technische universität dortmund

fakultät rehabilitationswissenschaften

Entwicklung und Erforschung inklusiver Bildungsprozesse

Masterarbeit

Ist das schon auffällig?

Bestimmung kriterialer Bezugsnormen für den Einsatz des verhaltensverlaufsdiagnostischen Instruments DBR-MIS in der schulischen Praxis

vorgelegt von

Franziska Weniger
Steiler Weg 8
58453 Witten
Franziska.weniger@tu-dortmund.de
Master Lehramt sonderpädagogische Förderung LABG 2009
Matrikelnr. 176280

Erstgutachter: Prof. Dr. Michael Schurig
Zweitgutachter: Dr. Johannes Hasselhorn

Ausgabedatum: 10.06.2020
Abgabedatum: 29.07.2020

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIC	Aikake-Information-Criterion
BIC	Bayersan-Information-Criterion
bzw.	beziehungsweise
CMB	Curriculum Based Measurements
CFI	Comparative Fit Index
DAV	Depressives und ängstliches Verhalten
DAV07	Item „Wirkt besorgt, betrübt oder bedrückt“
DAV08	Item „Wirkt ängstlich/ fürchtet sich“
DAV09	Item „Wirkt nervös (sucht Nähe zu Erwachsenen)“
DBR	Direct Behavior Rating(s)
DBR-MIS	Direct Behavior Rating–Multi Item Scale
d. h.	das heißt
DRC	Daily Report Cards
DVB	Direkte Verhaltensbeobachtung
ebd.	ebenda
EP	Emotionale Probleme
EXT	Externalisierendes Verhalten
FS ESE	Förderschwerpunkt emotionale und soziale Entwicklung
GT	Generalisierbarkeitstheorie
i. d. R.	In der Regel
INT	Internalisierendes Verhalten
KMK	Kultusministerkonferenz
KTT	Klassische Testtheorie
LEVUMI	**Le**rn**v**erla**u**fs-**M**on**i**toring
MIS	Multi-Item Skalen
ML	Maximum-Likelihood (Schätzmethode)
PS	Prosoziales Verhalten
PS17	Item „Verhält sich anderen gegenüber rücksichtsvoll“
PS18	Item „Verhält sich anderen gegenüber hilfsbereit“
PS19	Item „Verhält sich in Partner- und Gruppensituationen kooperativ“
PSI	Probleme in sozialen Interaktionen

PSI10	Item „Arbeitet/spielt meist alleine“
PSI11	Item „Wird von Mitschüler_innen gehänselt oder geärgert, lässt sich provozieren“
PSI12	Item „Arbeitet/spielt häufiger mit Erwachsenen als mit Mitschüler_innen“
PSV	Positives Schulverhalten
RMSEA	Root Mean Square Error of Approximation
RTI	Response to Intervention
SAV	Störendes und auflehnendes Verhalten
SAV01	Item „Verhält sich wütend und aufbrausend“
SAV02	Item „Missachtet Regeln und hört nicht auf die Lehrkraft“
SAV03	Item „Streitet sich mit Mitschüler_innen/provoziert durch eigenes Verhalten seine Mitschüler_innen“
SDQ	Strenghts and Difficulties Questionnaire
SIS	Single-Item Skala
SuS	Schülerinnen und Schüler
SV	Schulbezogenes Verhalten
SV13	Item „Meldet sich im Unterricht“
SV14	Item „Hält sich an Gesprächsregeln“
SV15	Item „Richtet Aufmerksamkeit/Konzentration auf die Bearbeitung der Aufgabe“
SV16	Item „Arbeitet ruhig am Platz und verweigert nicht die Mitarbeit“
TLI	Tucker-Lewis Index
VP	Verhaltensprobleme
VPG	Verhaltensprobleme mit Gleichaltrigen
VPL	Verhaltensprobleme beim Lernen
VPL04	Item „Zappelt, ist (motorisch) unruhig/ überaktiv“
VPL05	Item „Bricht Aufgaben häufig früh ab“
VPL06	Item „Lässt sich schnell und leicht ablenken“
z. B.	zum Beispiel

1. Einleitung

Der Umgang mit verhaltensauffälligen oder -schwierigen Kindern und Jugendlichen stellt in deutschen Schulen einen besonderen Brennpunkt dar, der durch die zunehmende Heterogenität in den Klassenzimmern im Rahmen der Inklusion verstärkt in den Fokus schulpädagogischer Diskurse gerückt ist (Stein & Müller, 2015a). Das Verhalten von Schülerinnen und Schülern (SuS) hat in schulischen Kontexten nicht nur einen Einfluss auf das allgemeine Klassen– oder Schulklima, sondern ist zudem von großer Bedeutung für den individuellen Lernerfolg und die sozialen Beziehungen eines Kindes oder Jugendlichen (Jantzer et al., 2012). Schwierigkeiten und Auffälligkeiten im Verhalten können dahingehend das schulische Lernen von SuS stark beeinträchtigen und außerdem Anzeichen für dahinterliegende verhaltensbezogene oder psychische Störungen sein (Haller et al., 2016). Um dem Fortschreiten möglicher Störungsentwicklungen entgegenzuwirken, sind eine frühzeitige Erkennung von Verhaltensproblemen sowie individuelle Fördermaßnahmen in schulischen Kontexten von großer Bedeutung (Hillenbrand, 2008). Diese Identifikation von Verhaltensauffälligkeiten und Planung von Fördermaßnahmen erfolgt in der schulischen Praxis auf der Basis diagnostischer Prozesse. Bislang vorherrschend sind dabei statusdiagnostische Prozesse, die der Erhebung eines Ist-Zustands dienen. Im Rahmen der Umsetzung des Inklusionsgedankens steht diese Form der Diagnostik allerdings in der Kritik, da im Hinblick auf das Ziel einer optimalen Förderung aller SuS eine Diagnostik gefordert wird, die eine stetige Evaluation der Wirkung von Fördermaßnahmen ermöglicht. Dieser Anforderung entsprechen Verfahren und Ansätze der Verlaufsdiagnostik, welche eine Überprüfung von Fördermaßnahmen anhand der Erhebung individueller Entwicklungsverläufe vorsehen (Hartmann, 2017). Einen konkreten Ansatz für die Verlaufsdiagnostik im Bereich Verhalten stellt die aus dem englischsprachigen Raum stammende Methode des Direct Behavior Rating (DBR) dar. Da die Forschung zu dieser Methode im deutschsprachigen Raum noch relativ jung ist, gibt es bisher nur wenige, bereits evaluierte Instrumente (Casale, Hennemann, Huber & Grosche, 2015b). Ein konkretes Instrument, das im Rahmen des Forschungsprojekts LEVUMI (Lernverlaufs-Monitