhaltens beobachtet und festgehalten wird, werden bei der Erfassung von Kategoriensystemen breite Bereiche berücksichtigt. Diese können potentiell an den Grenzen überschneidenden Kategorien zugeordnet werden. In beiden Fällen muss die beobachtende Person direkt oder indirekt anwesend sein, das heißt der zu observierenden Situation entweder aktiv oder passiv beiwohnen oder diese mit Hilfe eines Videomitschnitts bewerten (Schmidt-Atzert & Amelang, 2012).

Neben der direkten systematischen Verhaltensbeobachtung sind auch Bereiche der Verhaltensbeurteilung im DBR enthalten. Hierbei handelt es sich im Gegensatz zur direkten systematischen Verhaltensbeobachtung um eine eher abstrahierende Methode, da der Beobachter in Bezug auf eine zuvor definierte Verhaltensweise oder einen zuvor definierten Verhaltensbereich in einem bestimmten Zeitraum nach seiner Einschätzung gefragt wird (Schmidt-Atzert & Amelang, 2012). Mit Hilfe in der Regel mehrstufig aufgebauter Ratingskalen wird die Einschätzung der Rater_innen erfasst. Dabei werden keine konkreten Verhaltensweisen gezählt, sondern es wird die Einschätzung und Wahrnehmung einer in der Beobachtungssituation anwesenden Person festgehalten. Die zeitliche Entfernung zwischen der Beobachtung und der Dokumentation der Verhaltenseinschätzung kann variieren und sich möglicherweise über einige Tage oder sogar Wochen erstrecken. Man kann mit der Verhaltensbeurteilung dementsprechend auch die Beobachtungen mehrerer Unterrichtsstunden, Schultage oder Wochen zusammengefasst auswerten und festhalten (Schmidt-Atzert & Amelang, 2012; Huber & Rietz, 2015). Die Verhaltensbeurteilung steht unter anderem aufgrund der möglichen zeitlichen Verzögerung zwischen beobachteter Situation und deren Erfassung immer wieder in der Kritik, da diese Verzögerung die Verhaltensbeurteilung fehleranfällig machen kann. Die direkte systematische Verhaltensbeobachtung wiederum bietet zwar eine exaktere Erfassung von Verhaltensweisen, ist jedoch nicht so ökonomisch wie die Verhaltensbeurteilung (u.a. Huber & Rietz, 2015).

Wie bereits erwähnt, vereint das DBR diese beiden oben genannten Methoden der systematischen Verhaltensbeobachtung und der Verhaltensbeurteilung. Erstmals dargestellt wurde es von Steege, Davin und Hathaway (2001) und Chafouleas, Riley-Tillman und McDougal (2002). Die praktische Umsetzung von DBRs in der Schule sieht zunächst die Festlegung eines Verhaltensmerkmals vor, das die untersuchenden Personen interessiert und durch Förderung

in einer bestimmten Situation verbessert werden soll. Direkt im Anschluss an die Situationsbeobachtung wird das Verhalten in einer Ratingskala bewertet. Diese Beurteilung kann und sollte mehrfach wiederholt werden, sodass bereits nach einigen Durchgängen ein Verlauf abgebildet werden kann. Durch die Erstellung anschaulicher Diagramme, die Entwicklungsverläufe visualisieren, können zum Beispiel die Lehrkräfte Rückschlüsse auf die Effektivität von Fördermaßnahmen ziehen (Casale et al., 2015a). Das DBR hat also die Evidenzbasierung konkreter, individueller Fördermaßnahmen in Form einer engmaschigen und ressourcenorientierten Begleitung der Verhaltensentwicklung von Schüler_innen zum Ziel (Casale et al., 2015a).

Während die beiden oben genannten Verfahren, die direkte systematische Verhaltensbeobachtung und die Verhaltensbeurteilung, hauptsächlich der Statusdiagnostik dienen und für sich alleine stehend der für die Verhaltensverlaufsdiagnostik geforderten Ökonomie nicht genügen können, sind sie in Kombination als DBR durchaus geeignet (Huber & Rietz, 2015). Aufgrund ihrer Flexibilität und Praktikabilität, die DBRs dank der Einsetzbarkeit in verschiedenen Situationen mit verhältnismäßig wenig Aufwand zur Erfassung unterschiedlicher Verhaltensweisen mit sich bringen, können sie diese Anforderung an die Ökonomie für den schulischen Kontext besser erfüllen (Christ, Riley-Tillman & Chafouleas, 2009). Durch das direkte Ausfüllen der Ratingskalen im Anschluss an die beobachtete Situation oder noch in der zu beobachtenden Situation wird das Problem der zeitlichen Verzögerung, welches die klassische Verhaltensbeurteilung beinhaltet, minimiert. Außerdem kann die unökonomische Seite der direkten systematischen Verhaltensbeobachtung durch die Nutzung von Ratingskalen reduziert werden.

Bestehend aus drei Bereichen (Direktheit, Verhaltensbezogenheit und Beurteilung), hat das DBR drei wesentliche Eigenschaften. Die Direktheit bezieht sich in diesem Fall auf die zeitnahe Beobachtung und Beurteilung des Verhaltens in Bezug zum tatsächlichen Auftreten der Verhaltensweisen. Die im Vorhinein definierte Zeitspanne kann sich dabei von wenigen Sekunden bis hin zu einem gesamten Schultag strecken. Das DBR wird immer verhaltensbezogen durchgeführt. Vorher definierte, konkrete, beobachtbare Verhaltensweisen sollen erfasst werden. Es wird von abstrakten hypothetischen Konstrukten Abstand genommen und es werden durch genaue Operationalisierung für alle Personen, die an der Förderung des Schülers oder der Schülerin beteiligt sind, beobachtbare und nachvollziehbare Items formuliert. Außer-

dem wird durch die Beurteilung die Quantifizierung der Wahrnehmung des Beobachters verschiedener Verhaltensweisen in Ratingskalen vorgenommen. Hierbei geht es weniger um eine genaue Erfassung der Häufigkeit des Auftretens bestimmter Verhaltensweisen als um die Einschätzung der Verhaltensweisen in einer Skala (Casale et al., 2015b).

Unterschieden wird bei DBRs zwischen Multi-Item-Skalen (im Folgenden MIS) und Single-Item-Skalen (im Folgenden SIS). Unter MIS werden Skalen verstanden, die mehrere Items erfassen, wobei die Anzahl der Items je nach Bedarf und Notwendigkeit variiert und individuell an die Bedürfnisse der jeweiligen diagnostischen Erhebung und den entsprechenden Schüler oder die entsprechende Schülerin angepasst werden kann. SIS beinhalten nur ein einzelnes Item, welches im Rahmen des DBRs erfasst werden soll. Eine solche Skala ist sehr ökonomisch und ermöglicht ebenfalls eine kontinuierliche Erfassung einer Verhaltensweise im Unterricht (Hartmann, 2017).

Die Formulierung der Items kann ungeachtet dessen, ob es sich um eine SIS oder eine MIS handelt, global oder spezifisch sein. Unter global formulierten Items versteht man die eher umfassende, grobe Formulierung eines Verhaltensbereichs. Das Item „Zeigt störendes Verhalten“ wäre ein eher global formuliertes Item, welches verschiedene Bereiche umfassen würde. „Ruft herein anstatt sich zu melden“ wäre hingegen ein spezifisch formuliertes Item (Casale et al., 2015b). Neben der Unterscheidung zwischen global und spezifisch formulierten Items wird außerdem zwischen nomothetischen und idiografischen Ansätzen zur Formulierung der Items differenziert. Werden die Items auf Grundlage idiografischer Ansätze ausgewählt, bekommt jede Schülerin und jeder Schüler eigene, individuell angepasste Items. Der idiografische Ansatz geht vom Kind aus und ist auf den Einzelfall ausgerichtet. Die Passung der Items zum Kind ist dementsprechend exzellent, während die Testgüte aufgrund der geringen Stichprobengröße (N=1) sehr gering ausfällt (Casale et al., 2015b). Dem steht der nomothetische Ansatz gegenüber, welcher nicht das Individuum in den Vordergrund stellt, sondern sich an Studien mit großen Stichproben orientiert. Die Items werden dann anhand ihrer psychometrischen Eignung in den Studien ausgewählt. Die Testgüte dieser Items ist sehr gut, die Eignung im Einzelfall kann dagegen unpassend sein (Casale et al., 2015b).

Demzufolge ist die Gestaltung von Instrumenten, die für DBRs genutzt werden sollen, sehr flexibel und immer auch abhängig vom Zweck der Erhebung von Verhaltensverläufen.

Sowohl Huber und Casale (2015) als auch Casale et al. (2015a) geben Empfehlungen für den Einsatz von DBRs. Diese sind insbesondere auf die Nutzung der DBR-Skalen in der Praxis bezogen. Empfohlen wird, an Zahlen oder qualitativer Bewertung orientierte Skalen mit sechs bis elf Skalenpunkten einzusetzen. Wichtig ist außerdem, dass das DBR regelmäßig und immer von derselben Lehrkraft durchgeführt wird. Die Anzahl der Beobachtungen sollte im Vorfeld festgelegt werden und bei der Erfassung von mehreren Werten pro Tag sollte das arithmetische Mittel dieser Werte gebildet werden. Die Erstellung einer Verhaltensverlaufskurve ist, wie bereits in Kapitel 2.2.2 erwähnt, sinnvoll, um die Entwicklungen im Verhalten zu visualisieren.

Im Anschluss soll nun anhand aktueller Studien zusammengefasst werden, inwiefern das aus der direkten systematischen Verhaltensbeobachtung und der Verhaltensbeurteilung konzipierte DBR die Gütekriterien der Verhaltensverlaufsdiagnostik erfüllt, welche in Kapitel 2.2.2 bereits erläutert und in Kapitel 2.1.2 in Bezug auf die eben genannten Methoden, aus denen sich das DBR zusammensetzt, dargestellt wurden.

Da viele der auszuwertenden Studien auf der Generalisierbarkeitstheorie basieren, soll diese hier kurz skizziert werden. Der Ansatz dieser Theorie geht auf die Idee zurück, dass Varianzen in den einzelnen Messwerten sich auf verschiedene Einflussfaktoren zurückführen lassen, die Auswirkungen auf die Messung haben. Diese Einflussfaktoren werden als Facetten bezeichnet. Häufig geht es bei den Studien, denen die Generalisierbarkeitstheorie zugrundegelegt wird, um Untersuchungen zur Interrater-Reliabilität von Instrumenten zum DBR, also den Anteil der Varianz, der auf die verschiedenen Rater_innen zurückzuführen ist (Huber & Rietz, 2015). Eingeführt in die Sozialwissenschaften wurde die Generalisierbarkeitstheorie von Cronbach, Gleser, Nanda und Rajaratnam im Jahr 1972. Der Ausgangspunkt für die Theorie war die immer wiederkehrende Frage nach Reliabilität und Validität von Verhaltensmessungen sowie nach dem Einfluss verschiedenster, auf die Beobachter_innen, das Testinstrument, die Beobachtungssituation und andere Bereiche zurückzuführende Fehlerquellen auf die Ergebnisse der Erfassungen (Casale, Hennemann, Volpe, Briesch & Grosche, 2015c). Diesem kann die Klassische Testtheorie nicht gerecht werden, was zur Entwicklung der Generalisierbarkeitstheorie führte (Cronbach et al., 1972). Die Generalisierbarkeitstheorie bietet also eine Erweiterung der Klassischen Testtheorie an, indem sie von einer Zerlegung des globalen Fehlerwertes in oben erwähnte einzelne Facetten ausgeht (Brennan, 2001). Hierbei werden die

Varianzkomponenten, aus denen sich der Messfehler zusammensetzt, sowie ihre Interaktion untereinander geschätzt. Dadurch ergibt sich ein entscheidender Vorteil im Vergleich zur Klassischen Testtheorie, die ausschließlich eine einzelne systematische Fehlerquelle einbezieht und die Überprüfung von Interaktionen nicht zulässt (Casale et al., 2015c). Die Generalisierbarkeitstheorie besteht aus zwei Schritten. Zuerst wird eine sogenannte Generalisierbarkeitsstudie durchgeführt, in der die Varianz der einzelnen Facetten (Rater_innen, Messzeitpunkt usw.) und die Interaktionen dieser Facetten geschätzt wird. Durchgeführt wird dies durch eine Varianzanalyse. Hierbei werden die Facetten als Faktoren betrachtet und die Bedingungen der Facetten als Faktorstufen behandelt (Brennan, 2001). Daran anschließend werden die Ergebnisse aus der Generalisierbarkeitsstudie in der Entscheidungsstudie weitergenutzt mit dem Ziel, das Messinstrument im Hinblick auf praktische Entscheidungen zu optimieren (Casale et al., 2015c). In der Entscheidungsstudie wird simuliert, inwiefern sich die Varianzaufklärung, das heißt der Anteil des jeweils betrachteten Einflusses an der Gesamtvarianz, verändert, wenn die Anzahl der zulässigen Bedingungen innerhalb bestimmter Facetten variiert (Casale et al., 2015c). Es werden Informationen über die Reliabilität der Messung für relative und absolute Vergleiche geliefert. Relative Vergleiche beziehen sich hierbei auf den normorientierten Vergleich zum Beispiel innerhalb einer Schulklasse, während ein absoluter Vergleich intraindividuell angelegt ist und sich auf ein spezifisches Kriterium einer bestimmten Person richtet. Dementsprechend eignet sich die Generalisierbarkeitstheorie sehr gut für die Entwicklung und Untersuchung von Instrumenten zur Verhaltensverlaufsdiagnostik, da sie zum einen für eben genau diesen Bereich, die Verhaltensmessung, konzipiert wurde und zum anderen die Möglichkeit bietet, Aussagen hinsichtlich der individuellen Bezugsnorm zu treffen (Casale et al., 2015c).

Casale et al. (2015b), Huber und Rietz (2015) sowie einige andere Autor_innen haben Studien des Forscherteams um Chafouleas, Christ und Riley-Tillman aus den Jahren 2007-2014 zusammengefasst und stellen diese in verschiedenen Artikeln dar. Diese relativ wenigen Studien beziehen sich hauptsächlich auf SIS. Im Bereich der MIS ist der Forschungsbedarf dementsprechend noch höher als bei den SIS (Casale et al., 2015b).

In der Veröffentlichung von Casale et al. (2015b) werden akzeptable Werte für die drei Hauptgütekriterien (Validität, Objektivität und Reliabilität) festgestellt, basierend auf Studien von

Christ et al. aus dem Jahr 2009 und von Chafouleas, Christ, Riley-Tillman, Briesch und Chanese von 2007. Huber und Rietz (2015) geben in Bezug auf die Anwendung klassischer Testgütekriterien bei der Arbeit mit DBRs geteilte Meinungen wieder. So kann auf der einen Seite die Güte von DBRs nicht durch die klassischen Testgütekriterien der Psychometrie abgebildet werden, da sie grundsätzlichen Einschränkungen wie zum Beispiel Wahrnehmungs- und Urteilsfehlern unterliegt. Auf der anderen Seite jedoch existieren im Vergleich zwischen Verhaltensbeobachtungen und psychometrischen Tests höchstens konzeptuelle, aber keine methodischen Unterschiede, was dafür spräche, dass Objektivität, Reliabilität und Validität relevante Hauptgütekriterien zur Bestimmung der Eignung von DBRs sind (Huber & Rietz, 2015). Cone (1988) spricht sich für eine Übertragung der Gütekriterien auf Verhaltensbeobachtungen aus. Eine solche Übertragbarkeit sei aufgrund der Annahmen machbar und sinnvoll, dass eine Verhaltensbeobachtung immer möglichst genau mit dem tatsächlich gezeigten Verhalten übereinstimmen sollte, dass ein und dieselbe Person das gleiche Verhalten zu zwei Zeitpunkten auch ähnlich bewertet und dass gleiches gezeigtes Verhalten in unterschiedlichen Settings stabil bewertet wird. Schmidt-Atzert und Amelang (2006) fordern, die Güte von Verhaltensbeobachtungen in jeder neuen Untersuchung erneut zu beurteilen. Dies begründen sie damit, dass sich sowohl der Beobachtungsgegenstand als auch die Beobachtungssituation und die beobachtende Person bei jeder neuen Durchführung ändern können.

Auf Basis dieser unterschiedlichen Meinungen in Bezug auf die Gültigkeit der Hauptgütekriterien werfen Huber und Rietz (2015) die Frage nach dem Nutzen der Erfüllung der Hauptgütekriterien in der Verhaltensverlaufsdiagnostik auf. Im Rahmen von Verhaltensverlaufsdiagnostik geht es hauptsächlich um die Abbildung der Entwicklung von Verhalten über die Zeit. Deshalb stellt sich die Frage, ob eine intersubjektive Übereinstimmung mit einem Kriterium oder einer Norm überhaupt gewährleistet werden muss. Insgesamt ist die messtheoretische Fundierung von DBRs noch nicht ausreichend diskutiert worden. Deshalb sollte die Verwendung klassischer Konzepte der Testtheorie vorerst nur unter Vorbehalt geschehen (Huber & Rietz, 2015).

Im Folgenden sollen nun die Ergebnisse der verschiedenen Studien, beginnend mit grundlegenden Studien zur Untersuchung der Beobachtungsgüte von DBRs und dann mit Konzentration auf spezifische Aspekte, zusammengefasst werden. Briesch, Chafouleas und Riley-Till-

man konnten 2010 in einer Studie feststellen, dass DBRs deutlich stärker von der beobachtenden Person abhängen als direkte systematische Verhaltensbeobachtungen. 20 % der Varianz waren bei DBRs durch Lehrkräfte mit der Interaktion zwischen der beurteilenden Lehrkraft und den beurteilten Zielschüler_innen zu erklären. Bei direkten systematischen Verhaltensbeobachtungen beträgt dieser Bereich der Varianz nur circa 1 %. Steege, Davin und Hathaway (2001), Riley-Tillman, Chafouleas, Sassu, Chanese und Glazer (2008) sowie Christ, Riley-Tillman, Chafouleas und Jaffery (2011) analysierten diese Beurteilungsfehler genauer und kamen zu dem Ergebnis, dass die beurteilten Verhaltensaspekte stark auf die Beobachtungsgüte einwirken. So konnten Hathaway et al. (2001) beispielsweise feststellen, dass bei der Analyse stereotypen Verhaltens und aktiver Teilnahme der Schüler_innen eine Übereinstimmung von 94 % beziehungsweise 95 % zwischen den Rater_innen vorlag, die das DBR nutzten, und denen, die anhand direkter systematischer Verhaltensbeobachtungen bewerteten. Dies spricht für eine gute Kriteriumsvalidität und eine vergleichsweise hohe Interrater-Reliabilität (Huber & Rietz, 2015).

Insgesamt kann festgehalten werden, dass noch keine grundsätzliche Ableitung in Bezug auf die Beobachtungsgüte von DBRs möglich ist, aber die aktuellen Studienergebnisse darauf hinweisen, dass die Unsicherheit hinsichtlich statusdiagnostischer Entscheidungen auf Basis von DBRs aufgrund des Einflusses der beobachtenden Person auf die Ergebnisse erhöht ist (Huber & Rietz, 2015). Riley-Tillman, Christ, Chafouleas, Boice und Briesch (2011) konnten bei der Überprüfung der Test-Retest-Reliabilität wiederum erkennen, dass es eine vergleichsweise hohe Übereinstimmung zwischen den Messzeitpunkten gab und die Urteile, die mit Hilfe des DBRs getroffen werden, keiner Willkür unterliegen. Christ, Riley-Tillman und Chafouleas (2009) sprechen sich für die Nutzung von DBRs in prozessdiagnostischen und statusdiagnostischen Bereichen aus. Da das DBR aber insbesondere im Rahmen dieser Arbeit hauptsächlich für die Verlaufsdiagnostik von Schülerverhalten eingesetzt werden soll, ist dieses Ergebnis eher in Bezug auf die Beurteilungsgüte verschiedener Verhaltensbereiche und Interrater-Reliabilität interessant.

Die Zusammenfassung der Studien, die spezifische Aspekte des DBRs untersucht haben, ist bei Huber und Rietz (2015) in verschiedene Bereiche aufgeteilt. Es werden das Skalendesign, die Anzahl der Items, die Anzahl der Messzeitpunkte, die Wahl des Beobachtungsziels, die

Formulierung des Beobachtungsziels, die Valenz der Zielformulierung, die Länge der Verhaltensstichproben und die Auswirkungen eines Beobachtertrainings betrachtet und später diskutiert. Auch Casale et al. (2015b) erwähnen in ihrem Artikel, wenngleich weniger ausführlich, ähnliche Bereiche.

Skalendesign

Da Messungen mit DBRs immer über Ratingskalen erfolgen, stellt sich diesbezüglich die Frage nach einer angemessenen Anzahl von Ausprägungen und der Beschriftung (numerisch, verbal, grafisch) der jeweiligen Skala. Briesch, Kilgus, Riley-Tillman, Christ und Chafouleas (2012), Chafouleas et al. (2009) und Christ, Riley-Tillman, Chafouleas und Boice (2010), deren Studien sämtlich auf Basis der Generalisierbarkeitstheorie (siehe oben) ausgewertet wurden, konnten keinen nennenswerten Einfluss der Skalenbreite bei sechs-, zehn- oder vierzehnstufigen Skalen auf die Varianz nachweisen, die durch die Lehrkraft aufgeklärt werden konnte. Auch auf die Beobachtungsgüte hat die Länge der Skalen keinen signifikanten Einfluss. Die Autor_innen hielten fest, dass die alleinige Veränderung der Skalierung die Lehrkräfte nicht zu genaueren oder ungenaueren Rater_innen macht. In einer Sekundäranalyse geben Christ, Riley-Tillman und Chafouleas (2009) an, dass eine mindestens sechsstufige Skala angebracht sei. Für diese Empfehlung lieferten sie keine empirische Grundlage (Huber & Rietz, 2015). Riley-Tillman, Christ, Chafouleas, Boice-Mallach und Briesch (2011) untersuchten außerdem die Art der Skalenbeschriftung und kamen zu dem Schluss, dass die Beschriftung, obgleich der Zeitanteil des zu beurteilenden Verhaltens prozentual oder absolut angegeben werden soll, keinen nennenswerten Einfluss auf die Interrater-Reliabilität hat. Unter den von Huber und Rietz (2015) in ihrem Review betrachteten Studien, befindet sich keine, die eine Skala mit zeitlicher Einschätzung mit einer Skala mit qualitativer Bewertung in Bezug auf die Erfüllung des Zielverhaltens vergleicht. Diesbezüglich bedarf es weiterer Forschung (Huber & Rietz, 2015). Casale et al. (2015) nennen außerdem Studien von Christ und Boice (2009) und von Volpe und Briesch (2012), in denen zum einen die Nutzung grafischer, unipolarer Ratingskalen mit kategorialen Items für den praktischen Einsatz und zum anderen die Verbindung jedes Skalengradienten mit einer eigenen Beschreibung empfohlen wird.

Anzahl der Items

Generell wird, wie oben bereits erläutert, bei DBRs zwischen Single-Item-Skalen und Multi-Item-Skalen unterschieden. Volpe und Briesch (2012) verglichen in einer Studie die Kriteriums-Validität von SIS und MIS. Dabei ergab sich, dass bei SIS (33 %) der Anteil der ungeklärten Varianz höher ist als bei MIS (5-26 %). Volpe und Briesch (2012) fanden außerdem heraus, dass der Unterschied zwischen SIS und MIS in Bezug auf die Messzeitpunktanzahl darin liegt, dass zur Herstellung einer akzeptablen Messgenauigkeit bei MIS weniger Messzeitpunkte notwendig sind als bei SIS. Auch insgesamt konnte für MIS eine höhere Messgenauigkeit festgestellt werden. Die Interpretation dieses Ergebnisses ist jedoch kritisch zu sehen, da es sich bei einer Skala mit vier verschiedenen Items zum einen natürlich um eine MIS handelt, aber, aus einem anderen Blickwinkel betrachtet, durchaus auch eine SIS mit vier Messungen vorliegen kann. Die Berechnung von Mittelwerten aus verschiedenen Messungen liefern erwartungsgemäß in der Regel stabilere Ergebnisse als einzelne Messungen (Huber & Rietz, 2015).

Anzahl der Messzeitpunkte

Um eine akzeptable Übereinstimmung mit Rater_innen zu erhalten, die den Verhaltensbereich „Teilnahme am Unterricht“ von Schüler_innen mit Hilfe direkter systematischer Verhaltensbeobachtungen beurteilen, benötigen Rater_innen, die mit DBRs arbeiten, fünf Beurteilungen mit einer jeweiligen Länge von 15 Minuten (Briesch,et al., 2010). Generell kann laut Christ et al. (2010) und Volpe und Briesch (2012) festgehalten werden, dass bei SIS zur Erreichung einer akzeptablen Beobachtungsgüte mehr Bewertungen nötig sind als bei MIS. Insgesamt sind die Ergebnisse der Studien sehr unterschiedlich und deshalb schwierig zu vergleichen (Huber & Rietz, 2015). Für gut erforschte Verhaltensbereiche, die im engen Bezug zum schulischen Kontext stehen, wie zum Beispiel Unterrichtsteilnahmen oder störendes Verhalten, sind bereits mindestens fünf Beobachtungen mit einer SIS ausreichend, um eine akzeptable Validität und Reliabilität zu erreichen. Dennoch bietet es sich an, zu längeren Verhaltensstichproben zu tendieren, um eine möglichst hohe diagnostische Güte zu erreichen (Huber & Rietz, 2015). Casale et al. (2015b) weisen außerdem darauf hin, dass die Urteilsgenauigkeit der Beurteilung des Verhaltens einer Schülerin oder eines Schülers durch die Bewertung mehrerer Personen erhöht werden kann.

Wahl des Beobachtungsziels

Hierbei stellt sich die Frage nach der Abhängigkeit der Messgenauigkeit von DBRs vom zu beobachtenden Zielverhalten. Wie bereits im Punkt „Anzahl der Messzeitpunkte“ angesprochen, gibt es tatsächlich Unterschiede zwischen verschiedenen Verhaltensweisen in Bezug auf die Validität und die Interrater-Reliabilität. Diese sind höher beziehungsweise die Diskrepanz zwischen den Ergebnissen aus DBRs und direkten systematischen Beobachtungen ist niedriger, wenn schulbezogene, leicht operationalisierbare Verhaltensweisen wie „Unterrichtsteilnahme“ und „störendes Verhalten“ beurteilt werden sollen (Christ et al. 2011). Etwas größere Unterschiede und eine damit verbundene geringere Beobachtungsgüte ergeben sich, wenn „angemessenes“ oder „respektvolles“ Verhalten beobachtet und bewertet werden soll. Noch gravierender sind die Unterschiede in den Bereichen „Interaktion mit der Lehrkraft“ oder „Interaktion mit den Klassenkameraden und Klassenkameradinnen“ sowie im motorischen und sprachlichen Bereich (Huber & Rietz, 2015).

Formulierung des Beobachtungsziels

Bei der Formulierung des Beobachtungsziels wird zwischen einer globalen und einer spezifischen Formulierung unterschieden (s. o.). Diese wurde von Christ et al. (2011) in Bezug auf ihren Einfluss auf die Validität und Reliabilität untersucht. Zur Bestimmung der Validität wurde auch hier wieder die Übereinstimmung mit zwei geschulten Rater_innen gewählt, die eine direkte systematische Verhaltensbeobachtung durchführten. Die globale Formulierung des Zielverhaltens ist hierbei in fast allen untersuchten Verhaltensbereichen im Rahmen von SIS einer spezifischen überlegen. Auch die Interrater-Reliabilität ist deutlich höher, wenn die Items global formuliert sind (r=0,61 – r=0,81 zu r=0,09 – r=0,60). Volpe und Briesch (2012) widersprechen diesen Ergebnissen und geben an, konkrete Items führten zu besseren Interraterübereinstimmungen und seien auch insgesamt genauer. Hier sind dementsprechend weitere Nachforschungen notwendig.

Valenz der Zielformulierung

Die Valenz der Formulierung des Zielverhaltens bestimmt, ob der Fokus der Items auf positivem, erwünschtem oder negativem, unerwünschtem Verhalten liegt. Riley-Tillman, Chafouleas, Christ, Briesch und LeBel (2009) und Chafouleas, Jaffery, Riley-Tillman, Christ und

Sen (2013) haben den Einfluss positiver und negativer Zielformulierungen bei SIS im Rahmen eines Vergleichs von DBRs und direkten systematischen Verhaltensbeobachtungen untersucht und herausgefunden, dass sich bei Beobachtungen von Unterrichtsteilnahme eine positive globale Zielformulierung günstig auf die Beobachtungsgenauigkeit auswirkt. Ebenso positiv wirkt sich die negative Formulierung des Verhaltensbereichs „störendes Verhalten“ auf die Beobachtungsgenauigkeit aus. Insgesamt kann festgehalten werden, dass es einfacher ist, die Länge beziehungsweise die Anwesenheit des störenden Verhaltens zu beurteilen als die Länge des nicht-störenden Verhaltens oder dessen generelle Abwesenheit (Huber & Rietz, 2015). Uneindeutige Resultate ergeben die Untersuchungen für den Bereich „Respektvolles Verhalten“. Hier empfehlen Chafouleas et al. (2013) eine negative Zielformulierung, also die Erfassung „Nicht respektvollen Verhaltens“. Die Ergebnisse aus einem Verhaltensbereich sind nicht auf andere Verhaltensbereiche übertragbar (Huber & Rietz, 2015).

Länge der Verhaltensstichproben

Riley-Tillman et al. (2011) untersuchten in einer Studie den Einfluss der Länge der Verhaltensstichproben auf die Beobachtungsgüte. Dabei stellten sie fest, dass die Test-Retest-Reliabilität zwischen den beobachteten Situationen und der Beobachtungszeit stark variiert. Außerdem ergaben die Analysen, dass die Test-Retest-Reliabilität für zehnminütige Zeiträume minimal geringer ist als für zwanzigminütige. Das Bilden von Mittelwerten mehrerer Ratings kann zu höheren Reliabilitäten führen als bei einzeln durchgeführten und ausgewerteten Ratings. Insgesamt waren die in oben genannter Studie ermittelten Test-Retest-Reliabilitäten eher nicht zufriedenstellend und bedürfen weiterer Forschung. Es kann daher festgehalten werden, dass zurzeit keine genauen Empfehlungen für die Länge der Verhaltensstichproben vorliegen.

Auswirkungen von Beobachtertrainings

Die Frage nach der Auswirkung von Beobachtertrainings auf die Messgenauigkeit von Instrumenten zum DBR stellten sich Schlientz, Riley-Tillman, Briesch, Walcott und Chafouleas (2009). Sie konnten herausfinden, dass die Messgenauigkeit trainierter Rater_innen signifikant höher war als die Beobachtungen einer untrainierten Kontrollgruppe. Bei den trainierten Beobachtern konnte eine höhere Interrater-Reliabilität und eine höhere Übereinstimmung mit den ebenfalls trainierten Beobachter_innen, die eine direkte systematische Verhaltensbe-

obachtung durchgeführt haben, festgestellt werden. Beim Vergleich drei verschiedener Gruppen, von denen eine kein Training, eine Gruppe eine kurze Schulung und eine ein langes und intensives Training erfahren hat, konnten LeBel, Kilgus, Briesch und Chafouleas (2009) erkennen, dass eine kurze Schulung möglicherweise bereits ausreicht, um bessere Ergebnisse zu erzielen. Zwischen der Gruppe mit kurzer Schulung und der mit intensivem Training ließen sich in der Studie von LeBel et al. (2009) keine signifikanten Unterschiede feststellen.

Insgesamt stellen sowohl Huber und Rietz (2015) als auch Casale, Hennemann, Huber und Grosche (2015) auf Grundlage oben aufgeführter Studienergebnisse fest, dass sich DBRs prinzipiell für die Verhaltensverlaufsdiagnostik eignen, da sie zum einen, wie bereits erwähnt, durch ihre Ökonomie die schnelle Entstehung einer soliden Datengrundlage ermöglichen und zum anderen in Bezug auf die Anforderungen, die im Rahmen von Testgütekriterien an die Verhaltensverlaufsdiagnostik gestellt werden, insgesamt positiv bewertet werden können. Die Forschungsergebnisse weisen zwar daraufhin, dass die Beobachtungsgüte von DBRs im Vergleich zu der direkter systematischer Beobachtungen geringer ausfällt und stärker von der beobachtenden Person abhängig ist, jedoch lassen sich aus den genannten Forschungsarbeiten Möglichkeiten zur Verbesserung der Beobachtungsgüte ablesen (Huber & Rietz, 2015). So hängt die Testqualität in erheblichem Maße von Testdesign und Vorgehen ab (Casale et al., 2015). Die Validität und die Reliabilität von DBRs lassen sich durch das zu beobachtende Zielverhalten beeinflussen. Diese Faktoren steigen bei der Fokussierung von Zielverhalten an, welches im schulischen Kontext gut beobachtbar und operationalisierbar ist. Auch die Kriteriumsvalidität kann durch das Testdesign beeinflusst werden. Zu ihrer Erhöhung empfiehlt es sich, mehrere spezifische Verhaltensaspekte im Rahmen einer MIS zu beurteilen. Dies erhöht jedoch die Komplexität und damit die Schwierigkeit für die Rater_innen. Insgesamt sind positive und globale Formulierungen sinnvoll, bei denen eher die Anwesenheit eines Verhaltens erfasst werden soll. Die Skalenbreite hat wiederum keinen signifikanten Einfluss auf die Beobachtungsgüte. In Bezug auf die Messzeitpunkte konnten die Studien zeigen, dass es sich empfiehlt, zu mindestens fünf Zeitpunkten zu erheben, da bei weniger als fünf Zeitpunkten die Streuung der Messwerte der unterschiedlichen Rater_innen zu groß ist. Die Varianzaufklärung durch die Rater_innen wäre dann höher als die durch die Verhaltensunterschiede. Abschließend kann außerdem erwähnt werden, dass sich eine kurze Schulung der Rater_innen positiv auf die Validität und Reliabilität auswirkt (Huber & Rietz, 2015).

Trotzdem bedarf es insbesondere im deutschsprachigen Raum weiterer Forschungen in Bezug auf DBRs, ihre Eignung in verschiedenen Bereichen, die Art des Testdesigns und dessen Auswirkung auf die Testgüte.

3. Fragestellung

Die vorliegende Arbeit untersucht die Möglichkeit, die Fremdbeurteilungsversion des deutschsprachigen Strengths and Difficulties Questionnaires (SDQ) in ein praktikables Instrument abzuwandeln, mit dem eine direkte Verhaltensbeurteilung im Rahmen von Verhaltensverlaufsdiagnostik im schulischen Kontext geleistet werden kann. Hierbei soll der Fokus auf der Anwendbarkeit in der Praxis, der Interrater-Reliabilität und der stetigen Veränderung des Instruments unter dem Aspekt seiner Durchführbarkeit liegen. Der dreigeteilten Fragestellung entspricht die Struktur des vierten Kapitels, in welchem der Prozess bis zur vorerst finalen Fertigstellung des Instruments eines DBRs präsentiert wird. Zu Beginn soll in Kapitel 4.1 in einer theoretischen Rahmung die Frage geklärt werden, ob sich der SDQ in ein im schulischen Kontext anwendbares DBR umwandeln lässt. Die Frage nach der Anwendbarkeit beziehungsweise Praktikabilität des DBRs in der dieser Arbeit zugrundeliegenden Form wird im gesamten vierten Kapitel wiederholt aufgegriffen werden müssen, da sich diesbezüglich nach jeder Erprobung des Instruments in der Praxis neue Erkenntnisse ergeben.

Im zweiten Schritt wird die so entwickelte Version des DBRs inhaltlich validiert. Dies geschieht basierend auf den Rückmeldungen stichprobenartig ausgewählter Expertinnen aus dem schulischen Kontext. Hierbei wird der Blick speziell auf die Frage nach der Anwendbarkeit und der inhaltlichen Durchführbarkeit des DBRs gerichtet. Dieser Prozess wird in Kapitel 4.2 erläutert.

Abschließend wird in Kapitel 4.3 dargestellt, wie die in Kapitel 4.2 überarbeitete Version des DBRs erneut erprobt und mit Fokus auf seine Interrater-Reliabilität untersucht sowie ein weiteres Mal auch inhaltlich revidiert wird. Als Ergebnis liegt eine finale Version des Instruments zur direkten Verhaltensbeurteilung im schulischen Kontext vor, welche im Anschluss an seine Entwicklung im Rahmen einer Pilotierungsstudie untersucht wird.

In Kapitel 4.4 werden die Untersuchungsergebnisse aus der Pilotierungsstudie zusammengefasst, welche in diversen Grund- und Gesamtschulen durchgeführt wurde, wobei das Resultat des geschilderten Prozesses als finales Instrument zum DBR zur Anwendung kam. Im Fokus

soll hier einerseits die methodische Weiterentwicklung des vorliegenden verhaltensverlaufsdiagnostischen Instruments stehen, andererseits soll erneut bedacht werden, welche Schlüsse aus der Studie für die generelle Entwicklung und Gestaltung eines DBRs gezogen werden können.

Die Forschungsfragen stehen unter der übergreifenden Fragestellung:

Wie kann der Strengths and Difficulties Questionnaire zu einem Direct Behavior Rating umgewandelt werden und in der Praxis gute Ergebnisse erzielen?

Die einzelnen Forschungsfragen lauten entsprechend:

1) Wie wird das entwickelte Instrument zum Direct Behavior Rating von den Lehrkräften angenommen?
2) Was muss verändert und angepasst werden, um das entwickelte Instrument zum Direct Behavior Rating in der Praxis anwendbar zu machen?
3) Wie reliabel ist das entwickelte Instrument zum Direct Behavior Rating, wenn es im schulischen Kontext durchgeführt wird?

   3 a) Wie hoch ist die Interrater-Reliabilität des entwickelten Instruments im schulischen Kontext?

   3 b) Wie muss das vorliegende Instrument weiterentwickelt werden, um im Rahmen einer Verhaltensverlaufsdiagnostik reliable Ergebnisse zu liefern?

4. Überarbeitung und Erprobung eines Instruments zum Direct Behavior Rating

Zur Untersuchung der oben genannten Fragestellungen wurde das Instrument zum DBR in mehreren Stufen überarbeitet und in der Praxis erprobt. Im folgenden Kapitel wird dieser Prozess in einzelnen Schritten dargestellt. Am Anfang stand die Adaption eines Instruments zur Verhaltensverlaufsdiagnostik in Anlehnung an den SDQ (siehe hierzu Kapitel 2.3). Erste Schritte in diese Richtung machten bereits Gebhardt, Casale, Jungjohann und DeVries (2017). Diese wurden im theoretischen Rahmen und unter Berücksichtigung zusätzlicher Verhaltensbereiche nachvollzogen und weitergeführt. Daran anschließend wurde das zu diesem Zeitpunkt aktuelle Instrument zum DBR in eine Schule gegeben und von stichprobenartig ausgewählten Lehrkräften zur Untersuchung von Fragestellung 1) erprobt. Die Rückmeldungen der Lehrkräfte wurden im Rahmen von Expert_inneninterviews im Anschluss an die Erprobung gewonnen. Anschließend fand im Rahmen von Fragestellung 2) eine Anpassung der DBRs

auf Grundlage der in den Interviews gesammelten Informationen statt. Die überarbeitete Version des Instruments wurde im nächsten Schritt genutzt, um seine Interrater-Reliabilität im schulischen Kontext mit Blick auf Fragestellung 3 a) zu überprüfen. Danach wurde auf Basis erneuter qualitativer Rückmeldungen der Rater_innen und Hinweisen eines externen Experten eine weitere Anpassung der Items durchgeführt und eine vorerst finale Version des DBRs erstellt, welche in einer an diese Arbeit anschließenden Pilotierungsstudie im größeren Rahmen untersucht wurde und zur Beantwortung von Fragestellung 3 b) beitrug. Alle diese Überprüfungen und Anpassungen mündeten schließlich in der Beantwortung der übergeordneten Fragestellung: Wie kann der SDQ zu einem DBR umgewandelt werden und in der Praxis gute Ergebnisse erzielen?

4.1 Adaption des Strengths and Difficulties Questionnaires

Der SDQ, auf den in Kapitel 2.3 im Detail eingegangen wurde, wurde zur Statusdiagnostik konzipiert und diesbezüglich vielfach im Hinblick auf seine Testgüte überprüft und evaluiert. Mit seiner Hilfe können Problemfelder und Ressourcen in verschiedenen Verhaltensbereichen rückwirkend erfasst werden. Er ermöglicht eine Einordnung der Verhaltensprobleme in einen oder mehrere bestimmte Bereiche, die im Anschluss genauerer Betrachtung und eventuell einer Intervention bedürfen. Es bietet sich daher an, mit Hilfe des SDQs vor der Erfassung von Verhaltensverläufen statusdiagnostisch zu überprüfen, in welchen Verhaltensbereichen Schwierigkeiten vorliegen und ob deren Überprüfung sich im Rahmen eines DBRs anbieten würde. Gebhardt und Voß diskutieren in ihrem Artikel „Monitoring der sozial-emotionalen Situation von Grundschülerinnen und Grundschülern – Ist der SDQ ein geeignetes Verfahren?“ (2017) deshalb die Frage, ob der SDQ in ein Instrument zur Verhaltensverlaufsdiagnostik weiterentwickelt werden kann. Sie kommen zu dem Schluss, dass dies durchaus möglich ist, es jedoch der Konstruktion weiterer Items bedarf. Der vorliegenden Arbeit liegt ein Prototyp eines solchen aus dem SDQ entwickelten Instruments zum DBR zugrunde (siehe Anhang A). Auf Basis dieses Prototyps soll als Beitrag zur Untersuchung von Fragestellung 1) im Folgenden eine erste Version eines DBRs erstellt werden. Diese erste Version wird im Anschluss zur Überprüfung in eine Schule gegeben und dort in der Praxis erprobt.

4.1.1 Beschreibung der Ausgangsversionen

Als Ausgangsversion für diese Arbeit dienen die deutschsprachige Fremdbeurteilungsversion des SDQs (sdqinfo.com), die von Gebhardt et al. (2017) um den Bereich des Schulbezogenen

Verhaltens erweitert wurde (siehe Anhang G), und eine prototypische Version eines daraus entstandenen Instruments für die Verhaltensverlaufsdiagnostik, die ebenfalls von Gebhardt et al. (2017) entwickelt wurde (siehe Anhang A). Die Ausgangsversion des SDQs umfasst in der Originalversion die Bereiche Emotionale Probleme (EP), Verhaltensprobleme mit Gleichaltrigen (VPG), Verhaltensprobleme (VP), Hyperaktivität (HY) und Prosoziales Verhalten (PS). In der vorliegenden Version wurde der Verhaltensbereich Schulbezogenes Verhalten (SV) hinzugefügt, da ein enger Bezug zwischen Verhaltensauffälligkeiten und Lernschwierigkeiten gegeben ist, die sich auch in ebensolchen, auf den schulischen Kontext bezogenen Verhaltensweisen widerspiegeln können (siehe hierzu Kapitel 2.1 oder auch DeVries, 2018). Der SDQ umfasst dementsprechend in dieser Version 30 Items. Im Original kann beim Ausfüllen des SDQs zwischen „nicht zutreffend“, „teilweise zutreffend“ und „zutreffend“ entschieden werden. Die Kategorie „teilweise zutreffend“ wurde für diese Arbeit in Rückbezug auf Gebhardt und Voß (2017) und im Sinne der Vereinfachung von Vergleichen über die Zeit und Interpretationen von Verhaltensverlaufskurven entfernt. Aus dieser Version des SDQs wurde in Anlehnung an die vorliegende prototypische Version (siehe Anhang A), die nur vier der Verhaltensbereiche abdeckte (SV, VP, HY, EP), ein DBR erstellt, das mit jeweils drei Items sechs verschiedene Verhaltensbereiche abdeckt (EP, VPG, VP, HY, PS, SV). Um das Instrument möglichst ökonomisch zu gestalten, wurde die Anzahl der Items des SDQs von fünf auf drei reduziert.

Die Skalierung des Prototyps wurde übernommen und besteht aus einer siebenstufigen Likert-Skala. Eine Likert-Skala beschreibt eine nach R. Likert benannte Methode zur Skalierung persönlicher Urteile. Dies erscheint im Bereich der Verhaltensverlaufsdiagnostik adäquat, da hierbei ebensolche Einschätzungen durch Lehrkräfte zum Tragen kommen. Bei der Likert-Skala wird eine gestufte, unipolare Antwortskala mit einer Aussage verknüpft, z. B. mit „Das Kind verhält sich aggressiv.“ Meist sind die verschiedenen Stufen etikettiert mit „trifft voll und ganz zu“, „trifft zu“, „teils/teils“, „trifft nicht zu“ und „trifft überhaupt nicht zu“ oder ähnlichem (Wirtz, 2017). Die schon bei der Gestaltung der Ursprungsversion des DBR-Instruments getroffene Entscheidung für eine siebenstufige Likert-Skala basiert unter anderem auf Untersuchungen von Preston und Colman (2000). Diese ergaben, dass eine siebenstufige Antwortskala in Bezug auf Reliabilität und Validität am vorteilhaftesten ist. Zwar steigen Reliabilität und Validität weiter an, wenn mehr als sieben Antwortkategorien gegeben sind, jedoch fällt in Folge dessen die Modellpassung schlechter aus (Preston & Colman, 2000). Für den

Einsatz einer Likert-Skala spricht zudem, dass der SDQ ebenfalls mit ihr in dreistufiger Form arbeitet und auch Christ et al. (2009) eine mindestens sechsstufige Likert-Skala empfehlen.

Die Skala ist unipolar aufgebaut, beginnend bei 1 beziehungsweise „Nie“ und aufsteigend bis 7 beziehungsweise „Immer“. Diese beiden Extremwerte sind, einer unipolaren Ratingskala entsprechend, durch gegensätzliche Begriffe beschrieben und gehen in eine Richtung. Die Beschriftung der Skala ist numerisch mit Ziffern von eins bis sieben mit Randmarkierungen („Nie“ und „Immer“) gewählt, sodass von einer Häufigkeitsskala zu sprechen ist (Bühner, 2011). Die Zahlen zwei bis sechs sind verbal unbeschriftet. In jeder Zeile, die ein Item beinhaltet, finden sich die Zahlen eins bis sieben wieder und können zur Erhöhung der Praktikabilität beim Ausfüllen direkt, z. B. durch Ankreuzen oder Einkreisen, markiert werden.

Es handelt sich bei der prototypischen Version und allen in der vorliegenden Arbeit folgenden Versionen des DBRs um MIS. Die insgesamt 12 Items der prototypischen Version des DBRs sind nach Verhaltensbereichen gruppiert und durchnummeriert. Zuerst findet sich in jedem Abschnitt die den Bereich benennende Überschrift des entsprechenden Verhaltens, der die jeweils drei zugehörigen Items folgen. Diese sind, unabhängig von ihrer Zuordnung zu einem Verhaltensbereich, fortlaufend von eins bis zwölf nummeriert. Die Abkürzungen der einzelnen Verhaltensbereiche sind hinter der ausgeschriebenen Version in Klammern angegeben.

Da immer nur drei der fünf Items, häufig abgeändert, aus dem SDQ in das prototypische Instrument zum DBR übernommen wurden, mussten jeweils zwei Items aussortiert werden. Das Schulbezogene Verhalten wurde mit folgenden drei Items in den Prototypen aufgenommen: „Meldet sich im Unterricht“ (1), „Arbeitet ruhig und konzentriert im Unterricht“ (2) und „Bleibt während des Unterrichts ruhig am Platz sitzen, wenn dies erforderlich ist“ (3). In der Originalversion ist der Bereich des Schulbezogenen Verhaltens nicht enthalten. Deshalb wurden, wie von Gebhardt und Voß (2017) gefordert, Items aus diesem Bereich hinzugefügt, die eine Anwendung des SDQs im schulischen Kontext sinnvoller gestalten und eine Erfassung dieses Verhaltensbereichs ermöglichen.

Die Items „Verhält sich wütend und aufbrausend“ (4), „Hört nicht auf die Lehrkraft und verhält sich nicht regelkonform“ (5) und „Streitet sich mit anderen Kindern/schikaniert seine MitschülerInnen“ (6) wurden zur Beschreibung von Verhaltensproblemen in das DBR aufgenommen. Im SDQ heißen die Items zu Verhaltensproblemen „Hat oft Wutanfälle; ist aufbrausend“, „Im Allgemeinen folgsam; macht meist, was Erwachsene verlangen“, „Streitet sich oft mit

anderen Kindern oder schikaniert sie“, „Lügt oder mogelt häufig“ und „Stiehlt zu Hause, in der Schule oder anderswo“. Die beiden letztgenannten Items wurden nicht in den Prototypen des DBRs übernommen. Dieser Entscheidung lag zum einen die Überlegung zugrunde, dass das Lügen und Mogeln in der Regel nur in bestimmten Situationen, zum Beispiel bei der Kontrolle (nicht-)angefertigter Hausaufgaben, in Klassenarbeiten oder in Spielsituationen gezeigt wird und dementsprechend nicht häufig beobachtet und beurteilt werden kann. Zum anderen wurde durch sie der Tatsache Rechnung getragen, dass Lügen und Mogeln zwar zu den externalisierenden Verhaltensauffälligkeiten und damit zum Bereich der Verhaltensprobleme zählen, diese aber im schulischen Kontext häufiger in Form von oppositionellem und aggressivem Verhalten auftreten (siehe Kapitel 2.1.1). Deshalb erschien es günstiger, nur jene in der Schule häufiger gezeigten Verhaltensweisen zu erfassen. Das Item „Stiehlt zu Hause, in der Schule und anderswo“ kann, bei ausschließlichem Bezug auf das schulische Vorgehen, nur zu einem Drittel von Lehrkräften ausgefüllt werden, da diese in der Regel zwar, beispielsweise durch Elterngespräche, Einblicke in das häusliche und familiäre Umfeld der Kinder bekommen, aber in der Regel nicht unbedingt erfahren, ob das Kind dort oder gar „anderswo“ stiehlt. Außerdem zählt auch das Stehlen, ähnlich wie das Mogeln und Lügen, zu den Norm- und Eigentumsverletzungen und damit zu einer seltener im schulischen Kontext auftretenden Form der Verhaltensprobleme. Auch hier erschien es deshalb passend, es vorerst nicht in ein zur Verlaufsdiagnostik von Verhalten konzipiertes Instrument aufzunehmen.

Im Bereich der Hyperaktivität finden sich im Prototypen des DBRs die Items „Zappelt und ist motorisch unruhig“ (7), „Führt Aufgaben zu Ende“ (8) und „Unruhig und überaktiv“ (9). Im SDQ werden diese Items unter folgenden Namen geführt: „Unruhig, überaktiv, kann nicht lange stillsitzen“, „Ständig zappelig“, „Führt Aufgaben zu Ende; gute Konzentrationsspanne“. Außerdem sind im SDQ die Items „Leicht ablenkbar, unkonzentriert“ und „Denkt nach, bevor er/sie handelt“ vorhanden. Die letztgenannten Items sind nicht im Prototypen des DBRs enthalten. Die Nichtaufnahme des Items „Denkt nach, bevor er/sie handelt“ kann darauf zurückgeführt werden, dass es zum einen kaum beobachtbar ist und die soziale Validität dementsprechend eingeschränkt wäre und es zum anderen nicht direkt zum Bereich des Hyperaktiven Verhaltens gezählt werden kann. Zwar erscheinen die Handlungen hyperaktiver Kinder häufig unbedacht und planlos, jedoch stellt sich die Frage, ob das Kind mit vorheriger Reflexion die Handlung unterdrücken könnte beziehungsweise zu einem anderen Verhalten tendieren würde

oder ob im Rahmen der Hyperaktivität manche Verhaltensweisen derartig impulsiv hervortreten, dass ein Nachdenken kaum möglich wäre (siehe Kapitel 2.1.1). Das Item „Leicht ablenkbar, unkonzentriert“ wurde in dieser Form nicht in den Prototypen aufgenommen, weil mit dem Item „Führt Aufgaben zu Ende; gute Konzentrationsspanne“ ein ähnliches, aber in der Valenz der Zielformulierung gegensätzliches Item aufgeführt ist. Die Unruhe der beobachteten Kinder, die das äquivalente Item aus dem SDQ außerdem durch den Begriff „ablenkbar“ registriert, wird in den anderen beiden im Prototypen enthaltenen Items angesprochen.

Zu den emotionalen Problemen finden sich im SDQ die Items „Klagt häufig über Kopfschmerzen, Bauchschmerzen oder Übelkeit“, „Hat viele Sorgen; erscheint häufig bedrückt“, „Oft unglücklich oder niedergeschlagen, weint häufig“, „Nervös oder anklammernd in neuen Situationen, verliert leicht das Selbstvertrauen“ und „Hat viele Ängste; fürchtet sich leicht“. Aus diesen fünf Items wurden, wie bei den anderen Verhaltensbereichen, drei abgeändert und in den Prototypen des DBRs übernommen. Diese drei Items lauten „Wirkt besorgt, betrübt oder bedrückt“ (10), „Verhält sich ängstlich/fürchtet sich“ (11) und „Verhält sich nervös und klammert sich an Erwachsene“ (12). Die Items in Bezug auf Sorgen, Bedrücktheit und Niedergeschlagenheit sind in einem Item zusammengefasst worden. Nicht übernommen wurde in diesem Fall das Weinen, welches insbesondere in kurzen Beobachtungssituationen (z. B. eine Schulstunde oder weniger) über einen begrenzten Zeitraum (z. B. eine Schulwoche) selten in beobachtbarem Maße auftritt und deshalb ein eher statusdiagnostisch zu ermittelndes Merkmal darstellt. Ebenso schwierig verhaltensverlaufsdiagnostisch zu erfassen sind Übelkeit, Kopf- und Bauchschmerzen, welche in der Regel nicht mehrfach täglich oder wöchentlich auftreten und deshalb nicht explizit in den Items erwähnt werden. Die Gefühle der Angst und Furcht werden aus dem SDQ in gekürzter Version übernommen, ebenso Nervosität und Anklammern. Der Aspekt des Selbstvertrauens wurde ausgelassen, da es sich hierbei um ein relativ schwer zu beobachtendes Konstrukt handelt, bei dessen Beobachtung die Lehrkraft wahrscheinlich stark auf Informationen aus den vorhergehenden Wochen zurückgreifen würde und das sich daher eher für statusdiagnostische Erfassungen anbietet.

Die prototypische Version enthält eine erste Version einer Instruktion für Lehrkräfte, welche die DBR-Bögen ausfüllen sollen. Diese umfasst eine kurze Erklärung zur Intention der Durchführung von Verhaltensverlaufsdiagnostik, eine Erklärung zur Durchführungsweise und einen Hinweis auf die im Vorhinein mit Hilfe des SDQs zu erhebenden, für die Erfassung mit dem

DBR-Instrument relevanten Verhaltensbereiche. Die Instruktion befindet sich ebenfalls in Anhang A.

4.1.2 Eingliederung von Verhaltensbereichen des Strengths and Difficulties Questionnaires in das Instrument zum Direct Behavior Rating

Die erste Überarbeitung des Instruments zum DBR fand auf theoretischer Basis statt. Hierbei wurden die deutschsprachige Fremdbeurteilungsversion des SDQs und der Prototyp des DBRs zusammengeführt. Der Prototyp des DBRs enthielt die Bereiche SV, VP, HY und EP. Der SDQ, welcher als statusdiagnostisches Instrument zum einen vor der Anwendung des DBRs als Screening und Einschätzung der schwierigen Verhaltensbereiche durchgeführt werden sollte und zum anderen als Vorbild bei der Entwicklung des Prototyps genutzt wurde, enthielt zusätzlich die Bereiche „Verhaltensprobleme mit Gleichaltrigen“ und „Prosoziales Verhalten“. Diese beiden Bereiche fehlten der prototypischen DBR-Version und wurden deshalb nachträglich hinzugefügt. Hierbei wurden, wie oben ausgeführt, aus jeweils fünf Items des SDQs drei Items für das DBR ausgewählt.

Der Verhaltensbereich des prosozialen Verhaltens ist im SDQ in folgende Items unterteilt: „Rücksichtsvoll“, „Teilt gerne mit anderen Kindern (Süssigkeiten [sic!], Spielzeug, Buntstifte usw.)“, „Hilfsbereit, wenn andere verletzt, krank oder betrübt sind“, „Lieb zu jüngeren Kindern“ und „Hilft anderen oft freiwillig (Eltern, Lehrern oder anderen Kindern)“ (sdqinfo.com). Die Items „Rücksichtsvoll“, „Hilfsbereit, wenn andere verletzt, krank oder betrübt sind“ und „Teilt gerne mit anderen Kindern (Süssigkeiten [sic!], Spielzeug, Buntstifte usw.)“ wurden hier ausgewählt und umformuliert, sodass in die erweiterte Version des Prototyps die Items eingegangen sind: „Verhält sich anderen gegenüber rücksichtsvoll“ (13), „Verhält sich anderen gegenüber hilfsbereit“ (14) und „Teilt gerne mit anderen Kindern“ (15).

Diese Auswahl erfolgte im Hinblick auf den Einsatz des Instruments im schulischen Kontext, weshalb das Item „Lieb zu jüngeren Kindern“ nicht übernommen wurde. Da der Unterricht im deutschsprachigen Raum in der Regel im Klassenkontext und in verhältnismäßig altershomogenen Gruppen stattfindet, können Situationen mit jüngeren Kindern fast ausschließlich in Pausensituationen beobachtet werden. DBR-Instrumente können zwar grundsätzlich auch auf Situationen außerhalb des Klassenkontexts angewendet werden. Da in der Erprobungsphase jedoch hauptsächlich auf den Unterricht innerhalb der Klassengemeinschaft eingegangen wurde, bot es sich an, dieses Item von vornherein auszuschließen. Ebenfalls nicht (wörtlich)

übernommen wurde das Item „Hilft anderen oft freiwillig (Eltern, Lehrern oder anderen Kindern)“; es war mit einem weiteren Item aus dem SDQ („Hilfsbereit, wenn andere verletzt, krank oder betrübt sind“) im neu formulierten Item „Verhält sich anderen gegenüber hilfsbereit“ enthalten. Da Prosozialität, wie in Kapitel 2.1.1 beschrieben, grundsätzlich auf Freiwilligkeit basiert, wurden diese beiden Items zusammengefasst.

Auch der Verhaltensbereich „Verhaltensprobleme mit Gleichaltrigen“ ist im SDQ in fünf Items unterteilt, von denen wiederum drei ausgewählt und umformuliert wurden, um die Struktur zu wahren und das DBR-Instrument möglichst ökonomisch zu gestalten. Heißen die SDQ-Items „Einzelgänger, spielt meist alleine“, „Hat wenigstens einen guten Freund oder eine gute Freundin“, „Im Allgemeinen bei anderen Kindern beliebt“, „Wird von anderen Kindern gehänselt oder schikaniert“ und „Kommt besser mit Erwachsenen aus als mit Kindern“, lauteten die im DBR aufgeführten Items nun „Spielt meist allein“ (16), „Ist bei anderen Kindern beliebt“ (17) und „Kommt besser mit Erwachsenen als mit anderen Kindern aus“ (18). Die nicht übernommenen Items „Hat wenigstens einen guten Freund oder eine gute Freundin“ und „Wird von anderen Kindern gehänselt oder schikaniert“ erschienen für eine verhaltensverlaufsdiagnostische Beurteilung nicht geeignet. Das Item „Hat wenigstens einen guten Freund oder eine gute Freundin“ bezieht sich auf einen eher statusdiagnostischen Bereich. Da zudem insbesondere bei jüngeren Schüler_innen häufige Wechsel von Freundschaftsbeziehungen stattfinden können, die eine Einordnung dieses Bereichs erschweren, wurde es nicht aufgenommen (Reinders, 2003). „Wird von anderen Kindern gehänselt oder schikaniert“ wurde ebenfalls nicht in das DBR integriert. Das liegt daran, dass das Item sehr passiv formuliert ist und dementsprechend nicht ausschließlich dem beobachteten Kind zugewiesen werden kann, sondern immer zu einem großen Teil auch von den Mitschüler_innen abhängt. Eine Beurteilung der Gründe für Schikanen und Hänseleien wäre notwendig, damit das Item in Bezug auf den zu beurteilenden Verhaltensbereich aussagekräftig wäre. Es müsste immer geklärt werden, inwiefern das geärgerte Kind eine solche Behandlung durch seine Mitschüler_innen provoziert hat oder ihr unverschuldet zum Opfer gefallen ist. Nur so könnte eingeschätzt werden, ob es sich tatsächlich um ein Problem handelt, welches beim beobachteten Kind selbst liegt oder das eher den Mitschüler_innen, dem Klassenkontext oder anderen Faktoren zuzuschreiben ist.

4.1.3 Beschreibung der ersten überarbeiteten Version des Direct Behavior Ratings

Die erste überarbeitete Version des Prototyps des DBRs (Version 1), welche in dieser Form im nachfolgenden Schritt zum ersten Mal zur Erprobung in die Schule gegeben wurde, enthält alle fünf Verhaltensbereiche des SDQs sowie das Schulbezogene Verhalten (siehe Anhang B). Jeder Verhaltensbereich wird von drei wie im Prototypen von eins bis achtzehn nummerierten Items repräsentiert. Die Verhaltensbereiche sind in der gleichen Reihenfolge angeordnet wie im Prototypen (SV, VP, HY, EP). Die Items zum Prosozialen Verhalten und zu den Verhaltensproblemen mit Gleichaltrigen folgen im Anschluss an die vier im Prototypen enthaltenen DBR-Instrumente.

Die Items aus den Bereichen SV, VP, HY und EP sind aus der prototypischen Version übernommen. Neu eingegliedert wurden, wie in Kapitel 4.1.2 erläutert, die Items 13-18 aus den Verhaltensbereichen Prosoziales Verhalten (PS) und Verhaltensprobleme mit Gleichaltrigen (VPG).

Die erste überarbeitete Version des DBRs kann auf unterschiedliche Weise zusammengefasst werden und entweder als ein Instrument gesehen werden, das aus vier Dimensionen besteht (Externalisierendes Verhalten (VP und HY), Internalisierendes Verhalten (VPG und EP), Prosoziales Verhalten und Schulbezogenes Verhalten), oder aber als ein Instrument, das sich aus sechs Dimensionen zusammensetzt (SV, VP, HY, EP, VPG, PS). Auch kann dieses prinzipiell als MIS konzipierte Rating ebenfalls als SIS betrachtet werden, indem immer nur ein einzelnes Item herausgenommen und bewertet wird.

Die Skalierung und die Beschriftungen wurden ebenfalls vom Prototypen übernommen, sodass es sich auch bei der ersten überarbeiteten Version um eine siebenstufige Likert-Skala mit Beschriftungen („Nie“ und „Immer“) an den Extremwerten und einer Durchnummerierung mit den Ziffern 1-7 handelt (siehe Kapitel 4.1.1).

Bei den Items der ersten Version des DBRs, welche aus dem SDQ übernommen und dementsprechend in Anlehnung an Studien konzipiert wurden, handelt es sich um nomothetische Items. Ein idiografischer Ansatz wäre für die Entwicklung eines Instruments zur Verhaltensverlaufsmessung einer größeren Kohorte nicht umsetzbar. Das Individuum steht infolgedessen eher im Hintergrund, während die Testgüte der Items bereits im Vorhinein auf ihre psychometrische Eignung im Rahmen der Überprüfung der Qualität des SDQs getestet wurde (siehe

hierzu Kapitel 2.4). Des Weiteren sind die Items der vorliegenden ersten Version des DBRs weitgehend eher spezifisch als global formuliert. Dies kommt zum einen der Inferenz zugute, zum anderen sind spezifische Items in ihrer Bewertung weniger aufwendig als global formulierte und bieten sich deshalb im Rahmen der ökonomischen Gestaltung eines Instruments zur Verhaltensverlaufsdiagnostik an. Laut Christ et al. (2011) ist die Interrater-Reliabilität jedoch für global formulierte Items höher (siehe hierzu Kapitel 2.4). Volpe und Briesch (2012) widersprechen dem und geben an, dass konkrete Items zu besseren Interraterübereinstimmungen führen. Die Interrater-Reliabilität wird in Kapitel 4.3 im Rahmen der Beantwortung von Forschungsfrage 3 a) untersucht.

Auch die Instruktion wurde mit einigen kleinen Veränderungen übernommen: Sie wurde gekürzt und um die Bitte ergänzt, mindestens drei Verhaltensbereiche auszufüllen, damit trotz eines kleinen Stichprobenumfangs verwertbare Ergebnisse erzielt werden. Auch diese Instruktion findet sich in Anhang B wieder.

4.2 Inhaltliche Validierung des Direct Behavior Ratings anhand von Expert_inneninterviews

Im Anschluss an die Überarbeitung des Prototyps wurde die daraus entstandene Version 1 für das DBR im schulischen Kontext zusammen mit der um den Bereich des sShulbezogenen Verhaltens erweiterten Version des SDQs in eine Schule gegeben. Hier sollte das Instrument zur Verhaltensverlaufsdiagnostik zum ersten Mal in der Praxis erprobt werden. Diese Erprobung hatte eine auf die Items und die Form des DBRs ausgerichtete inhaltliche Validierung und die Anpassung der Items entsprechend den Ergebnissen der Expert_inneninterviews zum Ziel. Der Fokus lag demnach auf den Rückmeldungen der Lehrkräfte, die im Anschluss an die Erprobung erhoben wurden. Indem eine Annäherung der Items an den pädagogischen Alltag angestrebt wurde (Casale et al. 2015b), sollte der häufig von Lehrkräften geäußerten Kritik begegnet werden, Diagnostik mangele es an sozialer Validität.

Im Folgenden wird die Durchführung dieser ersten Erprobung beschrieben. Zuerst wird die für die Erprobung der ersten Version des DBRs ausgewählte Schule kurz vorgestellt. Im Anschluss daran wird die stichprobenartige Wahl der Lehrkräfte erläutert sowie die Art und Weise, wie die Lehrkräfte das Instrument angewendet haben. Auch wird auf die den Lehrkräf-

ten im Vorhinein an die Hand gegebene Instruktion eingegangen. Danach werden die Ergebnisse der Interviews präsentiert und ausgewertet und die auf dieser Basis vorgenommene Adaption des DBR vorgestellt.

4.2.1 Beschreibung der Durchführung

In der vorliegenden Arbeit wurden zur Untersuchung von Fragestellung 2) (Was muss verändert und angepasst werden, um das entwickelte Instrument zum Direct Behavior Rating in der Praxis anwendbar zu machen?) sowohl die überarbeitete Version des SDQs als auch die aus dem Prototypen hervorgegangene Version 1 des DBRs von drei Lehrkräften an einer Förderschule für den Förderschwerpunkt Lernen durchgeführt. Die Wahl der Schule lässt sich zum einen damit erklären, dass bereits Kontakte zur Institution und einigen der dort unterrichtenden Lehrkräfte existierten. Zum anderen sollten die Lehrkräfte, die Rückmeldungen zu Version 1 des DBRs gaben, bereits einige Erfahrung mit Tests, Ratings und Diagnoseinstrumenten im Allgemeinen haben. Dies ermöglichte es, die Item- und Testdesignkritik auf Einschätzungen von Lehrkräften mit Erfahrungen in diesem Bereich zu basieren. Auf die Auswahl der Lehrkräfte und der Schüler_innen soll in Kapitel 4.2.2 näher eingegangen werden.

Die Lehrkräfte wurden auf elektronischem Weg schriftlich kontaktiert, erhielten erste Informationen zum Vorhaben und wurden nach ihrer Bereitschaft zur Teilnahme gefragt. Nach der positiven Rückmeldung aller drei Lehrkräfte wurden diesen die beiden Rating-Instrumente (SDQ und DBR) sowie detaillierte Instruktionen zum genauen Ablauf der Erprobung zugeschickt. Letztere beinhalteten Informationen zum Ziel dieser ersten Erprobungsphase und genaue Anweisungen zum Vorgehen, in denen. die Lehrkräfte aufgefordert wurden, für drei bis fünf Kinder zuerst den SDQ für eine statusdiagnostische Einschätzung des Verhaltens der Schüler_innen durchzuführen. Dabei wurde darauf hingewiesen, dass es sich bei den ausgewählten Schüler_innen um solche handeln sollte, bei denen sich eine Diagnostik im Bereich des Verhaltens aufgrund auffälligen externalisierenden oder internalisierenden Verhaltens anbot. Außerdem wurde betont, dass der SDQ sich auf das Verhalten der Schüler_innen in den letzten Schulwochen bezog. Aus den Ergebnissen des SDQs sollten die Lehrkräfte anschließend die relevanten Verhaltensbereiche ableiten, deren Überprüfung mit Hilfe des DBRs ihnen sinnvoll erschien. Den Lehrkräften wurde dabei freigestellt, den SDQ selbst auszuwerten, ihn zur Auswertung an die Autorin der vorliegenden Arbeit zu schicken oder auf eine detaillierte Auswertung des SDQs zu verzichten und alle Bereiche im DBR zu bearbeiten.

Im Anschluss an die Durchführung des SDQs wurden die Lehrkräfte dazu angehalten, das Instrument zur Verhaltensverlaufsdiagnostik für alle mit dem SDQ beurteilten Kinder mindestens zweimal und maximal fünfmal durchzuführen. Betont wurde, dass die Beobachtungszeiträume vergleichbar gewählt werden sollten. Beispielhaft aufgeführt waren die wiederholte Beobachtung des Mathematikunterrichts, der ersten beiden Unterrichtsstunden oder des gesamten Schultags. Außerdem wurden die Lehrkräfte darauf hingewiesen, dass das Ausfüllen der DBRs möglichst zeitnah im Anschluss an die beobachtete Situation und unbedingt von ihnen persönlich erfolgen musste.

Im weiteren Verlauf wurden die Lehrkräfte über ein Interview im Anschluss an die Durchführung der beiden Ratings informiert, in dem sie Rückmeldungen zu beiden Instrumenten mit Fokus auf dem DBR geben sollten.

Stichprobenwahl I

Zur Untersuchung der Fragestellung wurden drei Lehrerinnen einer Förderschule mit dem Förderschwerpunkt Lernen als Raterinnen ausgewählt. Alle drei Lehrkräfte haben eine sonderpädagogische Ausbildung im Bereich Lernen. Lehrkraft 1 (L1) hat neben dem Förderschwerpunkt Lernen außerdem den Förderschwerpunkt Emotionale und Soziale Entwicklung studiert sowie Deutsch und Englisch und unterrichtet seit 28 Jahren an Förderschulen. Zum Zeitpunkt der Durchführung der ersten Erprobung war sie als Klassenlehrerin in einer jahrgangsübergreifenden 4./5. Klasse tätig. Bei Lehrkraft 2 (L2) handelt es sich ebenfalls um eine Sonderpädagogin mit den Schwerpunkten Lernen und Emotionale und Soziale Entwicklung; ihre Fächer sind Mathematik und Deutsch. Sie unterrichtet seit 18 Jahren an Förderschulen und führte die Erprobung in einer 5. Klasse durch. Außerdem arbeitete sie im Rahmen der Inklusion immer wieder parallel im gemeinsamen Unterricht an Regelschulen. Die dritte Lehrkraft (L3) hat 20 Jahre Berufserfahrung an Förderschulen. Ihr zweiter Förderschwerpunkt ist Geistige Entwicklung, ihre Fächer sind ebenfalls Mathematik und Deutsch. Die Erprobung wurde von ihr in einer sechsten Klasse durchgeführt.

Die Lehrerinnen wurden als Expertinnen ausgewählt, da sie umfangreiche Erfahrung im Umgang mit Schüler_innen im sonderpädagogischen Bereich aufweisen und zwei der drei Lehrkräfte außerdem im Bereich der Emotionalen und Sozialen Entwicklung ausgebildet sind. Sie kennen sich dementsprechend gut mit Schüler_innen mit Verhaltensauffälligkeiten aus und

sind geübt im Umgang mit der Beurteilung und Einschätzung von (schwierigen) Verhaltensweisen. Hinzu kommt, dass Sonderpädagog_innen im Allgemeinen aufgrund ihrer diagnostischen Tätigkeiten im Rahmen der Bedarfsermittlung an sonderpädagogischer Unterstützung bereits vielfältige Berührungspunkte zu Tests, Ratings, Beurteilungs- und Beobachtungsbögen haben. Entsprechend können sie im schulischen Bereich im Vergleich zu Lehrkräften an Regelschulen, die wenig bis keine Erfahrung mit Diagnose- und Rating-Instrumenten haben, als Expert_innen angesehen werden, da sie über den für Expert_inneninterviews erforderlichen Wissensvorsprung verfügen. Sie werden deshalb als Expert_innen angesprochen, weil sie vom Interviewenden in einem gewissen Forschungszusammenhang als solche eingeordnet werden und über eine Art von Wissen verfügen, welches zwar auch weitere Personen besitzen können, das aber nicht jedem im betroffenen Handlungsfeld zugänglich ist (Meuser & Nagel, 2009). Daher bietet es sich an, ebendiese sonderpädagogischen Fachkräfte mit der Erprobung eines in der Entwicklung befindlichen Instruments zur Verhaltensverlaufsdiagnostik zu beauftragen und im Anschluss zu befragen.

Die Raterinnen wurden dazu aufgefordert, drei bis fünf Kinder aus ihrer jeweiligen Klasse zu beobachten und zu beurteilen. Die Auswahl der Schüler_innen wurde den Lehrkräften dabei freigestellt bis auf die Vorgabe, für die Erprobung Kinder zu wählen, die im Vergleich zu ihren Klassenkamerad_innen ein eher auffälliges Verhalten zeigten. Im bestmöglichen Fall sollten Kinder mit externalisierenden und Kinder mit internalisierenden Schwierigkeiten im Bereich Verhalten beurteilt werden. Alle Lehrkräfte bestimmten jeweils drei Kinder für die Untersuchung. Von diesen waren fünf männlichen und vier weiblichen Geschlechts. Diese neun Schüler_innen befanden sich zum Zeitpunkt der Erhebung in den Klassenstufen vier bis sechs und waren zwischen 9;8 und 12;7 Jahre alt. Das Durchschnittsalter lag bei 10;64 Jahren. Alle Kinder hatten den vorrangig diagnostizierten Förderschwerpunkt Lernen, zwei außerdem den Förderschwerpunkt Sprache. Ein Kind hatte zusätzlich zum Bereich Lernen Förderbedarf im Bereich der Emotionalen und Sozialen Entwicklung.

Instruktion der Lehrkräfte

Die Instruktion der Lehrkräfte erfolgte, wie in Kapitel 4.2.1 beschrieben, größtenteils schriftlich und im Vorfeld. Die drei erhebenden Lehrerinnen erhielten außerdem auf den jeweiligen Rating-Bögen (SDQ und DBR) eine kurze Instruktion. Die Instruktion des DBR-Instruments wurde in den Interviews im Anschluss an die Erprobung ebenso wie das Rating-Instrument

selbst evaluiert. Die genauen Instruktionen finden sich im Anhang (siehe Anhang B). Die Lehrkräfte mussten folglich vorher an keinem Training teilnehmen. Ein solches Training hätte aufgrund der zeitlichen Kapazitäten der Lehrerinnen nicht realisiert werden können, war allerdings auch nicht erforderlich, da aufgrund der langjährigen Diensterfahrung (siehe *Stichprobenwahl I*) im sonderpädagogischen Bereich von mehr als grundlegenden Kenntnissen sowohl in Bezug auf Verhalten und Verhaltensauffälligkeiten als auch im Hinblick auf Ratings, Beurteilungen und Beobachtungen ausgegangen werden konnte. Bei Fragen konnten die Lehrkräfte sich zudem an die Autorin der vorliegenden Arbeit wenden.

Expert_inneninterviews

Im Anschluss an die Durchführung der Ratings zur status- und verlaufsdiagnostischen Erfassung des Schüler_innenverhaltens durch die Lehrkräfte fanden leitfadengestützte Expert_inneninterviews statt, in denen die Lehrkräfte Rückmeldungen zu den Instrumenten geben sollten. Der Fokus lag hierbei aufgrund des Forschungsschwerpunkts und der diesbezüglich formulierten Forschungsfragen auf dem DBR. Die Interviews fanden zeitnah nach der Beendigung der Erprobungsphase im Rahmen von Einzelbefragungen statt und wurden schriftlich aufgezeichnet. Die Interviewfragen waren chronologisch geordnet und in zwei Blöcke aufgeteilt. Im ersten Block wurden Fragen zum Zeitraum vor der Durchführung gestellt. Diese bezogen sich auf die angesprochene Information per Email, offen gebliebene Fragen, allgemeine Verständnisfragen und Unklarheiten. Im zweiten Block des Interviews wurde sowohl nach der Durchführung beider Rating-Instrumente als auch nach Kritikpunkten gefragt. In beiden Blöcken wurde den Lehrkräften Zeit zur Äußerung sonstiger, nicht in den Fragen beinhalteter Punkte gegeben, die sie in Bezug auf die Erprobung hatten. Eine Auflistung der Interviewfragen findet sich im Anhang (siehe Anhang H).

4.2.2 Allgemeine Ergebnisse der Expert_inneninterviews

Die Expert_inneninterviews ergaben neben Rückschlüssen auf die Anpassung der Items des DBR-Instruments zusätzliche, für das weitere Vorgehen und den Verlauf der vorliegenden Arbeit bedeutsame Informationen. Diese Ergebnisse sollen im Folgenden für jede der befragten Lehrkräfte kurz vorgestellt werden. Eine Zusammenfassung und Interpretation der itemspezifischen Ergebnisse erfolgt in Kapitel 4.2.3. Die Notizen aus den protokollgestützten Interviews sind in den Anhängen I, J und K zu finden.

Allgemeine Ergebnisse des Interviews mit L1

Block 1 – Vor der Durchführung:

L1 gab an, die Instruktionen und beide Rating-Instrumente mehrmals gelesen zu haben, um ein besseres Verständnis des Ablaufs der Erprobung und des Inhalts der Rating-Bögen zu bekommen. Außerdem erwähnte sie, dass es ihr anfänglich schwergefallen sei herauszufinden, welches Instrument mehrfach und welches Instrument einfach durchgeführt werden sollte.

Block 2 – Während/nach der Durchführung:

L1 füllte den SDQ für zwei Schüler und eine Schülerin aus. Die Entscheidung zur Beobachtung und Beurteilung dieser Kinder fiel auf Grundlage der in der Email angesprochenen Empfehlung, Schüler_innen mit auffälligem Verhalten zu beurteilen. L1 gab an, Schüler_innen ausgewählt zu haben, die derartiges Verhalten zeigten; außerdem habe sie drei in der Auffälligkeit ihres Verhaltens divergierende Kinder gewählt. So wählte sie ein Kind aus, welches im schulischen Kontext durch häufige (Arbeits-)Verweigerung auffällt, eines, welches an einer diagnostizierten Aufmerksamkeits-Defizit-Hyperaktivitäts-Störung leidet und eines, das sich durch eine familienbedingte psychische Störung auffällig verhält. L1 bewertete beim Ausfüllen des SDQs das Verhalten für den gesamten Zeitraum seit dem Kennenlernen der Schüler_innen (vier Monate). Sie füllte den SDQ in der Schule nach dem Unterricht aus, wobei sie pro Kind circa zehn Minuten Zeit benötigte. Negativ aufgefallen sind ihr die durch das Ankreuzen einmal negativ („nicht zutreffend“), einmal positiv („zutreffend“) zu bewertenden, in ihrer Valenz jedoch durchweg positiv formulierten Items. Ihr fiel es außerdem schwer, bei einem Kind, welches sich beispielsweise wenig rücksichtsvoll (Item 1, Version 1) und überaktiv und unruhig (Item 2, Version 1) verhielt, nicht die gleiche Ausprägung auf der Skalierung anzuwählen. Beide Items seien aus ihrer Sicht gleich zu bewerten, da sie sich negativ auf das Verhalten des Schülers auswirken. L1 gab dementsprechend an, sie habe sich erst mit dem SDQ vertraut machen müssen, bevor ein Ausfüllen überhaupt möglich gewesen sei.

Das Instrument zum DBR füllte L1 für die gleichen Schüler_innen aus wie den SDQ und bewertete dabei alle sechs Verhaltensbereiche anstatt, wie in der Instruktion angeboten, einen Teil auszulassen. Dies begründete sie damit, dass sie bis auf einige wenige Items (siehe Kapitel 4.2.3) alle für „sinnvoll“ halte. L1 füllte den DBR für zwei Schüler jeweils dreimal innerhalb von drei Tagen aus und für eine Schülerin viermal. Sie wählte als Beobachtungszeitraum für alle Beurteilungen mit Hilfe des DBRs den gesamten Schultag aus. Allerdings erklärte sie , an keinem Tag von der ersten bis zur letzten Unterrichtsstunde in der Klasse gewesen zu sein.

Die Auswahl des Beobachtungszeitraumes fiel ihr leicht. Aufgrund von kleinen Unterbrechungen des Schultages (durch Fachwechsel, Raumwechsel, Lehrer_innenwechsel), empfand sie es als effektiver und sinnvoller, erst am Ende des Schultages das Verhalten der Kinder zu bewerten, wenn sie sich ein umfassenderes Bild gemacht hatte, anstatt dies zwischendurch zu tun. Sie füllte das Instrument immer im direkten Anschluss an den Unterricht am Ende des Schultags aus. Für die Beurteilung der Kinder mit Hilfe des DBRs benötigte sie zwischen fünf und sieben Minuten.

Allgemeine Ergebnisse des Interviews mit L2

Block 1 – Vor der Durchführung:

L2 gab an, im Vorfeld der Durchführung alles mehrfach genau durchgelesen zu haben, um Unklarheiten zu beseitigen.

Block 2 – Während/nach der Durchführung:
L2 füllte den SDQ für zwei Schüler und eine Schülerin aus, die sie nach der Auffälligkeit ihres Verhaltens aussuchte. Sie gab an, den SDQ außerhalb der Schulzeit von zu Hause aus in circa fünf Minuten pro Kind ausgefüllt zu haben. Dabei habe sie die verschiedenen Verhaltensbereiche im Rückblick auf das letzte Schulhalbjahr bewertet. L2 kritisierte Item Nummer 22 aus dem SDQ („Stiehlt zu Hause, in der Schule oder anderswo.“). Sie gab, das Ausfüllen dieses Items aufgrund des Bezugs auf das häusliche Verhalten der Lernenden schwierig gefunden zu haben. Darüber hinaus ergab das Interview, dass L2 den SDQ nach der Durchführung nicht weiter ausgewertet hatte. Die Ergebnisse des SDQs bewertete die Lehrkraft eher neutral und sah ihn vor allem als Bestätigung für das ihr bereits vor der Anwendung des Rating-Instruments Bekannte, da sie ihr Hauptaugenmerk generell auf das Verhalten der Schüler_innen lege. Zum Abschluss der Befragung zum SDQ regte die Lehrkraft an, dem Beurteilungsbogen eine explizite Erklärung der Abkürzungen für die verschiedenen Verhaltensbereiche hinzuzufügen.

Das DBR-Instrument wurde von L2 für die drei auch mit dem SDQ beurteilten Kinder jeweils dreimal an drei aufeinanderfolgenden Tagen durchgeführt. L2 entschied sich dafür, alle Verhaltensbereiche auszufüllen und nicht auf Grundlage des SDQs einzelne, stärker problembehaftete Bereiche auszuwählen. Den Bereich der Verhaltensprobleme mit Gleichaltrigen beur-

teilte L2 für die drei Schüler_innen nur im Anschluss an einen der drei Beobachtungszeiträume, da sie in den anderen beiden Fällen keinen Bezug zwischen Beobachtungssituation und Verhaltensbereich erkennen konnte. L2 wählte als Beobachtungszeitraum immer zwei aufeinanderfolgende Unterrichtsstunden im Wochenplansystem und füllte die DBR-Bögen am Ende des Schultages im Lehrerzimmer aus. Dabei war es gleich, ob die Wochenplanstunden am Anfang oder am Ende des Tages lagen. Die Auswahl der Beobachtungszeiträume fiel ihr leicht. Sie benötigte für die Beurteilung mit dem DBR pro Kind circa fünf Minuten.

Allgemeine Ergebnisse des Interviews mit L3

Block 1 – Vor der Durchführung:
L3 gab an, weder in Bezug auf die Instruktionen noch auf den SDQ oder das Instrument zum DBR im Vorfeld Schwierigkeiten oder Verständnisprobleme gehabt zu haben.

Block 2 – Während/nach der Durchführung:
Der SDQ wurde von L3 für zwei Schülerinnen und einen Schüler ausgefüllt, die in Bezug auf ihre Verhaltensauffälligkeiten ausgesucht worden waren. Ein besonderes Augenmerk legte L3 dabei darauf, dass die Schwierigkeiten in möglichst unterschiedlichen Verhaltensbereichen auftraten. Den SDQ füllte sie zu Hause nach der Schule für alle drei zu beurteilenden Lernenden hintereinander aus. Sie benötigte dafür ungefähr drei Minuten pro Kind und gab an, das Ausfüllen sei schnell gegangen. Den SDQ empfand sie als ausreichend und alle Bereiche abdeckend; ihr habe nichts gefehlt. Wie L1 und L2 hat auch L3 keine Auswertung des SDQs vorgenommen. Sie gab an, sie habe sich nicht auf einzelne Bereiche beschränken wollen, um ein ganzheitliches, vollständiges Bild zu bekommen. Das Ergebnis des SDQs bewertete L3 als passend, da ihre Einschätzungen der beurteilten Schüler_innen bestätigt wurden.

Das DBR-Instrument füllte L3 für dieselben Schüler_innen aus. Um die oben bereits angesprochene Vollständigkeit zu wahren, füllte sie das DBR für alle Verhaltensbereiche aus. Insgesamt beurteilte die Lehrkraft jedes Kind nur einmal mit dem DBR, bearbeitete also insgesamt drei DBR-Bögen. Sie begründete ihre Entscheidung damit, dass sich das Verhalten ihrer Meinung nach nicht innerhalb weniger Tage ändere und es deshalb ausreiche, einen Bogen pro Kind auszufüllen. L3 beobachtete immer das Verhalten der Schüler_innen an einem gesamten Unterrichtstag und beurteilte es direkt im Anschluss an die letzte Stunde im Lehrerzimmer. Sie benötigte dafür pro Kind circa drei Minuten.

Des Weiteren gab L3 am Ende des Interviews an, dass sie das Ausfüllen des DBRs mehrmals pro Woche für wenig sinnvoll halte (wörtlich: „Arbeits-Beschaffungs-Maßnahme“ (siehe Interview mit L3, ANHANG)). Sie begründete dies damit, dass ihrer Meinung nach kaum Verhaltensveränderungen in kurzen Zeiträumen stattfinden und in drei bis fünf Tagen gar keine. Auch nach einem erklärenden Hinweis der Interviewleiterin, dass die Verhaltensverlaufsdiagnostik dazu konzipiert sei, Förderungen auf ihre Effektivität zu prüfen sowie eine evidenzbasierte Rückmeldung zu geben und diese Erprobung nur zur Evaluation des DBR-Instruments angesetzt sei, blieb die Lehrkraft bei ihrer negativen Einschätzung der Verhaltensverlaufsdiagnostik. Zudem kritisierte sie den Rahmen, in dem das DBR durchgeführt wurde. Ihrer Meinung nach handelte es sich bei dieser Form der teilnehmenden Beobachtung und der anschließenden Beurteilung um stark subjektive Methoden, die außerdem durch eine bereits im Vorhinein gefasste Meinung beeinflusst würden. Sie schlug eine genaue zeitliche Erfassung der auffälligen Verhaltensweisen als objektiveren Weg vor.

Auswertung der allgemeinen Ergebnisse der Expert_inneninterviews

Alle drei Lehrkräfte gaben an, den SDQ zwar vor der Bearbeitung des DBRs ausgefüllt, ihn aber nicht weitergehend ausgewertet zu haben und die Durchführung des DBRs nicht von den im SDQ als problematisch eingestuften Verhaltensweisen abhängig gemacht zu haben, da sie von der Beurteilung aller Verhaltensbereiche ein umfassenderes, vollständigeres Bild erwartet hätten. Diesen Rückmeldungen entsprechend wurde der SDQ als Instrument, welches vor dem DBR angewendet werden und eigentlich durch das Einschränken auf einige Verhaltensbereiche zur Ökonomisierung beitragen sollte, im weiteren Verlauf der Überarbeitung und Erprobung des DBRs herausgenommen und nicht mehr angewendet. Da die Lehrkräfte offenbar dazu bereit und daran interessiert waren, wenn möglich und zur Beobachtungssituation passend alle Verhaltensbereiche auszufüllen, erscheint der SDQ im Nachhinein als überflüssiger Mehraufwand.

Eine der drei Lehrkräfte kritisierte das Instrument zum DBR offen und bezweifelte sowohl seine allgemeine Sinnhaftigkeit als auch seine Ökonomie. Ihrer Meinung könne man innerhalb weniger Tage keine Veränderung des Verhaltens feststellen. Außerdem seien die Beurteilungen zu subjektiv. Da diese direkte Kritik zum einen nur bei einer der drei Lehrkräfte derartig konkret zur Sprache kam und auch an keine konkreten Bereiche des Instruments geknüpft,

sondern auf Verhaltensverlaufsdiagnostik an sich bezogen war, hatte diese Kritik keinen Einfluss auf den DBR-Bogen.

Alle Lehrerinnen füllten den SDQ und das Instrument zum DBR für drei Kinder aus. In den Instruktionen wurde empfohlen, das Rating für drei bis fünf Kinder durchzuführen. Dass keine der Lehrkräfte mehr als drei Schüler_innen mit dem DBR ratete, ist vermutlich auf Zeitmangel zurückzuführen.

L2 füllte den Bereich der Verhaltensprobleme mit Gleichaltrigen in einigen Fällen nicht aus, da dieser ihrer Aussage nach nicht zur beobachteten Situation passte. Dies sollte für eventuell nachfolgende Studien berücksichtigt werden, da die soziale Validität (siehe Kapitel 2.2.2) eine große Rolle für die Auswahl der Items spielt. So könnte den beurteilenden Lehrkräften ein Spielraum gelassen werden, in dem sie selbst entscheiden können, welche Verhaltensbereiche es sich in Bezug auf die beobachtete Situation zu bewerten anbietet.

Die Wahl zu vergleichender Beobachtungszeiträume fiel allen Lehrerinnen leicht. Dies lässt den Schluss zu, dass die Instruktionen diesbezüglich ausreichten. Auch der in den Instruktionen gegebene Hinweis, die DBR-Bögen möglichst zeitnah im Anschluss an die beobachtete Situation auszufüllen, wurde von den Lehrkräften in der Regel beachtet und umgesetzt. Bei L1, welche als einzige Lehrerin nicht den gesamten Schultag beziehungsweise ihre gesamte gemeinsame Zeit mit den Schüler_innen an einem Schultag bewertete, sondern den Beobachtungszeitraum durch zwei Schulstunden mit Wochenplanarbeit definiert hatte, bestand, je nach Lage der Wochenplanstunden, ein etwas größerer zeitlicher Abstand zwischen der beobachteten Situation und dem Ausfüllen der DBR-Bögen.

Bei Betrachtung der ausgefüllten SDQ-Bögen konnte festgestellt werden, dass alle drei Lehrkräfte, wie in den Instruktionen empfohlen, für die Beobachtung Kinder mit unterschiedlichem Verhalten gewählt haben. Die Auswahl von Schüler_innen mit divergierenden Verhaltensschwierigkeiten scheint den Lehrerinnen nicht schwer gefallen zu sein. Dies kann an der generellen Erfahrung sonderpädagogisch ausgebildeter Lehrkräfte im Umgang mit verhaltensauffälligen Kindern und an der spezifischen Ausbildung zweier Lehrkräfte im Bereich der Emotionalen und Sozialen Entwicklung liegen. Da das DBR-Instrument in der Regel nur für Kinder mit solchen Verhaltensauffälligkeiten genutzt werden soll und dementsprechend nur im Bedarfsfall zum Einsatz kommt, müssen Lehrkräfte, die eine solche Verhaltensverlaufsdiagnostik durchführen, nicht dazu in der Lage sein, eine Klasse für eine Diagnostik bereits im

Vorhinein in Kinder mit auffällig externalisierendem oder internalisierendem Verhalten einzuteilen, wie es in der Erprobung der Fall war. Sie füllen das DBR-Instrument nur für die Kinder aus, die ihrer Meinung nach einer Überprüfung und einer eventuell anschließenden Anpassung der Fördermaßnahmen bedürfen. Zusammenfassend ist es also von eher geringer Bedeutung, wie einfach oder schwer den Lehrkräften die Auswahl der Schüler_innen fiel, da dies nur für die Erprobung und eventuell anschließende Pilotierungsstudien relevant ist, nicht aber für die eigentliche Anwendung von DBRs in der Praxis.

Alles in allem empfanden die Lehrerinnen den Ablauf der Erprobung schlüssig und sinnvoll. Abgesehen von genereller Kritik an der Methode der Verhaltensverlaufsdiagnostik gab es größtenteils positive Rückmeldungen zu der Durchführung und Anwendung der beiden Rating-Instrumente. Die Instruktionen wurden deshalb nicht geändert, sondern für die zweite Erprobung übernommen. Auch die Skalierung wurde von keiner Lehrkraft negativ bewertet. Weder die Extremwerte „Nie“ und „Immer“ noch die siebenstufige Einteilung wurden kritisiert, was für eine Beibehaltung dieser Skala spricht.

4.2.3 Itemspezifische Ergebnisse der Expert_inneninterviews

Nachdem in Kapitel 4.2.2 die allgemeinen Ergebnisse der Expert_inneninterviews vorgestellt und ausgewertet wurden, werden in diesem Kapitel nun die Rückmeldungen präsentiert, welche die Lehrkräfte in den Interviews zu den Items des DBR-Bogens (Version 1) gegeben haben. In diesem Kapitel stehen ausschließlich die Äußerungen der Lehrkräfte in Bezug auf die DBR-Items im Vordergrund, da es bei der Beantwortung oben genannter Forschungsfragen eben nicht um die Weiterentwicklung des SDQs geht, sondern um die des DBRs.

Itemspezifische Ergebnisse des Interviews mit L1

Item 5: Durch die negative Formulierung des fünften Items fiel L1 dessen Beurteilung verhältnismäßig schwer. Sie gab an, sie habe einige Male über das Item nachdenken müssen, bevor ihr eine Beurteilung möglich gewesen sei. Deshalb schlug sie eine Umformulierung von „Hört nicht auf die Lehrkraft und verhält sich nicht regelkonform“ zu der positiven Fassung „Hört auf die Lehrkraft und verhält sich regelkonform“ vor.

Item 6: L1 regte im Interview an, diesem Item einen Zusatz anzufügen, in dem auf die mögliche Provokation anderer Kinder durch das Verhalten der beobachteten und beurteilten Schüler_innen hingewiesen wird, und diese beurteilbar zu machen.

Item 16: Dieses Item wurde von L1 kritisiert, da in vielen Unterrichtsstunden nicht gespielt würde und es deshalb nur sehr selten ausfüllbar sei. Bei der Beurteilung der drei Lernenden füllte sie das Item nur dann aus, wenn den Schüler_innen auch tatsächlich eine Spielmöglichkeit in der Unterrichtsstunde eingeräumt wurde.

Item 17: L1 gab an, sie finde das Item ungünstig gewählt, da es ihrer Meinung nach eher in den Bereich der Statusdiagnostik gehöre. Außerdem stufte sie den Faktor der „Beliebtheit" von Schüler_innen im Klassen- und Schulkontext als insgesamt schwierig ein, da sich ihres Erachtens insbesondere bei jüngeren Kindern die Beliebtheit sehr schnell und vor allen Dingen häufig ändere. Obendrein sei sie in vielen Fällen nicht nur vom Verhalten und Charakter der Lernenden abhängig, sondern häufig auch von materiellen Gegebenheiten.

Item 18: Auch in Bezug auf Item 18 gab L1 an, dass sie dieses eher ungeeignet für ein verlaufsdiagnostisches Rating finde und es eher im Bereich der Statusdiagnostik ansiedeln würde.

Weitere Ergebnisse: Neben der direkten Kritik oben genannter Items formulierte L1 den Vorschlag, weitere Bereiche in Form zusätzlicher Items in das DBR einzugliedern. Hier nannte sie den Fall, dass ein Kind in Folge einer Verweigerung den Klassenraum verlässt und nicht wieder betritt sowie direkte Streitigkeiten mit anderen Schüler_innen. Insgesamt betonte sie, dass sie die fehlende Möglichkeit, Verweigerung zu beurteilen, negativ sehe.

Itemspezifische Ergebnisse des Interviews mit L2

Items 2 und 3: Diese beiden Items gleichen einander laut Aussage von L2 zu sehr und erschienen ihr wie eine Wiederholung.

Weitere Ergebnisse: L2 gab an, einige Items bei der Beurteilung von Beobachtungszeiträumen mit Wochenplanarbeit als überflüssig einzuschätzen. Außerdem meldete L2 zurück, Probleme mit der teilweise positiven, teilweise negativen Formulierung einiger Items gehabt zu haben. Item 1 beispielsweise („Meldet sich im Unterricht") ist in Bezug auf das sich dahinter verbergende und zu beurteilende Verhalten positiv einzuschätzen. Eine Bewertung mit „7/Immer" wäre deshalb gut. Eine gleiche Bewertung von Item 4 („Verhält sich wütend und aufbrausend") mit „7/Immer" wäre dahingegen in Bezug auf das Verhalten negativ. L2 wünschte sich laut eigener Aussage, dass die Bewertung eines Items mit „7/Immer" tatsächlich die bestmögliche Ausprägung eines Verhaltens darstellte. L2 forderte außerdem eine „praktischere Formulierung" der Items. Hierfür konnte sie auf Nachfrage keine Beispiele nennen.

Itemspezifische Ergebnisse des Interviews mit L3

Item 15: Dieses Item wurde von L3 kritisiert, da im schulischen Kontext das Teilen an sich zwar des Öfteren beobachtet werden könne, die Komponente der Freiwilligkeit („gerne“) hingegen jedoch eher im Hintergrund stehe. Da das Teilen häufig von Lehrkräften und anderen pädagogischen Fachkräften gefordert würde, könne nicht klar beurteilt werden, ob das im DBR bewertete Kind tatsächlich gerne teile oder nur, um nicht negativ aufzufallen. Außerdem gab L3 zu bedenken, Teilen sei zwar prinzipiell gut, jedoch in Bezug auf manchmal mittellose Schüler_innen nicht immer durchweg positiv zu bewerten. Ein Kind könnte vielleicht nur teilen, um dazuzugehören, beliebt zu sein oder Aufmerksamkeit zu erregen, und eventuell Dinge teilen, die es eher behalten sollte. L3 schlug eine Umformulierung oder das Austauschen dieses Items vor.

Auswertung der itemspezifischen Ergebnisse der Expert_inneninterviews

Die itemspezifischen Ergebnisse aus den Expert_inneninterviews geben Rückschlüsse auf notwendige Anpassungen des DBR-Instruments. Insbesondere die Rückmeldungen von L1 waren in Bezug auf die von ihr angesprochenen Items sehr aufschlussreich. Die Erläuterungen zu den Veränderungen der Items auf Grundlage der Interviews finden sich in Kapitel 4.2.4. Insgesamt kritisierten die Lehrkräfte hauptsächlich Formulierungen, wobei sie in einzelnen Fällen Vorschläge für Zusätze innerhalb bereits existierender Items machten. Außerdem erwähnten die Lehrkräfte persönliche Einschätzungen bezüglich der Eignung einzelner Items in der Verhaltensverlaufsdiagnostik und verwiesen auf einige eher statusdiagnostische Komponenten. Auch die Beobachtbarkeit der Items wurde von den Lehrerinnen evaluiert und bei manchen Items als ungünstig bewertet. Ein weiterer Kritikpunkt bezog sich auf positive und negative Formulierungen sowie auf die gesamte Bewertung der Items auf der Skala. Hier stach insbesondere die Formulierung von Item 5 heraus. Wie in Kapitel 2.4 in Bezug auf die Valenz der Zielformulierung beschrieben, fällt es Rater_innen laut Huber und Rietz (2015) leichter, die Anwesenheit störender Verhaltensweisen zu beurteilen als deren Abwesenheit. Dem widerspricht die Aussage von L1. Ihrer Meinung nach ist es einfacher, die Anwesenheit regelkonformen, den Anweisungen der Lehrkraft folgenden Verhaltens zu bewerten als die Anwesenheit störenden Verhaltens. Dies mag einerseits an der Formulierung liegen, welche in Item 5 als einzige im Bereich der Verhaltensprobleme negativ gewählt wurde, andererseits an der

in manchen Fällen notwendigen Nutzung des linken Skalierungsbereichs, welcher am Extrempunkt mit „Nie“ beschrieben ist und infolgedessen durch die negative Formulierung des Items eine doppelte Verneinung nach sich ziehen würde. Bei einer Umformulierung des Items wäre dieses zwar, wie die anderen beiden Items aus dem Bereich der Verhaltensprobleme, ebenfalls positiv formuliert, würde jedoch als einziges zudem ein positiv zu bewertendes Verhalten messen. Um den Lehrkräften jedoch eine lange Auseinandersetzung mit der Formulierung zu ersparen und außerdem dem Bestreben nachzukommen, ein ökonomisches Instrument zu entwickeln, erscheint eine Umformulierung wie die in Kapitel 4.2.4 beschriebene angebracht.

Die Items 17 und 18 wurden in der Folge der Interviews ebenfalls abgeändert (siehe Kapitel 4.2.4), da sie von L1 als zu starr und wenig geeignet im Bereich der Verlaufsdiagnostik bewertet wurden. Weder der Beliebtheitsgrad noch die Hinwendung zu Erwachsenen und entsprechende Abwendung von gleichaltrigen Schüler_innen sind in der in Version 1 des DBR-Instruments versprachlichten Form passend. Eine sichtbare Veränderung auf einer Verhaltenskurve, die aus den Ergebnissen mehrerer innerhalb eines kurzen Zeitraums ausgefüllter DBR-Bögen gebildet würde, erscheint in diesen beiden Bereichen eher unwahrscheinlich. Außerdem stellt sich die Frage, ob die Beliebtheit eines Kindes (Item 17) gefördert werden kann oder sollte und ob eine Überprüfung dieses Items generell sinnvoll ist. Auch das eher global formulierte Auskommen mit Erwachsenen (Item 18) bedarf, um es ausfüllen zu können, einiger zusätzlicher Überlegungen und kann die beurteilenden Lehrkräfte zu einem Einbezug vergangener Schultage, Wochen oder sogar Monate verleiten. Dementsprechend wäre, wie in Kapitel 2.2.2 in Bezug auf Verhaltensverlaufsdiagnostik angegeben, eine etwas spezifischere Formulierung sinnvoller, mit der eine niedrig inferente Herangehensweise ermöglicht würde.

Festgehalten werden sollte, dass der Bereich der Verhaltensprobleme mit Gleichaltrigen eher kritisch zu betrachten ist. L1 bewertet zwei der drei in diesem Bereich enthaltenen Items negativ und L2 füllt diesen Bereich in vielen Fällen gar nicht aus, da sie solches Verhalten in den von ihr beobachteten Situationen nicht beobachten könne. Die negative Bewertung der Items 17 und 18 wurde oben bereits thematisiert. Die laut L2 nicht mögliche Beurteilung der drei Items 16-18 wiederum erscheint im Hinblick auf die nicht besonders ausgefallen gewählte Beobachtungssituation im Rahmen der Wochenplanarbeit, wie sie an vielen (Förder-)Schulen in den Schulalltag integriert wird, kritisch, denn ebendiese Anwendung des DBRs in alltäglichen Situationen im Schulkontext stellt den Regelfall dar. Die Einordnung der Items aus dem

Bereich der Verhaltensprobleme mit Gleichaltrigen weist dementsprechend auch auf eine mangelnde soziale Validität der Items hin, die in einer Überarbeitung des Ratings zu berücksichtigen ist. Eine Eliminierung des Verhaltensbereichs aus dem Rating erscheint allerdings wenig sinnvoll, da dieser insbesondere für Schüler_innen mit Verhaltensschwierigkeiten von Bedeutung ist und es im schulischen Kontext ständig zu Konfrontationen mit Gleichaltrigen kommen kann.

In Bezug auf Forschungsfrage 1) und 2) kann zusammenfassend festgestellt werden, dass einige Bereiche des DBR-Instruments von den Lehrkräften angenommen wurden, andere, insbesondere in Bezug auf die Formulierung der Items, noch einer Anpassung bedürfen. Insgesamt waren die Rückmeldungen der Lehrkräfte vor allen Dingen in Bezug auf die Durchführbarkeit und Praktikabilität des DBRs vorerst zufriedenstellend.

4.2.4 Anpassung des Direct Behavior Ratings auf Grundlage der Expert_inneninterviews

Auf Grundlage der Interviewergebnisse wurden die Items 3, 5, 6, 15, 16, 17 und 18 aus Version 1 des DBRs angepasst. Im Folgenden wird die Änderung der Items genau beschrieben und die neue Version des DBRs vorgestellt.

Item 3 wurde in den Expert_inneninterviews als redundant kritisiert. Die Ähnlichkeit zu Item 2 war insbesondere L2 zu hoch. Deshalb wurde Item 3 von „Bleibt während des Unterrichts ruhig am Platz sitzen, wenn dies erforderlich ist“ komplett inhaltlich umformuliert und zu „Redet im Unterricht oft dazwischen“ geändert. Dieses Item wurde aus dem Bereich „Schulbezogenes Verhalten“ des SDQs übernommen. Die positive Formulierung bleibt bestehen. Auch wird weiterhin die Anwesenheit und nicht die Abwesenheit eines Verhaltens beurteilt, was Rater_innen laut Huber und Rietz (2015) leichter fällt (siehe Kapitel 2.4) und sich entsprechend günstig auf Inferenz und Ökonomie auswirkt. Während die Formulierung positiv bleibt, verändert sich die Valenz des Items in Richtung der Beobachtung eines negativ zu bewertenden Verhaltens. Während die anderen beiden Items des Bereichs „Schulbezogenes Verhalten“ erwünschte Verhaltensweisen messen, fokussiert Item 3 nun eine Form störenden Verhaltens.

Item 5 wurde von der negativen Formulierung „Hört nicht“ und „Verhält sich nicht“ in ein positiv formuliertes Item überführt. Um eine vereinfachte Bewertung des Items zu gewährleisten und einer Verwirrung durch eine mögliche doppelte Verneinung („nie“ und „nicht“) zu

vermeiden, wird hier von der oben als günstiger beschriebenen Erfassung der Anwesenheit eines störenden Verhaltens zu der eines erwünschten Verhaltens gewechselt. Durch diese Änderung wird Item 5 als einziges der im Bereich der Verhaltensprobleme verankerten Items 4, 5 und 6 auf die Bewertung eines gewünschten Verhaltens hin ausgerichtet. Item 4 und Item 6 erfassen dagegen unerwünschte, negative Verhaltensweisen. Des Weiteren überprüft das umformulierte Item nun keine externalisierende Verhaltensweise im Rahmen des Bereichs der Verhaltensprobleme mehr, sondern das Gegenteil – ein wünschenswertes Verhalten. Ein solcher Wechsel ist als problematisch zu beurteilen und bedarf weiterer Überprüfung und Anpassung. Das geänderte Item 5 lautet nach der Änderung „Hört auf die Lehrkraft und verhält sich regelkonform“.

Item 6 wurde von Version 1 mit einem Zusatz in Version 2 übernommen. Da den Lehrkräften der Aspekt der Provokation anderer Schüler_innen durch das Verhalten des beurteilten Schülers oder der beurteilten Schülerin fehlte, wurde es um den Satz „Provoziert durch eigenes Verhalten die anderen Kinder“ erweitert. Im Zuge dieser Erweiterung des Items wurde der Teil „Schikaniert seine MitschülerInnen“ der Länge und Übersichtlichkeit halber herausgenommen, sodass Item 6 in Version 2 des DBRs „Streitet sich mit anderen Kindern/provoziert durch eigenes Verhalten seine Mitschüler_innen“ lautet.

Die Veränderung von Item 15 basiert auf verschiedenen Überlegungen. Zum einen wurde von den Lehrkräften das Wort „gerne“ als schwer beobacht- und messbar kritisiert. Es stellte sich die Frage, ob ein Kind tatsächlich mit Freude und absolut freiwillig teilt oder eher, weil es von Lehrkräften und den schulinternen Normen und Werten dazu gedrängt und vom Schulkontext gezwungen wird. Das Teilen wäre dann kein Indiz für prosoziales Verhalten mehr, sondern eher für Anpassungsfähigkeit oder ähnliches. Außerdem wurde von einer der Lehrkräfte der Gedanke angeregt, dass Teilen an sich nicht immer ausschließlich positiv sei (siehe Kapitel 4.2.3). Das Item wurde deshalb komplett geändert, wobei der Fokus darauf lag, den Aspekt der Freiwilligkeit beizubehalten, welcher für das prosoziale Verhalten ausschlaggebend ist. Das neue Item („Bietet anderen Unterstützung an“) enthält diesen Aspekt. Wörter wie das in der Vorgängerversion genutzte „gerne“ wurden aus den genannten Gründen nicht eingefügt.

Item 16 wurde um den Aspekt des Arbeitens erweitert und lautet nun „Spielt/arbeitet meist alleine“, sodass nun nicht nur bewertet und beurteilt werden kann, ob das Kind eher mit Er-

wachsenen als mit anderen Kindern spielt, sondern auch, ob es lieber mit Erwachsenen arbeitet. Dies erschien sinnvoll, weil nicht jede Unterrichtsstunde spielerische Phasen enthält. Das Item wird so im Rahmen des Kriteriums der sozialen Validität besser beobachtbar und ist aufgrund der nun globaler gewählten Formulierung einfacher auszufüllen.

Die Änderung von Item 17 von „Ist bei anderen Kindern beliebt“ in „Wird von anderen Kindern gehänselt oder geärgert“ soll vom eher statusdiagnostischen Fokus des Items auf die Beliebtheit eines Kindes hin zu einem verlaufsdiagnostisch überprüfbaren Item führen. Das Item wurde deshalb weiter operationalisiert und praktischer formuliert. Eine Abwendung vom direkten Bezug zur Beliebtheit legte sich zudem nahe, da diese insbesondere im schulischen Kontext häufig von Faktoren wie dem sozialen Status, materiellen Aspekten und der jeweiligen Situation abhängig ist. Die Ausrichtung des Items bleibt letztlich dennoch erhalten, da Kinder mit einem hohen Beliebtheitsgrad seltener gehänselt oder geärgert werden als eher unbeliebte Kinder, wobei sich die Valenz des Items ändert.

Item 18 wurde von der globalen Formulierung „Kommt besser mit Erwachsenen als mit anderen Kindern aus“ zur etwas engeren, spezifischeren Formulierung abgewandelt „Arbeitet/spielt lieber mit Erwachsenen als mit anderen Kindern“. Diese Umformulierung hat zur Folge, dass aufgrund einer genaueren und praxisbezogeneren Operationalisierung ein besserer Situationsbezug hergestellt sowie eine höhere soziale Validität geschaffen werden können und das Item infolgedessen einfacher zu beobachten ist (vgl. hierzu Kapitel 2.2.2). Eine solche genauere Operationalisierung bewirkt zudem, dass die beobachteten Verhaltensweisen für alle am Förderprozess beteiligten Personen nachvollziehbar sind.

Die vollständige zweite überarbeitete Version des DBR-Instruments findet sich im Anhang (siehe Anhang C). Einige Items wurden weiter operationalisiert, zwei wurden unter Berücksichtigung des übergeordneten Verhaltensbereichs komplett ausgetauscht. Alle Veränderungen basieren auf den Rückmeldungen der befragten Expert_innen und der Erprobung des DBRs in der Praxis. Die oben auf dem DBR-Bogen befindliche Kurzversion der Instruktion wurde, wie in Kapitel 4.2.2 beschrieben, nicht verändert. Im Rahmen einer Untersuchung zur Interrater-Reliabilität wurden die Items der zweiten Version einer erneuten Überprüfung unterzogen, wie im Folgenden dargelegt werden wird.

4.3 Revision und Erprobung des Direct Behavior Ratings mit Fokus auf seine Interrater-Reliabilität

Nach der Erprobung der ersten Version des DBRs und der Überarbeitung und Anpassung des Instruments wurde die zweite Version zur weiteren Überarbeitung erneut in der Praxis getestet und im Hinblick auf ihre Interrater-Reliabilität untersucht. Ziel dieser Revision war es, die Items noch einmal anzupassen und so zu ändern, dass die Praktikabilität sowohl für die Durchführung als auch für die spätere Auswertung erhöht würde. Das Ziel der Erprobung des DBRs im Hinblick auf seine Interrater-Reliabilität war einerseits, eine Rückmeldung über die Formulierung und bereits durchgeführte Operationalisierung der Items zu erhalten. Andererseits galt es zu erfahren, wie reliabel das Instrument ist, wenn es von verschiedenen Rater_innen in der Praxis angewendet wird. Der Fokus lag also erneut auf der Beantwortung von Forschungsfrage 2, welche sich mit den notwendigen Anpassungen und Veränderungen im Hinblick auf die Anwendbarkeit des DBRs befasst und mit deren Beantwortung bereits in Kapitel 4.2 begonnen wurde. Außerdem wurde Forschungsfrage 3 a) in den Fokus gerückt. Diese bezieht sich auf die Reliabilität des Instruments im schulischen Kontext und hier speziell auf die Interraterübereinstimmung bei der Anwendung des Instruments durch zwei oder drei Rater_innen. Abschließend sollte eine erneute Anpassung des Instruments auf Grundlage der qualitativen Rückmeldungen während der zweiten Erprobung und in deren Anschlussstattfinden.

4.3.1 Beschreibung der Durchführung

In diesem Kapitel werden die Stichprobenwahl und der Ablauf der Ratings dargestellt. Auf der Basis dieser zweiten Erprobung wird in Kapitel 4.3.2 die Interrater-Reliabilität berechnet und in Kapitel 4.3.3 eine erneute Anpassung der Items vorgenommen. Die Stichprobenwahl bezieht sich auf die Beobachter_innen, die das Rating mit der auf Grundlage der Expert_inneninterviews entstandenen zweiten Version des DBRs durchgeführt haben. Im Abschnitt zum Ablauf des Ratings wird skizziert, in welcher Form die Erhebung der Interrater-Reliabilität stattgefunden hat.

Stichprobenwahl II

Die Berechnungen zur Interrater-Reliabilität basieren auf den Beurteilungen dreier Beobachter_innen beziehungsweise Rater_innen. Bei Beobachterin 1 (B1) und Beobachterin 2 (B2) handelt es sich um geschulte Raterinnen, die sich bereits im Vorfeld sowohl mit dem Instru-

ment an sich als auch mit den Verhaltensbereichen auseinandergesetzt haben und dementsprechend mit den Items vertraut sind. Die dritte beobachtende Rolle wurde nacheinander von zwei verschiedenen Lehrkräften eingenommen, die im Folgenden unter B3 zusammengefasst werden. Ermöglicht wird dies dadurch, dass der Hauptfokus auf dem Faktor „Lehrkraft“ liegt und dieser in Bezug auf die Beurteilungsübereinstimmung mit den beiden geschulten Rater_innen überprüft werden soll. Es handelt sich bei den beiden Lehrkräften außerdem um vergleichbare Beobachter_innen, da beide die Förderschwerpunkte Lernen und Geistige Entwicklung studiert haben und die jeweils beurteilten Schüler_innen zum Zeitpunkt der Erhebung seit ungefähr anderthalb Jahren als Klassenleitungen unterrichteten. Hinzukommt, dass die Stichprobe für den Vergleich von B1 oder B2 mit B3 aufgrund von Zeitmangel der Lehrkräfte verhältnismäßig klein ausgefallen ist. Wären die beiden unter B3 zusammengefassten Lehrkräfte auf zwei einzelnen Ebenen erfasst worden, wäre die Stichprobe noch weitaus kleiner gewesen.

Beide Lehrkräfte B3 können als ungeschulte Rater_innen bezeichnet werden, da sie sich im Vorfeld weder tiefergehend mit der Verhaltensverlaufsdiagnostik beschäftigt haben noch in Bezug auf das Rating-Instrument oder die Verhaltensbereiche geschult wurden. Trotzdem kann davon ausgegangen werden, dass beide Lehrkräfte aufgrund ihrer Tätigkeit an einer Förderschule mit dem Förderschwerpunkt Lernen und dementsprechend langjähriger Erfahrung bereits sehr vertraut im Umgang mit Kindern und Jugendlichen mit auffälligen Verhaltensweisen in verschiedenen Bereichen sind.

Insgesamt wurden 16 Schüler_innen beurteilt, von denen vier weiblich und zwölf männlich sind. Drei der Schüler_innen waren zum Zeitpunkt der Beobachtung in einer jahrgangsübergreifenden Klasse 4/5, zehn der Schüler_innen in einer 5. Klasse und wiederum drei in einer 6. Klasse. Die Schüler_innen waren zwischen 9;8 und 11;9 Jahre alt. Die Auswahl der Schüler_innen wurde auf unterschiedliche Weise getroffen. In den Klassen 4/5 und 6 wurden die Schüler_innen in einem Vorgespräch mit der Lehrkraft ausgesucht. Hierbei lag der Hauptfokus darauf, Kinder mit Auffälligkeiten in unterschiedlichen Bereichen auszuwählen. Die Lehrkräfte wurden dazu angehalten Schüler_innen zu nennen, welche in externalisierenden oder internalisierenden Bereichen Auffälligkeiten zeigten. Alternativ konnten auch Schüler_innen gewählt werden, die entweder schwierig einschätzbare oder im Mittelfeld zwischen externalisierend und internalisierend befindliche Verhaltensweisen aufwiesen. Die Auswahl der zehn

Schüler_innen, welche in der 5. Klasse mit Hilfe des DBRs beurteilt wurden, war nicht im Vorhinein mit der Lehrkraft abgesprochen. Dies lag daran, dass vor dem Beginn der Unterrichtseinheit, welche direkt auf die große Pause folgte (siehe *Ablauf der Ratings*), aufgrund organisatorischer Gegebenheiten kein ausreichend großes Zeitfenster zur Rücksprache mit der Lehrkraft gegeben war. Von den zwölf an diesem Tag anwesenden Schüler_innen wurden deshalb zehn ausgewählt. Eine derart große Stichprobe sollte die Wahrscheinlichkeit erhöhen, ohne vorherige Absprache mit der Lehrkraft Schüler_innen mit unterschiedlichen Verhaltensweisen beobachten und beurteilen zu können.

Zu erwähnen ist, dass B2 die im Rahmen dieser zweiten Erprobung zur Untersuchung der Interrater-Reliabilität beobachteten und beurteilten Schüler_innen aus dem Praxissemester kannte, während B1 die Schüler_innen zum Zeitpunkt der Durchführung und Anwendung des DBR-Instruments zum ersten Mal sah.

Ablauf der Ratings

Alle Ratings fanden im Rahmen eines Schultages an einer Förderschule mit dem Förderschwerpunkt Lernen statt. Ausgewählt wurden, wie bereits erläutert, drei verschiedene Klassen, aus denen zum Teil vorher abgesprochene Kinder, zum Teil größere Stichproben beobachtet und beurteilt wurden. Sämtliche Ratings gingen über eine Unterrichtseinheit von 45 Minuten. Im Vorhinein waren drei Lehrkräfte angeschrieben und um die Möglichkeit gebeten worden, eine Verhaltensverlaufsdiagnostik mittels des DBR-Instruments durchzuführen. Zwei der Lehrkräfte gaben positive Rückmeldungen und stimmten einer solchen Beurteilung ausgesuchter Schüler_innen zu. Die dritte angeschriebene Lehrkraft sagte ab, da sie befürchtete, mit den zusätzlichen zwei Personen zu viele Erwachsene im Klassenraum zu haben und die Schüler_innen dadurch zu verunsichern. Da diese Absage erst relativ kurzfristig erfolgte, musste am Tag der Durchführung eine weitere Lehrkraft ausgewählt werden, die mit einer Beobachtung und Beurteilung einiger ihrer Schüler_innen einverstanden war. Diese konnte bei Vorgesprächen im Lehrerzimmer zeitnah gefunden werden.

Das erste von insgesamt vier Ratings fand in der zweiten Unterrichtsstunde in der Klasse 4/5 statt (9.00 – 9.45 Uhr). Die Lehrkraft teilte B1 und B2 im Vorhinein mit, welche Schüler_innen es ihrer Meinung nach aufgrund verschieden ausgeprägter Verhaltensweisen zu beobachten und beurteilen lohnte. Ausgewählt wurden drei Jungen, von denen laut Lehrkraft zwei eher externalisierende Verhaltensweisen zeigten, während einer im Bereich des internalisierenden

Verhaltens auffällig erschien. Die Klasse wurde während der Durchführung des DBRs von ihrer Klassenlehrerin im Fach Deutsch unterrichtet. Es handelte sich um eine relativ offene Form des Unterrichtens mit viel Differenzierung. Eine direkte Instruktion gab es nur am Anfang der Unterrichtseinheit. Einige Bereiche des Rating-Instruments konnten deshalb nicht oder nur teilweise bewertet werden. Hierzu zählen die Items 15 und 18 (Version 2), welche von keiner der Rater_innen bewertet wurden, sowie Item 14, das in einem Fall nur von B2 ausgefüllt wurde und in einem Fall gar nicht. B1 und B2 beobachteten die Schüler_innen während der Unterrichtseinheit und bewerteten sie direkt anschließend mit Hilfe der Rating-Skala. Eine Absprache bezüglich des Verhaltens und seiner Einschätzung fand vor dem Ausfüllen der DBR-Bögen nicht statt. Die unterrichtende Lehrkraft konnte wegen eines sich an die Stunde anschließenden Termins nicht raten.

Das nächste Rating wurde nach der ersten großen Pause in der dritten und vierten Stunde (10.15 – 11.45 Uhr) in der 5. Klasse vorgenommen. Hier wurden zwei Unterrichtseinheiten beobachtet und im Anschluss an jede Einheit jeweils fünf Kinder beurteilt. Vor der Beurteilung konnte keine Absprache zur Auswahl der Schüler_innen mit der Lehrkraft stattfinden, da diese bis kurz vor Stundenbeginn in die Lösung eines in der Pause aufgetretenen Konflikts eingebunden war. Von den zwölf anwesenden Schüler_innen wurden deshalb pro Unterrichtseinheit fünf mit Hilfe des DBRs beurteilt (siehe *Stichprobenwahl II*). Von diesen waren drei weiblich und sieben männlich. Bei den beiden beobachteten Unterrichtseinheiten handelte es sich um zweimal 45 Minuten Kunstunterricht, der durch die Klassenlehrerin erteilt wurde. Ähnlich wie in der ersten beobachteten Unterrichtseinheit liefen auch diese beiden Einheiten relativ offen ab. Eine Instruktion der Schüler_innen fand am Anfang der Stunde statt. Dies hatte zur Folge, dass auch im Anschluss an diese Beobachtungssituation mehrere Items für einige Schüler_innen nicht ausgefüllt werden konnten. Dies gilt wieder für die Items 15 und 18, zudem für Item 3, das in einem Fall ebenfalls von keinem Rater und keiner Raterin beurteilt wurde. Item 1 wurde in zwei Fällen, Item 15 in einem, Item 14 in einem Fall jeweils nur von B2 und Item 18 in je einem Fall nur von B1 beziehungsweise B3 ausgefüllt. Die Beurteilung der beobachteten Schüler_innen fand direkt im Anschluss an die jeweilige Unterrichtseinheit statt. Die drei von der Lehrkraft ausgewählten und von ihr im Hinblick auf das Verhalten in der zweiten Unterrichtseinheit beurteilten Kinder zählten zu den fünf von B1 und B2 in der zweiten Unterrichtseinheit beobachteten Schüler_innen.

Die vierte Beobachtung war im Erdkundeunterricht in der fünften Stunde (12.00 – 12.45 Uhr) einer 6. Klasse verortet. Deren Klassenlehrer, der auch die Beobachtung durchführte, hatte sich am Morgen spontan im Lehrerzimmer für die Durchführung des Ratings gemeldet. Im Vorfeld waren auf Nachfrage ein Schüler mit eher externalisierendem und ein Schüler mit eher internalisierendem Verhalten ausgewählt worden. Eine ebenfalls genannte Schülerin schwankt laut den Angaben des Lehrers je nach Verfassung zwischen den beiden Verhaltensrichtungen. Die Unterrichtsform kann als frontal beschrieben werden: Im gesamten Beobachtungszeitraum führte die Lehrkraft ein Unterrichtsgespräch mit den Schüler_innen. Das Ausfüllen der DBR-Bögen fand im direkten Anschluss an den Beobachtungszeitraum statt. Schwierigkeiten ergaben sich hierbei im Gegensatz zu B1 und B2 für B3, der zunächst wartete, bis die Schüler_innen den Klassenraum verlassen hatten, und mehrfach durch zwei Schüler unterbrochen wurde, welche aus Neugier auch nach Unterrichtsschluss in die Klasse kamen. Für Item 18 konnte erneut keinerlei Beurteilung erfolgen. Die Items 15 und 16 wurden in jeweils nur einem Fall und nur von B3 bewertet, die Items 11, 12 und 14 in eins bis drei Fällen von nur zwei Rater_innen ausgefüllt.

Während sich B1 und B2 ganz auf die Beobachtung und Beurteilung der Schüler_innen konzentrieren konnten, musste B3 parallel dazu den Unterricht gestalten und organisieren und konnte sich dementsprechend nicht ausschließlich der Beobachtung widmen.

4.3.2 Auswertung der Überprüfung der Interrater-Reliabilität

Zur Überprüfung der Interrater-Reliabilität der zweiten Version des DBRs beobachteten und beurteilten die oben genannten Beobachter_innen beziehungsweise Rater_innen das Verhalten von Schüler_innen im unterrichtlichen Kontext. Der Stichprobenumfang und die Anzahl *N* übereinstimmender Ratings zwischen den einzelnen Rater_innen werden in der untenstehenden Matrix veranschaulicht. B1 beurteilte insgesamt 256 Items, während B2 258 Items bewertete. B3 füllte mit einer Bewertung von 97 Items am wenigsten aus. Die meisten korrespondierenden Messungen wurden von B1 und B2 durchgeführt. Hierbei handelt es sich um 254 von 258 durchgeführten Ratings. Die Beurteilungen von B1 und B3 korrespondierten in 87 Fällen, die Beurteilungen von B3 und B2 in 89 Fällen.

$$N(\mathrm{B1}, \mathrm{B2}, \mathrm{B3}) = \begin{pmatrix} 256 & 254 & 87 \\ 254 & 258 & 89 \\ 87 & 89 & 97 \end{pmatrix}$$

Bei der zweiten Erprobung wurden 16 Schüler_innen durch B1 und B2 bewertet. Jede Schülerin und jeder Schüler konnte mit 18 Items beurteilt werden. Insgesamt gab es für B1 und B2 also die Möglichkeit, 288 Items zu beurteilen. B3 führte das DBR-Instrument für 6 Schüler_innen durch und hatte entsprechend eine mögliche Höchstzahl zu bewertender Items von 108. Es lässt sich demnach, wie im Abschnitt zur Stichprobenwahl (*Stichprobenwahl II*) bereits angedeutet, feststellen, dass weniger Items geratet wurden, als die Stichprobe ermöglichte. Daraus lässt sich folgern, dass es nicht in jedem Unterrichtskontext möglich ist, alle Items zu bewerten (siehe hierzu auch Kapitel 4.2.3). Viele Items konnten immer ausgefüllt werden, während einige wiederholt nicht bewertet werden konnten. Zu Letzteren zählten insbesondere die Items 15, 16 und 18 aus Version 2 des DBRs. Auch die Items 1, 3, 11, 12 und 14 wurden vereinzelt nicht beurteilt. Dies lässt auf eine geringere Allgemeingültigkeit der Items beziehungsweise ein selteneres Auftreten im unterrichtlichen Kontext schließen.

Nach der Beschreibung des Stichprobenumfangs und ersten Schritten in Richtung einer Auswertung und Begründung nicht bewerteter Items soll nun die Überprüfung des Rangkorrelationskoeffizienten nach Spearman erfolgen. Die Spearman-Rangkorrelation ist ein Maß zur Beschreibung der Zusammenhänge zwischen ordinal-skalierten Daten, wie sie hier vorliegen (Bühner, 2011). Im Folgenden wird dementsprechend der Korrelationskoeffizient nach Spearman für den Zusammenhang zwischen zwei beziehungsweise drei Variablen benannt, um die Übereinstimmung zwischen mehreren Beobachter_innen, das heißt die Interrater-Reliabilität anzugeben. Die Variablen sind in diesem Fall die Rater_innen beziehungsweise Beobachter_innen B1, B2 und B3. Mit Hilfe der Berechnung der Interrater-Reliabilität soll, wie in untenstehender Matrix dargestellt, herausgefunden werden, wie stark die jeweiligen Ratings der verschiedenen Beobachter_innen miteinander zusammenhängen.

$$\rho(\mathrm{B1}, \mathrm{B2}, \mathrm{B3}) = \begin{pmatrix} 1.0 & .837 & .724 \\ .837 & 1.0 & .853 \\ .724 & .853 & 1.0 \end{pmatrix}$$

Allgemein kann festgehalten werden, dass alle Werte größer als 0 und nahe an 1 verortet sind, was bedeutet, dass insgesamt eine positive Korrelation zwischen allen Rater_innen nachgewiesen werden konnte. Diese ist auf dem Niveau von .01 (zweiseitig) signifikant. Bei näherer Betrachtung wird deutlich, dass der wichtigste Wert durch die hohe Korrelation zwischen den beiden geschulten Rater_innen B1 und B2 beschrieben wird. Hier liegt der Wert bei ρ = .837,

was im Rahmen der Rangkorrelation nach Spearman für eine sehr hohe bis perfekte Korrelation spricht. Auch zwischen B1 und B3 gibt es mit einem Wert von $\rho = .724$ einen deutlichen Zusammenhang. Die Übereinstimmung zwischen B2 und B3 ist zwar mit einem Wert von $\rho = .853$ sogar noch höher als jene zwischen B1 und B2 und damit ebenfalls als sehr hoch bis perfekt einzuordnen, aufgrund der geringen Stichprobengröße jedoch ähnlich wie die Korrelation zwischen B1 und B3 in ihrer Aussagekraft beschränkter.

Insgesamt kann aufgrund der durchweg hohen Korrelationen (zwischen $\rho = .724$ und $\rho = .853$) insbesondere für die größte übereinstimmende Stichprobe zwischen B1 und B2 von einer ausgeprägten Interrater-Reliabilität gesprochen werden. Die durchgeführte Untersuchung von Forschungsfrage 3 a) ergibt also, dass die Interrater-Reliabilität des überprüften Rating-Instruments gegeben ist.

4.3.3 Finale Anpassung des Direct Behavior Ratings auf Grundlage qualitativer Rückmeldungen

Auf Basis von Rückmeldungen und Kritik am DBR-Instrument, die während der Durchführung der zweiten Erprobungsphase von den teilnehmenden Rater_innen genannt wurden, und der Einschätzung externer Experten (zwei Dozenten der TU Dortmund und der Universität Rostock) wurde das Instrument zur Verhaltensverlaufsdiagnostik in zwei Durchgängen (siehe Anhang D und Anhang E) erneut angepasst. Die daraus entstandene dritte Version wurde im letzten Schritt wiederum in einigen Punkten verändert, bevor in der an diese Arbeit anknüpfenden Studie eine Finalversion in verschiedene Grund- und Gesamtschulen ging. Hierauf soll in Kapitel 4.4 näher eingegangen werden. Gegenstand dieses Kapitels ist die in zwei Schritte aufgeteilte vorerst letzte Überarbeitung des DBR-Instruments. Im ersten Schritt wurde das DBR auf Basis der Rückmeldungen der geschulten Rater_innen angepasst und eine vorläufige Version erstellt, die anschließend in Zusammenarbeit mit externen Experten bearbeitet und angepasst wurde.

Während dieser Erprobungsphase standen diverse Fragen bezüglich der Überarbeitung des Rating-Instruments im Raum. Unter anderem wurde darüber gesprochen, ob eine weitere Spalte mit der Beschriftung „keine Angabe möglich“ im Bereich der Skala hinzugefügt werden sollte. Dieser Gedanke wurde im Gespräch jedoch wegen der Befürchtung verworfen, die beurteilenden Lehrkräfte oder anderen pädagogischen Fachkräfte könnten vorschnell auf die

Möglichkeit zurückgreifen könnten, „keine Angabe“ zu machen, obgleich ein Ausfüllen eventuell doch machbar wäre. Es wurde daraufhin beschlossen, den Rater_innen das Auslassen eines Items zuzugestehen, falls sie dieses ihrer Meinung nach für einen bestimmten Beobachtungszeitraum nicht ausfüllen können.

Insbesondere von B1 und B2 wurden acht Items kritisiert und infolgedessen überarbeitet. Item 1, das in der zweiten überarbeiteten Version „Meldet sich im Unterricht“ lautete, wurde in einem Zwischenschritt erweitert und lautete anschließend „Meldet sich im Unterricht, hält sich an Gesprächsregeln“. Diese Anpassung basierte auf der zu diesem Zeitpunkt gültigen, später von oben genannten Experten widerrufenen Annahme, jeder Verhaltensbereich solle im Rahmen einer übersichtlichen und gleichmäßigen Gestaltung des Rating-Instruments aus drei Items bestehen. Die Zusammenfassung zweier Items, in diesem Fall 1 und 3 („Meldet sich im Unterricht, hält sich an Gesprächsregeln.“), war auf dieser Basis unumgänglich, da ein weiteres Item im Verhaltensbereich des Schulbezogenen Verhaltens hinzugefügt werden sollte. Nur so konnte Platz für ein Item zur Verweigerung der Mitarbeit von Schüler_innen geschaffen werden, welches schon im Rahmen der Expert_inneninterviews (siehe Kapitel 4.2.3) angesprochen wurde. Item 3 wurde dementsprechend in „Verweigert die Mitarbeit“ umbenannt (vorher: „Redet im Unterricht oft dazwischen.“).

Auch Item 2 wurde in der vorläufigen dritten Version des DBRs angepasst (siehe Anhang D). Aus „Arbeitet ruhig und konzentriert im Unterricht“ wurde „Richtet seine Aufmerksamkeit/Konzentration auf die Bearbeitung der Aufgabe“. Diese Änderung basiert ebenfalls auf Rückmeldungen der beiden geschulten Beobachterinnen, welche in der Praxis die Notwendigkeit feststellten, die Attribute „ruhig“ und „konzentriert“ im Rahmen von Aufgabenfokussierung und Aufmerksamkeit zu trennen, da Schüler_innen, die auf den ersten Blick ruhig erschienen, nicht zwingend konzentriert an ihren Aufgaben arbeiteten. Vielfach saßen diese Schüler_innen an ihrem Platz ohne zu arbeiten, waren verträumt oder abgelenkt und dementsprechend wenig konzentriert. Das neue Item enthält deshalb die Aspekte der auf Aufgaben gerichteten Aufmerksamkeit und Konzentration.

Item 4, welches in der zweiten überarbeiteten Version „Verhält sich wütend und aufbrausend“ lautete, wurde in der vorläufigen dritten Version um den Bereich der geringen Frustrationsto-

leranz erweitert, welche in vielen Fällen mit wütendem und aufbrausendem Verhalten einhergeht. Das Item lautete im Folgenden „Verhält sich wütend und aufbrausend, hat eine geringe Frustrationstoleranz“.

Da insbesondere B1 in Bezug auf die Items 7 („Zappelt und ist motorisch unruhig“) und 9 („Unruhig und überaktiv“) eine Dopplung sah, wurde Item 9 vorerst aus dem DBR-Instrument herausgenommen. Dadurch entstand Platz für ein weiteres Item im Bereich der Hyperaktivität. In der vorläufigen dritten Version erscheint, mit Blick auf seine nicht zweifelsfrei erwiesene Eignung für den entsprechenden Verhaltensbereich nur unter Vorbehalt, das Item 9 „Spielt mit den Arbeitsmaterialien“.

Auch bei den Items 14 und 15, die sich auf Unterstützungs- beziehungsweise Hilfsbereitschaft beziehen, wurde in Bezug auf diese Dopplung eine Anpassung vorgenommen. Item 15 wurde aus dem Rating-Instrument entfernt und vorerst unbesetzt gelassen. Vor der neuen Eingliederung eines Items zum Prosozialen Verhalten der Schüler_innen sollte zum einen Rücksprache mit externen Experten geführt und zum anderen Literaturrecherche in diesem Bereich betrieben werden.

Im Bereich der Verhaltensprobleme mit Gleichaltrigen wurden zwei Items angepasst. Item 17 wurde um den Zusatz „lässt sich provozieren“ ergänzt, da dieser Aspekt bisher nicht aufgeführt worden war und in der zweiten Erprobungsphase einen verhältnismäßig großen Raum eingenommen hatte. Das neue Item 17 heißt dementsprechend „Wird von anderen Kindern gehänselt oder geärgert, lässt sich provozieren“. In Item 18 wurde das Wort „lieber“ durch das Wort „häufiger“ ersetzt. Dies soll eine objektivere Beurteilung zulassen und die Bewertung insgesamt erleichtern, da die Häufigkeit laut Rückmeldung von B1 und B2 einfacher zu beurteilen ist als die Vorliebe und so auch Fälle eingeschlossen werden, in denen Schüler_innen zwar nicht lieber, aber häufiger mit Erwachsenen aus dem Schulkontext arbeiten. Gründe hierfür können Mobbing oder andere Probleme mit Mitschüler_innen sein, welche in großem Maße für Probleme mit Gleichaltrigen stehen können.

Neben der Veränderung der oben erläuterten Items wurde über zusätzliche Items gesprochen. B1 und B2 diskutierten insgesamt über vier Items („Ist geistig abwesend, träumt“, „Fordert Hilfe von der Lehrkraft ein“, „Nutzt Hilfsmittel angemessen“ und „Fordert die Aufmerksamkeit der Lehrkraft ein“). Diese wurden letztlich nicht in das Rating-Instrument aufgenommen, da kaum Übereinstimmungen mit dem zugrundeliegenden SDQ vorliegen. Die Instruktionen

wurden vorerst nicht weiter verändert, da zu dem Zeitpunkt des Überarbeitungsschrittes bereits Gespräche über die in Kapitel 4.4 beschriebene Studie geführt wurden, welche ihrerseits eine adaptierte, aber erst zu einem späteren Zeitpunkt erstellbare Instruktion erfordert.

Aus der vorläufigen dritten überarbeiteten Version (siehe Anhang D) wurde im Gespräch mit einem externen Experten und nach weiterführender Literaturrecherche zum Verhaltensbereich der Hyperaktivität und des Prosozialen Verhaltens die dritte überarbeitete Version erstellt (siehe Anhang E), welche als Grundlage für die finale Version (siehe Anhang F) dient. Die weiterhin aus 18 Items bestehende dritte überarbeitete Version wurde erneut verändert. Insbesondere die Items 7, 9, 11, 12, 15 und 16 wurden modifiziert. Außerdem wurde der Begriff Kinder einheitlich durch das ebenfalls genderneutrale und altersübergreifende Wort „Mitschüler_innen“ ersetzt. Die Spalte „keine Angabe möglich“ ist aufgrund oben genannter Gründe in der dritten überarbeiteten Version nicht mehr enthalten.

Bei Item 7 wurde der Begriff „motorisch“ in Klammern gesetzt, sodass die von ihm anvisierte Unruhe nicht nur auf den körperlichen Bereich beschränkt ist, und der im Bereich Hyperaktiven Verhaltens typische Aspekt der Überaktivität hinzugefügt. Das Item lautet entsprechend nicht mehr „Zappelt und ist motorisch unruhig“ sondern „Zappelt, ist (motorisch) unruhig/ überaktiv“.

Item 9, welches in der vorläufigen dritten überarbeiteten Version vorerst als „Spielt mit den Arbeitsmaterialien“ formuliert wurde, konnte auf Basis der Literaturrecherche durch eine für Hyperaktivität passendere Verhaltensweise ersetzt werden („Lässt sich schnell und leicht ablenken“). Hierbei handelt es sich um die schnelle und leichte Ablenkbarkeit, welche hyperaktive Kinder in Bezug auf die Durchführung von Aufgaben und das Zuhören im Klassenkontext sowie im allgemeinen Schulalltag häufig zeigen (siehe Kapitel 2.1.1). Je nach dem Grund der Ablenkung kann der Bereich des Spielens mit Arbeitsmaterialien weiterhin unter diesem nun globaler formulierten Item erfasst werden.

Ebenfalls verändert wurden die Items 11 und 12, bei denen das Wort „verhält“ jeweils durch das Wort „wirkt“ ersetzt wurde. So konnte eine Angleichung an Item 10 herbeigeführt und vom definitiven Verhalten zur Wirkung der Schüler_innen gewechselt werden. Dies ist im Rahmen der immer subjektiv konnotierten Beurteilungen insofern sinnvoll, als die beobachtende und bewertende Person sich nie sicher sein kann, ob die beobachteten Lernenden sich

tatsächlich etwas nervös oder ängstlich verhalten oder es nur so auf den Betrachter wirkt. Außerdem wurde bei Item 12 von der Formulierung „klammert sich an Erwachsene“ zu „(sucht Nähe zu Erwachsenen)“ gewechselt. Die Formulierung des Anklammerns erschien etwas zu umgangssprachlich und zu ungenau gewählt. Die Items 11 und 12 lauten in der dritten überarbeiteten Version folglich „Wirkt ängstlich/ fürchtet sich“ und „Wirkt nervös (sucht Nähe zu Erwachsenen)“.

In der dritten überarbeiteten Version (siehe Anhang E) wurde auf Basis von Literaturrecherche ein weiterer schulrelevanter Bereich des Prosozialen Verhaltens mit in das Rating-Instrument aufgenommen, der Aspekt des kooperativen Verhaltens. Im schulischen Kontext findet auf vielfältige Weise Zusammenarbeit der Schüler_innen statt. Bei solchen Gruppen- und Partnerarbeiten kommt es auf die Fähigkeit an, kooperativ zu arbeiten, gemeinsam Lösungen zu finden, Konflikte zu lösen und am Ende ein (Gruppen-)Ergebnis präsentieren zu können. Diese Kompetenzen fallen unter den Begriff des Prosozialen Verhaltens (siehe Kapitel 2.1.1), weshalb Item 15 in „Verhält sich in Partner- und Gruppensituationen kooperativ“ umbenannt wurde.

Zur Vereinheitlichung und um das Arbeiten zu betonen, welches in zahlreichen schulischen Kontexten mehr Raum als das Spielen einnimmt, wurde in Item 16 die Reihenfolge von „spielen“ und „arbeiten“ an Item 18 angepasst. Item 16 lautet dementsprechend nun „Arbeitet/ spielt meist alleine“.

Die dritte Version des DBRs (siehe Anhang E) wurde erneut von beiden externen Experten begutachtet und modifiziert. So entstand die Finalversion (siehe Anhang F), welche in der im Anschluss an die vorliegende Arbeit durchgeführten Studie in größerem Rahmen überprüft wurde. Außerdem wurde eine Anpassung der Valenz der Zielformulierung vorgenommen, sodass die Items eines Verhaltensbereichs entweder alle erwünschte Verhaltensweisen oder alle von den im schulischen Kontext geltenden Regeln abweichendes Verhalten beurteilen.

Eine der größten Änderungen von der dritten (siehe Anhang E) zur vierten, finalen Version (siehe Anhang F) besteht darin, dass Item 1 aus der dritten Version in zwei Items aufgeteilt wurde. Dies geschah auf Grundlage einer Rückmeldung eines Experten der Universität Rostock, der dafür plädierte, im Bereich des Schulbezogenen Verhaltens vier anstatt wie vorher drei Items zur Beurteilung zuzulassen. Item 1, welches in der dritten Version „Meldet sich im Unterricht, hält sich an Gesprächsregeln“ lautete, wurde in zwei Items getrennt. Die neuen

Items 1 und 2 der Finalversion sind dementsprechend auf das Meldeverhalten der Schüler_innen und die im Schul- oder Klassenkontext geltenden Gesprächsregeln ausgerichtet und heißen „Meldet sich im Unterricht“ (1) und „Hält sich an Gesprächsregeln“ (2). Vor der Aufteilung des Items lag der Fokus hauptsächlich auf dem generellen (unterrichtsbezogenen) Kommunikationsverhalten der Schüler_innen und insbesondere auf der Frage, ob sich die Schüler_innen melden und warten, bis sie von der Lehrkraft zum Sprechen aufgefordert werden, oder ob sie in die Klasse rufen. Nach der Aufteilung ist Item 1 auf die generelle Beteiligung der Schüler_innen am Unterricht gerichtet, während Item 2 weiterhin auf die Kommunikation und die geltenden Gesprächsregeln bezogen ist. Es kann nun zwischen zwei verschiedenen in der Schule relevanten Verhaltensweisen, der aktiven Unterrichtsteilnahme und der Einhaltung von Regeln für die im Klassen- und Schulkontext stattfindende Kommunikation, unterschieden werden. Diese genauere Formulierung und Operationalisierung kann zu einer niedrigeren Inferenz führen, was die Beurteilung dieser Items für die pädagogischen Fachkräfte vereinfacht und einer ökonomischen Bearbeitung entgegenkommen kann (siehe hierzu Kapitel 2.2.2). In Folge der Aufteilung des Items 1 verschiebt sich die Nummerierung der folgenden Items um jeweils eine Stelle nach hinten (siehe Anhang F).

Item 3 und Item 5 aus der dritten überarbeiteten Version, welche in der Finalversion nun die Nummern 4 und 6 tragen, wurden erweitert und umformuliert. Um zu gewährleisten, dass alle Items innerhalb eines Verhaltensbereichs entweder ein positives, erwünschtes Verhalten oder ein negatives, unerwünschtes Verhalten beschreiben, und eine spätere Auswertung der ausgefüllten Rating-Bögen zu erleichtern, wurden beide Items ins Negative umformuliert. Die Kritik der Raterinnen im Rahmen der Expert_inneninterviews führte zu einer relevanten Veränderung. Es war angemerkt worden, dass die Komplexität der Beurteilung eines nicht auftretenden Verhaltens mit dem Extremwert „Nie“ der siebenstufigen Skala zu groß sei für die kurze Zeitspanne, die im schulischen Kontext für solche Beobachtungen zur Verfügung stehe. Die Items wurden deshalb am Anfang mit einem positiv formulierten Zusatz versehen, an dem die beobachtenden und bewertenden Personen ihre Beurteilung orientieren können. Im Fall des vierten Items aus der Finalversion handelt es sich um den Zusatz „Arbeitet ruhig am Platz“, für das sechste Item wurde der Verweis auf die Missachtung von Regeln hinzugefügt. Beide Items beginnen dementsprechend in der Finalversion mit einem positiv formulierten Teil und schließen mit einer negativen Formulierung („Arbeitet ruhig am Platz und verweigert nicht

die Mitarbeit“ (4) und „Missachtet Regeln und hört nicht auf die Lehrkraft“ (6)). Die Änderung dieser beiden Items ist in zweierlei Hinsicht als Verbesserung zu betrachten. Zum einen wird die durchgehende Bewertung eines erwünschten oder nicht erwünschten Verhaltens innerhalb eines Verhaltensbereichs durch die Anpassung der Items ermöglicht und zum anderen wird durch das Hinzufügen des positiv formulierten Zusatzes die Beurteilung der Items erleichtert.

Item 8 aus der überarbeiteten dritten Version (Item 9 aus der Finalversion) wurde ebenfalls so adaptiert, dass es nun wie die beiden anderen Items aus dem Verhaltensbereich der Hyperaktivität ein unerwünschtes Verhalten misst. Zu diesem Zweck wurde es von „Führt Aufgaben zu Ende“ zu „Bricht Aufgaben häufig früh ab“ umformuliert.

Da das DBR in der vorliegenden Form keine zeitlichen Angaben zulässt, sondern eher qualitativ misst, ob ein Verhalten auftritt, wurde Item 9 aus der überarbeiteten dritten Version (Item 10 aus der Finalversion) gekürzt, sodass es nun nicht mehr die schnelle und leichte Ablenkbarkeit der Schüler_innen in den Fokus stellt, sondern sich nur noch auf die generelle Ablenkbarkeit an sich bezieht und den zeitlichen Aspekt außen vor lässt.

Auch Item 13 (vorher 12) wurde gekürzt, sodass es von der eher spezifischen Formulierung „Wirkt nervös (sucht Nähe zu Erwachsenen)“, welche die Art des nervösen Verhaltens bereits durch den in Klammern stehenden Zusatz vorgibt, zu einer globaleren Formulierung („Wirkt nervös“) geändert wurde. Die Auswirkungen der globalen Formulierung von Items in der Verhaltensverlaufsdiagnostik sind noch nicht hinreichend erforscht. Die Umformulierung kann in diesem Fall entweder dazu führen, dass die Beurteilung für die das Rating anwendenden Lehrkräfte aufwendiger wird (siehe hierzu den Abschnitt zur Inferenz in Kapitel 2.2.2), oder dazu, dass die Beobachtungsgenauigkeit steigt (siehe hierzu den Abschnitt zur Formulierung des Beobachtungsziels in Kapitel 2.4). Die Entscheidung, das Item globaler zu formulieren, beruhte auf dem Gedanken, dass nervöses Verhalten sich nicht nur durch das Klammern an Erwachsene zeigen, sondern sich vielfältig ausdrücken kann. Um den Lehrkräften die Einordnung einer größeren Bandbreite an Verhaltensweisen in diesem Bereich zu ermöglichen, wurde eine globalere Formulierung gewählt. Diese Entscheidung muss in weiteren Untersuchungen überprüft werden, da, wie oben erwähnt und in Kapitel 2 an verschiedenen Stellen erläutert, die Studien diesbezüglich unterschiedliche Ergebnisse liefern.

Item 17 der Finalversion wurde um einen Zusatz erweitert, der in einem vorherigen Anpassungsschritt aus einem anderen Item entfernt wurde. Das Item „Arbeitet/spielt meist alleine“ wurde um den Aspekt „lieber“ ergänzt, sodass das Item nun „Arbeitet/spielt meist oder lieber alleine“ lautet. Dies ist damit zu begründen, dass die durch dieses Item zu messenden Verhaltensweisen nicht in jeder Unterrichtssituation beobachtet werden können und der Zusatz „lieber“ dem Item eine weitere Komponente zuteilt, welche die Beobachtung einer entsprechenden Verhaltensweise wahrscheinlich und damit die Beurteilung dieses Items leichter macht.

Auch im Bereich der Instruktion und des Layouts wurden Veränderungen vorgenommen. Während die dritte überarbeitete Version keinerlei Instruktion enthielt, da es sich bei dieser Version um eine reine Arbeitsvorlage handelte, wurde die Instruktion bei der Finalversion, welche im Rahmen der anschließenden Studie in einer Mappe mit anderen Rating-Bögen und Informationsmaterialien in Gesamt- und Grundschulen gegeben wurde, ausgegliedert und auf einem gesonderten Blatt am Anfang der Mappe verortet. Zudem wurde die Instruktion durch die Nutzung von Zwischenüberschriften und Spiegelstrichen übersichtlicher gestaltet. Auch das Layout der DBR-Bögen wurde angepasst. Die Überschriften der sechs Verhaltensbereiche wurden grau unterlegt, um auch hier eine übersichtliche Gestaltung zu generieren. Weiterhin stehen die Zahlen der siebenstufigen Skala in jeder auszufüllenden Zeile und die verbale Beschriftung der Extremwerte („Nie“ und „Immer“) ist nur am Kopf des Bogens zu finden.

Der gesamte Entwicklungsverlauf der einzelnen Items ist in Anhang L nachzuvollziehen. Nach der oben beschriebenen vorerst letzten Überarbeitung wurde das DBR in einer in Gesamt- und Grundschulen durchgeführten Pilotierungsstudie weiter untersucht. Diese wird in Kapitel 4.4 zusammengefasst dargestellt.

4.4 Zusammenfassung der anschließenden Pilotierungsstudie

Zur Beantwortung von Forschungsfrage 3 b), die sich insbesondere darauf bezieht, wie durch Weiterentwicklung des vorliegenden Instruments die Reliabilität der Ergebnisse erhöht werden kann, soll nun auf die an diese Arbeit anknüpfende und im Rahmen einer Masterarbeit von Hisker durchgeführte Pilotierungsstudie (im Druck) eingegangen werden. Ein Ziel der Pilotierungsstudie betrifft ebendiese in Forschungsfrage 3 b) angesprochene Weiterentwicklung der Methode des DBRs. Hierbei orientiert sich die Studie an den für die Verlaufsdiagnostik relevanten Gütekriterien (siehe Kapitel 2.2.2).

Die Daten wurden innerhalb der Pilotierungsstudie zum einen quantitativ durch die Ratingskalen und zum anderen qualitativ durch halbstrukturierte Expert_inneninterviews gewonnen. Die Auswertung der aus den Ratingskalen erhobenen Daten orientiert sich an einem Artikel von Gebhardt und Voß (2017), in dem die Autoren drei Forderungen an verhaltensverlaufsdiagnostische Instrumente stellen: Das Instrument muss erstens den psychometrischen Gütekriterien aus dem Bereich der Statusdiagnostik entsprechen, zweitens die psychometrischen Eigenschaften der Verlaufsdiagnostik berücksichtigen und drittens ökonomisch sein sowie einen positiven Einfluss auf schulische Unterrichtsprozesse haben. Auch Hisker (im Druck) orientiert sich in der Auswertung ihrer Studie an diesen Forderungen und untersucht die Reliabilität des Verfahrens, um Aussagen darüber treffen zu können, ob sich das vorliegende Instrument für statusdiagnostische Zwecke eignet. Um zu klären, ob auch eine Eignung im verlaufsdiagnostischen Rahmen besteht, wird zudem die Änderungssensibilität untersucht. Des Weiteren werden die Messbedingungen überprüft, unter denen das Instrument angewendet worden ist, und untersucht, ob diese eine zuverlässige Erfassung der Verhaltensweisen von Schüler_innen an Gesamt- und Grundschulen zulassen. Zur Beantwortung der Frage nach der Notwendigkeit einer Weiterentwicklung des Instruments wurden, wie bereits erwähnt, Expert_inneninterviews durchgeführt., Im Rahmen dieser Interviews wurden aber nicht nur praxisrelevante Einschätzungen der Lehrkräfte zu erforderlichen Adaptions- und Modifizierungsmaßnahmen erhoben, sondern auch Fragen zur Ökonomie und Nützlichkeit des Instruments gestellt. Konstatiert werden kann daher, dass Hisker (im Druck) in ihrer Pilotierungsstudie allen drei Forderungen von Gebhardt und Voß (2017) nachkommt.

Durchgeführt wurde die Pilotierungsstudie in einem Zeitraum von drei Monaten an fünf Grundschulen und vier Gesamtschulen. Insgesamt 39 Lehrkräfte beurteilten mit Hilfe der finalen Version des DBRs das Verhalten von Schüler_innen verschiedener Klassenstufen. Von den 39 Lehrkräften unterrichteten 23 an Grundschulen und 16 an Gesamtschulen. Vier der 39 Lehrkräfte sind als Sonderpädagog_innen im Primarbereich, drei an Gesamtschulen als tätig. 31 der Lehrkräfte hatten zum Zeitpunkt der Pilotierungsstudie die Klassenleitung der von ihnen für die Durchführung gewählten Klasse inne. Die Grundschullehrer_innen arbeiteten, wie erwartet, im Durchschnitt mehr Stunden pro Woche mit den von ihnen beurteilten Schüler_innen zusammen als die Gesamtschullehrkräfte. Ein Drittel der Lehrkräfte konnte Vorerfahrungen mit Ratingskalen vorweisen. Insgesamt wurden 205 Lernende, hiervon 108 Grundschüler_innen und 97 Gesamtschüler_innen, im Rahmen der Pilotierungsstudie beurteilt. 60

der beobachteten Lernenden waren weiblichen und 145 männlichen Geschlechts, wobei sich die Geschlechterverteilung bei Betrachtung der zwei Schulformen nicht nennenswert unterscheidet. Die Schüler_innen waren zum Zeitpunkt der Durchführung der Pilotierungsstudie zwischen 6;4 und 18;5 Jahre alt und befanden sich in den Klassenstufen eins bis zehn. 35 der Schüler_innen wiesen einen sonderpädagogischen Förderbedarf auf. Im Anschluss an die Studie wurden jeweils eine Regelschullehrkraft und eine sonderpädagogische Lehrkraft pro Schulform interviewt.

Zur Durchführung der Pilotierungsstudie erhielt jede Lehrkraft eine Mappe mit allgemeinen Informationen sowie Instruktionen und den Ratingbögen. Bei der Übergabe dieser Mappen wurde mindestens eine Lehrkraft pro Schule persönlich instruiert und darum gebeten, die erhaltenen Informationen an andere teilnehmende Kolleg_innen weiterzugeben. Lehrkräfte, die aufgrund der geographischen Verteilung der Schulen und der Größe der Stichprobe nicht persönlich instruiert werden konnten, hatten deshalb entweder über Kolleg_innen oder durch die Mappe Zugang zu den notwendigen Informationen. Alle Lehrkräfte wurden dazu aufgefordert, das Instrument für fünf Schüler_innen an fünf aufeinanderfolgenden Unterrichtstagen durchzuführen. Hierbei wurde betont, dass immer dieselbe Lehrkraft das DBR anwenden müsse.

Die Ergebnisse der Studie zeigen, dass die Werte bezüglich der internen Konsistenz für alle Skalen über alle Messzeitpunkte ausnahmslos im akzeptablen Bereich sind (Cronbachs $\alpha \geq .75$). Insbesondere für die Skalen Externalisierendes Verhalten, welche die Verhaltensbereiche Hyperaktivität und Verhaltensprobleme umfasst, und Prosoziales Verhalten konnten gute Werte erhoben werden (Cronbachs $\alpha \geq .86$), die auf eine stabile Reliabilität hinweisen. Dementsprechend können mit Hilfe der Ratingskala die mit ihr zu beurteilenden Verhaltensbereiche im inklusiven Setting von Grund- und Gesamtschule reliabel erfasst werden. Damit ist die erste Forderung erfüllt, welche Gebhardt und Voß (2017) an Instrumente zur Verhaltensverlaufsdiagnostik stellen.

In Bezug auf das Gütekriterium der Trennschärfe ergab die Studie für die Items 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 und 16 aus der finalen Version des DBRs hohe Werte über alle fünf Messzeitpunkte (korrigierte Item-Skala-Korrelation $r_i > .5$). Für die Items 17 (zu einem Messzeitpunkt), 18 und 19 (zu allen fünf Messzeitpunkten) konnten nur mittelmäßige Werte hinsichtlich der Trennschärfe festgestellt werden ($.3 < r_i < .5$). Allerdings sänke hier zumeist

die interne Konsistenz, wenn das jeweilige Item weggelassen würde. Ebenfalls nur mittelmäßige (zu drei Messzeitpunkten $.3 < r_i < .4$) bis niedrige Trennschärfen (zu zwei Messzeitpunkten $r_i < .3$) weist das Item 1 auf. Festgehalten werden kann daher, dass die Ratingskala in den Skalen Externalisierendes Verhalten, Internalisierendes Verhalten (Emotionale Probleme und Verhaltensprobleme mit Gleichaltrigen) und Prosoziales Verhalten insgesamt auch dem verlaufsdiagnostisch relevanten Gütekriterium der Trennschärfe entspricht. Die Interpretation der Ergebnisse im Bereich des Schulbezogenen Verhaltens gestaltet sich schwierig, da die Trennschärfe des ersten Items nur niedrig bis mittelmäßig ausgeprägt ist. Insbesondere hier als auch im Bereich der Verhaltensprobleme mit Gleichaltrigen müssen demnach weitere Studien zur Überprüfung folgen.

Die auf Basis der Pilotierungsstudie für das vorliegende Instrument erfassten Korrelationsergebnisse zeigen, dass schulformübergreifend für alle Skalen ein Reliabilitätsabfall der Verhaltensbeurteilungen über die fünf Messzeitpunkte gemessen werden kann. Dies spricht laut Klauer (2011) für ein änderungssensibles Testinstrument, das also in kurzen Zeiträumen bereits kleine relevante Veränderungen abbilden kann. Dieses Gütekriterium der Änderungssensibilität erfüllt die Ratingskala sowohl im Kontext der Grundschule als auch in dem der Gesamtschule. Eine Analyse der individuellen Verhaltensveränderungen von Schüler_innen und in der Folge auch das Treffen von Entscheidungen über Fördermaßnahmen ist dementsprechend mit Hilfe des Instruments möglich. Die Ratingskala erfüllt damit auch die zweite Forderung, die Gebhardt und Voß (2017) an verhaltensverlaufsdiagnostische Instrumente stellen.

Insgesamt kann also, basierend auf den oben zusammengefassten Ergebnissen der Pilotierungsstudie, konstatiert werden, dass dem vorliegenden Ratinginstrument eine akzeptable Testgüte zugeschrieben werden kann. Die Skala kann dementsprechend für die Erfassung von Verhaltensverläufen in den Bereichen des Externalisierenden, Internalisierenden, Prosozialen und Schulbezogenen Verhaltens genutzt werden.

Hisker untersuchte in ihrer Pilotierungsstudie (im Druck) auch die Art und Weise, wie Lehrkräfte das vorliegende Instrument zur Verhaltensverlaufsdiagnostik an Grund- und Gesamtschulen einsetzen. Das Rating wurde unter Alltagsbedingungen im natürlichen schulischen Kontext durchgeführt und lässt dementsprechend Ableitungen für den zukünftigen Einsatz des Instruments im Handlungsfeld von Grund- und Gesamtschulen zu. Insgesamt konnte festge-

stellt werden, dass Grundschullehrkräfte die fünf aufeinanderfolgenden Beurteilungen in einem kürzeren Zeitraum durchführen konnten als Gesamtschullehrer_innen. Ihnen ist es dementsprechend aufgrund ihres meist täglichen Kontaktes mit denselben Schüler_innen früher möglich, eine Aussage über die Effektivität der Förderung von Schüler_innen zu treffen. 65,2 % der Grundschullehrkräfte wählten einen Beobachtungszeitraum, der sich über den gesamten Schultag erstreckte. Damit beobachteten sie durchschnittlich längere Zeiträume als die Lehrkräfte an Gesamtschulen, von denen 82,9 % sich nur auf die Beobachtung einer einzelnen Schulstunde konzentrierten. Dies ist darauf zurückzuführen, dass wegen des an Grundschulen herrschenden Klassenlehrer_innenprinzips die dortigen Lehrkräfte deutlich mehr Zeit in einer Klasse verbringen als Gesamtschullehrkräfte, die in einer Klasse häufig nur ein Fach unterrichten.

Des Weiteren ergab sich im Rahmen der Interviews, die im Anschluss an die Durchführung des Instruments geführt wurden, dass alle Lehrkräfte dem vorliegenden Rating eine schnelle und häufige Durchführbarkeit zusprachen. Der Aspekt der Ökonomie ist damit ebenfalls erfüllt. Auch in Bezug auf den im Vergleich zu anderen, umfangreicheren Diagnoseinstrumenten verhältnismäßig geringen Materialaufwand kann von einem ökonomischen Verfahren gesprochen werden. Während die Ökonomie im Rahmen der Durchführbarkeit festgestellt werden konnte, konnte in Bezug auf die Auswertung seitens der Lehrkräfte keine Aussage getroffen werden, da zu dem Zeitpunkt noch kein entsprechendes Auswertungstool existierte. Eine der sonderpädagogischen Lehrkräfte äußerte den expliziten Wunsch nach einer computergestützten Möglichkeit der Auswertung, was in weiterführenden Arbeiten berücksichtigt werden muss.

In Bezug auf den Umfang der Ratingskala gaben zwei Gesamtschullehrkräfte an, sich deren Erweiterung um eine zusätzliche Seite und eine dementsprechend umfänglichere Erfassung vorstellen zu können. Dies ist kritisch zu beurteilen, da die Ökonomie sich vermindern, die Komplexität sich allerdings erhöhen würde. Des Weiteren gab eine Grundschullehrkraft an, die Ökonomie sei in der Durchführung der Skala durch einen Valenzwechsel beeinträchtigt worden. Die Items aus dem Bereich des schulbezogenen Verhaltens sind positiv, die darauffolgenden Items (VP, HY, EP) negativ und die des letzten Bereichs (PS) erneut positiv formuliert. In Kapitel 2.4 bereits erwähnte Studien geben an, dass die positive Formulierung für Items, welche die Teilnahme am Unterricht operationalisieren, und die negative Formulierung

von Items, die sich auf störendes Verhalten und respektvolles Verhalten beziehen, sinnvoll seien. Unter Berücksichtigung dieser Studienergebnisse wäre eine Umformulierung der Items wenig nachvollziehbar. Eine Neuanordnung der Verhaltensbereiche, bei der zuerst positive und dann negative Items aufgelistet werden, könnte hier allerdings Abhilfe leisten.

Ebenfalls positive Ergebnisse ergaben die Interviews mit den Lehrkräften hinsichtlich des Gütekriteriums der Nützlichkeit. Die Lehrkräfte bezeichneten das Instrument durchweg als nützlich und gaben an, es insbesondere zur Informationsgewinnung und für kollegiale Gespräche nutzen zu wollen.

Auch Hinweise auf den Grad der Inferenz des vorliegenden Instruments ergaben sich im Rahmen der Expert_inneninterviews. Die Lehrkräfte berichteten, sie hätten die Items 1, 2, 3, 5, 8, 9, 16, 17 und 19 der finalen Version des DBRs aufgrund der konkreten Operationalisierung schnell beurteilen und ausfüllen können. Dies spricht für eine niedrige, im Rahmen von Verhaltensverlaufsdiagnostik positiv zu bewertende Inferenz (siehe hierzu Kapitel 2.2.2). Die Items 4, 5, 6, 7 und 18 hingegen wurden von den Lehrkräften als schwierig beurteilt, weil sie verschiedene, nicht unbedingt direkt zusammenhängende Verhaltensweisen in einem Item aufführten. Dies kann in einer zukünftigen Version durch die ausschließliche Aufnahme eindeutiger Items umgangen werden. Insbesondere die Ökonomie muss dann allerdings erneut überprüft werden.

Insgesamt kann festgehalten werden, dass das in dieser Arbeit entwickelte Instrument viele der Gütekriterien erfüllt und ihm daher eine akzeptable Testgüte zugeschrieben werden kann. Die Rückmeldungen der Expert_innen der Pilotierungsstudie an Grund- und Gesamtschulen erlauben das Fazit, dass das Instrument den ersten größeren Praxistest erfolgreich durchlaufen hat. Aufgrund dieses positiven Analyseergebnisses kann das Instrument als durchaus geeignet für die Erfassung schulbezogenen, externalisierenden, internalisierenden und prosozialen Verhaltens bezeichnet werden.

5. Zusammenfassung der Ergebnisse und Ausblick

Das Ziel der vorliegenden Arbeit war die Entwicklung eines verhaltensverlaufsdiagnostischen Instruments zur Anwendung an Regelschulen, das es erlaubt, die Verhaltensentwicklung von Schüler_innen zu messen und die Wirkung eingesetzter Fördermaßnahmen zu überprüfen. Dabei wurde besonderer Wert auf die Praktikabilität und Ökonomie des Instruments gelegt, die

in zwei Erprobungsphasen an Förderschulen ausgewertet und durch Anpassungen der Items verbessert wurden. Anhand dreier Forschungsfragen auf Grundlage des theoretischen Hintergrundes wurde ein Instrument zur Verhaltensverlaufsdiagnostik von Schüler_innen entwickelt, das in einer sich dieser Arbeit anschließenden Pilotierungsstudie im Rahmen einer Masterarbeit von Hisker (i. D., 2018) ausführlich getestet wurde.

Die Basis für die Entwicklung eines Direct Behavior Ratings (DBR) legte der Strengths and Difficulties Questionnaire (SDQ), ein viel erprobtes Instrument zur statusdiagnostischen Erfassung verschiedener Verhaltensweisen. Zunächst wurden hier theoretisch motivierte Anpassungen vorgenommen, welche die Testgüte steigern sollten. Unter anderem wurden zusätzliche Verhaltensbereiche aus dem SDQ in der Ursprungsversion des DBRs ergänzt, da diese lediglich die Bereiche Schulbezogenes Verhalten, Verhaltensprobleme, Hyperaktivität und Emotionale Probleme berücksichtigte. Folgerichtig wurde das Instrument um die Bereiche Prosoziales Verhalten und Verhaltensprobleme mit Gleichaltrigen erweitert. Dazu wurden aus dem SDQ jeweils drei als relevant und aussagekräftig bewertete Items der entsprechenden Verhaltensbereiche ausgewählt und in die Ursprungsversion des DBRs implementiert. Sowohl vor dem Hintergrund der notwendigen ökonomischen Gestaltung eines Ratings als auch der in dieser Arbeit besonders in den Blick genommenen Annahme durch die Lehrkräfte wurde eine Beschränkung auf drei Items pro Verhaltensbereich vorgenommen. Daher umfasst das Instrument 18 Items, die alle sechs bekannten Verhaltensbereiche abdecken. Selbstverständlich wurde die Skalierung der hinzugefügten Items ebenfalls an die Ursprungsversion des DBRs angepasst. Die auf diese Weise entstandene erste Version des DBRs wurde zusammen mit einer überarbeiteten Version des SDQs zur statusdiagnostischen Erfassung von Schüler_innenverhalten in einer ersten Erprobungsphase an einer Förderschule angewendet. Die erste Forschungsfrage, welche die Annahme des Instruments durch die Lehrkräfte zum Inhalt hatte, wurde mithilfe von Expert_inneninterviews bearbeitet. So wurde neben der inhaltlichen Validität auch die soziale Validität berücksichtigt, indem die umfangreichen Rückmeldungen der erfahrenen Förderschullehrer_innen in einer detaillierten Überarbeitung des DBRs resultierten. Trotz eines kleinen Stichprobenumfangs konnten gewinnbringende Ergebnisse erzielt werden. In diesem Zuge wird die zweite Forschungsfrage beantwortet, welche die nötigen Veränderungen und Anpassungen zur Anwendbarkeit des DBRs in der Praxis beinhaltet. Es zeigte sich vor allem, dass eine sehr spezifische Formulierung der Items angebracht ist, was zu einer niedrigeren Inferenz führt und so die Ökonomie und Anwendbarkeit des Instruments

erhöht. Allgemein ist das vorgelegte Instrument von den Expert_innen positiv angenommen worden. Die Lehrkräfte schätzten die Beschäftigung mit Schüler_innenverhalten und die diesbezügliche Diagnostik ungeachtet ihrer vielfältigen Aufgaben grundsätzlich als überaus relevant ein. Das spricht dafür, dass die Konzeption eines praktikablen Ratings zur Verhaltensverlaufsdiagnostik für Lehrkräfte von hoher Bedeutung ist. Einzig der Bereich der Verhaltensprobleme mit Gleichaltrigen bereitete Schwierigkeiten in der Anwendung.

Insgesamt kann festgehalten werden, dass sich der Einbezug der Lehrkräfte als Expert_innen in die Entwicklung des Instrumentes als sehr ertragreich erwies. Besonders die Rückmeldungen zu den Formulierungen der Beobachtungsaufgabe an sich und der einzelnen Items hinsichtlich ihrer Komplexität konnten für die weitere Überarbeitung des DBRs zur besseren Anwendbarkeit in der Praxis genutzt werden. Der gewählte Fokus dieser Arbeit, welcher auf ebendieser Umformulierung und Überarbeitung auf Basis verschiedener Erprobungsschritte lag und in Forschungsfrage 2) zur Sprache kommt, wurde bestätigt.

Die nach der ersten Erprobungsphase vorgenommenen Änderungen wurden in einer zweiten Version des DBRs festgehalten, die erneut zur Erprobung an einer Förderschule getestet wurde. Hier wurde der Fokus neben der zweiten auch auf die dritte Forschungsfrage gelegt, welche die Reliabilität des DBRs im schulischen Kontext aufgreift. Dabei ist die Interrater-Reliabilität von besonderem Interesse, da der Unabhängigkeit von der beobachtenden Person unter dem Gesichtspunkt der Objektivität enorme Bedeutung zukommt. Daher werden an dieser Stelle zwei geschulte Raterinnen, von denen eine die Schüler_innen bereits aus dem Praxissemester kannte, und eine Lehrkraft zur Evaluation eingesetzt. Hierbei ergab sich zwischen den beiden geschulten Raterinnen ein Rangkorrelationskoeffizient nach Spearman von ρ = .837, was eine starke positive Korrelation bedeutet. Trotz eines geringen Stichprobenumfangs des Lehrkraft-Ratings erwies sich mit Spearman-Rangkorrelationen von ρ = .724 und ρ = .853 zu den beiden geschulten Raterinnen auch hier die Interrater-Reliabilität als gegeben. Somit ist die Unabhängigkeit der getesteten zweiten Version des DBRs von den Beobachter_innen gewährleistet. Eine erneute Überarbeitung wurde anschließend auf Grundlage der Rückmeldungen der beiden geschulten Rater_innen und der Einschätzung externer Experten vorgenommen. Insbesondere in Bezug auf die Formulierung der Items konnte festgestellt werden, dass spezifisch formulierte Items in der Praxis besser angenommen wurden. Für die weitestgehend spezifische Formulierung der Items spricht nicht nur die bei diesen erzielte sehr hohe

Interraterübereinstimmung, sondern zudem die Tatsache, dass diese zudem die Inferenz senkt und zu einer ökonomischen Gestaltung des Instruments verhilft. Dies bestärkt die in Kapitel 4.1.3 zuletzt aufgegriffenen Ergebnisse von Volpe und Briesch (2012) und widerspricht denen von Christ et al. (2011). Im letzten Schritt wurden die im Rahmen der Erprobungsphasen des Ratings gewonnenen Erkenntnisse sowie von Experten des Forschungsfeldes eingebrachte Veränderungsvorschläge in einer finalen Überarbeitung implementiert und so die endgültige Version des DBRs entwickelt. In diesem Zuge wurde insbesondere auch die Valenz der Items angepasst und es wurde ein zusätzliches Item hinzugefügt, bevor das so entstandene DBR-Instrument in einer Pilotierungsstudie an Grund- und Gesamtschulen weiter untersucht wurde.

In dieser im Rahmen einer weiteren Masterarbeit durchgeführten und sehr umfangreichen Pilotierungsstudie wurde der Forschungsfrage nachgegangen, wie das vorgelegte DBR-Instrument weiterentwickelt werden kann, um verhaltensverlaufsdiagnostisch reliable Ergebnisse zu erzielen. Dies geschah erneut mittels Expert_inneninterviews. Die Auswertung der gewonnenen Daten lassen auf eine gegebene interne Konsistenz und damit auf reliable Ergebnisse in allen betrachteten Verhaltensbereichen schließen (Cronbachs $\alpha \geq .75$). Des Weiteren erfüllen alle Verhaltensbereiche bis auf das Schulbezogene Verhalten das Gütekriterium der Trennschärfe (korrigierte Item-Skala-Korrelation $r_i > .5$). In Bezug auf das Schulbezogene Verhalten gestaltet sich die Interpretation der Ergebnisse schwieriger, da die Trennschärfe des ersten Items nur niedrig bis mittelmäßig ausgeprägt ist ($.27 < r_i < .37$). Gleichwohl konnte in allen Verhaltensbereichen festgestellt werden, dass ein Reliabilitätsabfall vom ersten zum letzten (hier: fünften) Messzeitpunkt stattfand, was bedeutet, dass das entwickelte Instrument änderungssensibel ist. Insgesamt kann der Schluss gezogen werden, dass eine akzeptable Testgüte gegeben ist und das DBR folglich zur Analyse von Verhaltensveränderungen von Schüler_innen und als Indikator für die Anwendung von Fördermaßnahmen verwendet werden kann. In den im Anschluss geführten Interviews mit den Lehrkräften wurde eine schnelle und häufige Durchführbarkeit des Instruments hervorgehoben. Diese ist im Rahmen des Unterrichts an Grundschulen noch stärker ausgeprägt als an Gesamtschulen, da Grundschullehrkräfte zumeist in täglichem Kontakt mit den Schüler_innen stehen, wohingegen die Lehrkräfte an den Gesamtschulen in der Regel nicht an fünf Tagen pro Woche in derselben Klasse unterrichten. Auch die Beobachtungszeiträume, die von den Grundschullehrer_innen gewählt wurden, unterscheiden sich von denen ihrer Kolleg_innen an den Gesamtschulen und fallen durchschnittlich länger aus. Die Unterschiede sowohl in der Häufigkeit als auch in der Dauer des Kontaktes

mit den Schüler_innen sind darauf zurückzuführen, dass an Grundschulen in der Regel nach dem Klassenlehrer_innenprinzip unterrichtet wird. Die grundsätzliche Durchführbarkeit in Zusammenhang mit dem verhältnismäßig geringen Materialaufwand zeigt, dass die hochrelevante Ökonomie des Instruments gegeben ist. Darüber hinaus wurde das Instrument stets als nützlich eingestuft. Somit kann dieser erste größere Praxistest als erfolgreich bewertet werden.

Durch die vorgenommenen Veränderungen konnte ein Instrument entwickelt werden, das den Anforderungen sowohl theoretischer Grundlagen als auch praktischer Arbeit gerecht wird. Das DBR stellt eine vielversprechende Methode zur Verhaltensverlaufsdiagnostik von Schüler_innen dar. Dennoch muss die Notwendigkeit einer stetigen Reflexion und Evaluation betont werden. Ein hierbei zu berücksichtigendes Feld ist die Gestaltung der Items und ihr Bezug zu einer oder mehreren Verhaltensweisen. Außerdem kann zum Beispiel die derzeit geführte Diskussion um die Bevorzugung globaler oder spezifischer Items anhand des entwickelten Instruments aufgegriffen und überprüft werden. Auch die Frage, ob das vorliegende DBR als Multi-Item-Skala betrachtet oder als eine Verknüpfung mehrerer Single-Item-Skalen aufgefasst werden muss, gilt es in kommenden Überlegungen zu beachten.

Aus Sicht der Forschung zu Schüler_innenverhalten stellt das DBR eine Möglichkeit dar, mit überschaubarem Aufwand große Datensätze über Schüler_innenverhalten zu gewinnen. Hierzu ist der evaluierte Einsatz an unterschiedlichen Schulformen mit mehreren Stichproben erstrebenswert. Den Ansatz für diese Forschung liefert das oben beschriebene Projekt mit seinem Bezug zu Grund- und Gesamtschulen. Die dort gewonnenen Daten könnten zum Beispiel unter dem Aspekt des Schüler_innenverhaltens ausgewertet und mit soziodemographischen Kriterien verglichen werden. Zur Auswertung seitens der Lehrkräfte konnte im Rahmen der bisherigen Untersuchungen keine Aussage getroffen werden, da weder zum Zeitpunkt der Entwicklung des Ratings noch bei der Durchführung der Pilotierungsstudie ein Auswertungstool existierte. In nachfolgenden Arbeiten muss dementsprechend die Erarbeitung eines Verfahrens zur Auswertung des Rating-Instruments in den Blick genommen werden.

Gerade vor dem Hintergrund der Inklusion kommt standardisierten Instrumenten zur Verhaltensverlaufsdiagnostik eine besondere Bedeutung zu. Das in der vorliegenden Arbeit entwickelte Direct Behavior Rating kann hier einen Beitrag leisten, indem es ermöglicht, den Bedarf von Fördermaßnahmen zu ermitteln und deren Wirksamkeit zu überprüfen. Die Nutzung derartiger Instrumente stellt vor allem Lehrkräfte an Regelschulen vor neue Herausforderungen.

Um diese zu bewältigen, benötigen sie Kompetenzen, deren Grundlagen schon im Studium zu legen sind. Daher kann die Implementierung der Vermittlung entsprechender Kenntnisse und Fertigkeiten als Desiderat künftiger Lehrer_innenausbildung angesehen werden.

6. Literaturverzeichnis

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders.* 4th ed. Text revision (DSM-IV-TR). Arlington: VA: American Psychiatric Association.

Beelmann, A. & Raabe, T. (2007). Dissoziales Verhalten von Kindern und Jugendlichen. Erscheinungsformen, Entwicklung, Prävention und Intervention. Göttingen u.a.: Ho grefe.

Blumenthal, Y. (2017). Ein Rahmenkonzept mit mehreren Förderebenen - Response to Intervention (RTI). In: B. Hartke (Hrsg.): *Handlungsmöglichkeiten Schulische In klusion. Das Rügener Modell kompakt. (*S. 20–32). Stuttgart: Kohlhammer.

Blumenthal, Y. & Marten, K. (2017). Mehrebenenkonzept zur Förderung der emotionalen und sozialen Entwicklung und des Verhaltens. In: B. Hartke (Hrsg.): *Handlungsmöglichkeiten Schulische Inklusion. Das Rügener Modell kompakt. (*S. 185-214). Stuttgart: Kohlhammer.

Brennan, Robert L. (2001). *Generalizability Theory*. New York, NY: Springer.

Brezinka, V. (2003). Zur Evaluation von Präventivinterventionen für Kinder mit Verhaltensstörungen. *Kindheit und Entwicklung*, 12 (2), 71-83.

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

Briesch, A., Kilgus, S., Chafouleas, S., Riley-Tillman, T. & Christ, T. (2012). 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.

Bühner, Markus (2011). *Einführung in die Test- und Fragebogenkonstruktion*. (3. Aktuali sierte und erw. Aufl.) München: Pearson Studium.

Casale, G., Grosche, M. & Hennemann, T. (2015a). Zum Beitrag der Verlaufsdiagnostik für eine evidenzbasierte sonderpädagogische Praxis am Beispiel des Förderschwer punktes der emotionalen und sozialen Entwicklung. *Zeitschrift für Heilpädagogik,* 66 (7), 325-334.

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

Casale, G., Briesch, A., Grosche, M., Hennemann, T. & Volpe, R. (2015c). Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse. *Empirische Sonderpädagogik,* 7 (3), 258-268.

Chafouleas, S., Riley-Tillman, T. & McDougal, J. (2002). Good, Bad, or In-Between. How Does the Daily Behavior Report Card Rate? *Psychology in the Schools,* 39 (2), 157–169.

Chafouleas, S., Christ, T., Riley-Tillman, T., Briesch, A. & Chanese, J. (2007). Generalizability and Dependability of Direct Behavior Ratings to Assess Social Behavior of Preschhoolers. *School Psychology Review,* 36 (1), 63-79.

Chafouleas, S., Jaffery, R., Riley-Tillman, T., Christ, T. & 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.

Christ, T. & Boice, C. (2009). Rating Scale Items. *Assessment for Effective Intervention,* 34 (4), 242-250.

Christ, T., Riley-Tillman, T. & Chafouleas, S. (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.

Christ, T., Riley-Tillman, T., Chafouleas, S. & Boice, C. (2010). Direct Behavior Rating (DBR). Generalizability and Dependability Across Raters and Observations. *Educational and Psychological Measurement,* 70 (5), 825–843.

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

Cone, J. (1988): Psychometric Considerations and the Multiple Models of Behavioural Assessment. In A. Bellack & M. Hersen (Hrsg.): *Pergamon General Psychology Series, Vol. 65. Behavioral Assessment: A Practical Handbook* (S. 42-66). Elmsford: Pergamon Press.

Costello, E., Mustillo, S., Erkanli, A., Keeler, G. & Angold, A. (2003). Prevalence and Development of Psychiatric Disorders in Childhoof and Adolescence. *Archives of General Psychiatry,* 60 (8), 837-844.

Cronbach, L., Gleser, G., Nanda, H. & Rajaratnam, W. (1972). *The Dependability of Behavioral Measurements. Theory of Generalizability for Scores and Profiles.* New York: Wiley.

DeVries, J., 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.

DeVries, J., Rathmann, K. & Gebhardt, M. (2018). How Does Social Behavior Relate to Both Grades and Achievement Scores? *Frontiers in Psychology,* 9, n.p.

Dickey, W. & Blumberg, S. (2004). Revisiting the Factor Structure of the Strengths and Difficulties Questionnaire. United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry,* 43 (9), 1159–1167.

Garner, P. (2010). Emotional Competence and its Influences on Teaching and Learning. *Educational Psychology Review,* 22 (3), 298-321.

Gebhardt, M., Casale, G., Jungjohann, J. & DeVries, J. (2017). *Lern-Verlaufs-Monitoring. LEVUMI Lehrerhandreichung. SDQ, DBR & PIQ. Version 1.0, September 2017.* (uneröffentlichtes Dokument).

Gebhardt, M., Diehl, K. & Mühling, A. (2016a). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www.LEVUMI.de. Zeitschrift für Heilpädagogik, 67(10), 444-454.

Gebhardt, M., Diehl, K. & Mühling, A. (2016b). Lern-Verlaufs-Monitoring LEVUMI Lehrerhandbuch. Version 1.1. http://dx.doi.org/10.17877/DE290R-17792

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

Gebhardt, M. (2015). Gemeinsamer Unterricht von Schülerinnen und Schülern mit und ohne sonderpädagogischen Förderbedarf – Ein empirischer Überblick. In E. Kiel (Hrsg.): *Inklusion im Sekundarbereich.* (S. 39-52). Stuttgart: Kohlhammer.

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

Goodman, R. (1997). The Strengths and Difficulties Questionnaire. A Research Note. *Journal of Child Psychology and Psychiatry,* 38 (5), 581-586.

Goodman, A., Lamping, D. & Ploubidis, G. (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.

Grosche, M. & Volpe, R. (2013). Response-to-Intervention (RTI) as a Model to Facilitate Inclusion for Students with Learning and Behaviour Problems. *European Journal of Special Needs Education*, 28 (3), 254-269.

Hartke, B., Sikora, S. & Voß, S. (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.

Hartmann, B. (2017). Verlaufsdiagnostik bei Verhaltens- und Lernschwierigkeiten. In A. Methner, B. Seebach & K. Popp (Hrsg.), *Verhaltensprobleme in der Sekundarstufe. Unterricht - Förderung - Intervention* (S. 74-83). Stuttgart: Kohlhammer.

Hathaway, M., Davin, T. & Steege, M. (2001). Reliability of a Performance-Based Behavioral Recording Procedure. *National Association of School Psychologists,* 30 (1), 252-261.

Hattie, J. (2009). *Visible Learning. A Synthesis of over 800 Meta-Analyses Relating to Achievement.* New York: Routledge.

Hellwig, N. (2010). *Verhaltensstörungen bei Kindern pädagogisch behandeln. Legasthenie, Dyskalkulie, Schulangst und Hyperaktivität erkennen und erfolgreich therapieren* (1. Aufl.). Augsburg: Brigg-Pädagogik.

Hensle, U. (1994). *Einführung in die Arbeit mit Behinderten. Psychologische, pädagogische und medizinische Aspekte* (5. erw. Aufl.). Wiesbaden: Quelle & Meyer.

Hillenbrand, C. & Melzer, C. (2017). Aggressives Verhalten. In A. Methner, B. Seebach & K. Popp (Hrsg.), *Verhaltensprobleme in der Sekundarstufe. Unterricht - Förderung – Intervention (*S. 167–187). Stuttgart: Kohlhammer.

Hisker, S. (im Druck). Veränderungen im Direct Behavior Rating (DBR) über die Zeit – Eine Pilotierung mit fünf Messzeitpunkten in Grund- und Gesamtschulen.

Huber, C. & Casale, G. (2015). Schülerverhalten systematisch erfassen. Die Methode Direct Behavior Rating (DBR). *Praxis fördern - Zeitschrift für individuelle Förderung und Inklusion,* (6), 17-23.

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.

Hussy, W., Schreier, M. & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor. Mit 23 Tabellen.* (2. überarb. Aufl.). Berlin u.a.: Springer (Springer-Lehrbuch).

Ihle, W. & Esser, G. (2008). Epidemiologie psychischer Störungen des Kindes- und Jugendalters. In: B. Gasteiger-Klicpera, H. Julius und C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung*, Bd. 3 (S. 49–62). Göttingen u.a.: Hogrefe.

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. http://dx.doi.org/10.17877/DE290R-18042

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

Klauer, K. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forschung*, (1), 16–26.

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

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

Knopf, H. & Gallschütz, C. (2006). *Prosozialität statt Aggressivität.* Berlin: Rhombos-Verlag.

LeBel, T., Kilgus, S., Briesch, A. & Chafouleas, S. (2009). The Impact of Training on the Accuracy of Teacher-Completed Direct Behavior Ratings (DBRs). *Journal of Positive Behavior Interventions*, 12 (1), 55–63.

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

Lohbeck, A., Schultheiß, J., Petermann, F. & Petermann, U. (2015). Die deutsche Selbstbeurteilungsversion des Strengths and Difficulties Questionnaire (SDQ-Deu-S). *Diagnostica*, 61 (4), 222–235.

Meuser, M. & Nagel, U. (2009). Experteninterview und der Wandel der Wissensproduktion. In A. Bogner (Hrsg.), *Experteninterviews. Theorien, Methoden, Anwendungsfelder* (3. überarb. Aufl.) (S. 35-60). Wiesbaden: Verlag für Sozialwissenschaft.

Pelham, W., Fabiano, G. & Massetti, G. (2005). Evidence-Based Assessment of Attention Deficit Hyperactivity Disorder in Children and Adolescents. *Journal of Clinical Child and Adolescent Psychology*, 34 (3), 449-476.

Preston, C. & Colman, A. (2000). Optimal Numer of Response Categories in Rating Scales. Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, 104 (1), 1-15.

Reinders, H. (2003). Freundschaften im Jugendalter. In: E. Fthenakis & M. Textor (Hrsg.), *Das Online-Familienhandbuch*. München.

Riley-Tillman, T., Chafouleas, S., Sassu, K., Chanese, J. & Glazer, A. (2008). Examining the Agreement of Direct Behavior Ratings and Systematic Direct Observation Data for On-Task and Disruptive Behavior. *Journal of Positive Behavior Interventions*, 10 (2), 136-143.

Riley-Tillman, T., Chafouleas, S., Christ, T., Briesch, A. & LeBel, T. (2009). The Impact of Item Wording and Behavioral Specificity on the Accuracy of Direct Behavior Ratings (DBRs). *School Psychology Quarterly*, 24 (1), 1–12.

Riley-Tillman, T., Christ, T., Chafouleas, S., Boice-Mallach, C. & 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.

Schlientz, M., Riley-Tillman, T., Briesch, A., Walcott, C. & Chafouleas, S. (2009). The Impact of Training on the Accuracy of Direct Behavior Ratings (DBR). *School Psychology Quarterly*, 24 (2), 73–83.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik. Mit 82 Tabellen.* (5. vollst. überarb. und erw. Aufl.). Berlin u.a.: Springer.

Schmischke, J. (2008a). Beobachtung. In R. Christiani & K. Metzger (Hrsg.), *Taschenlexikon Grundschulpraxis. 132 Beiträge zum Schulalltag. Pädagogik und Methodik. Deutsch und Mathematik* (1. Aufl.) (S. 24-25). Berlin: Cornelsen Scriptor.

Schmischke, J. (2008b). Unterrichtsstörungen. In R. Christiani & K. Metzger (Hrsg.), *Taschenlexikon Grundschulpraxis. 132 Beiträge zum Schulalltag. Pädagogik und Methodik. Deutsch und Mathematik* (1. Aufl.) (S. 192-193). Berlin: Cornelsen Scriptor.

Schmischke, J. (2008c). Verhaltensauffälligkeiten. In R. Christiani & K. Metzger (Hrsg.), *Taschenlexikon Grundschulpraxis. 132 Beiträge zum Schulalltag. Pädagogik und Methodik. Deutsch und Mathematik* (1. Aufl.) (S. 196-197). Berlin: Cornelsen Scriptor.

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

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2000). *Empfehlungen zum Förderschwerpunkt emotionale und soziale Entwicklung. Beschluss der Kultusministerkonferenz vom 10.03.2000*.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2016). *Sonderpädagogische Förderung in Schulen. 2005-2014*. Berlin.

Stein, R. (2017). Angst und sozial unsicheres Verhalten. In A. Methner, B. Seebach & K. Popp (Hrsg.): *Verhaltensprobleme in der Sekundarstufe. Unterricht - Förderung - Intervention* (S. 150–166). Stuttgart: Kohlhammer.

Stein, R. & Müller, T. (2015). Verhaltensstörungen und emotional-soziale Entwicklung: Zum Gegenstand. In R. Stein, T. Müller, E. Fischer, U. Heimlich, J. Kahlert & R. Lelgemann (Hrsg.), *Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung* (1. Aufl.) (S. 19–43). Stuttgart: Kohlhammer.

Strengths and Difficulties Questionnaire (2015). *Information for Researchers and Professionals about the Strengths & Difficulties Questionnaires.* Verfügbar unter http://www.sdqinfo.com/, zuletzt geprüft am 26.06.2018.

Tenorth, H.-E. & Tippelt, R. (Hrsg.). (2007). *Beltz Lexikon Pädagogik.* Weinheim: Beltz.

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

Wirtz, M. (Hrsg.) (2006). *Dorsch - Lexikon der Psychologie* (18. überarb. Aufl.). Bern: Hogrefe.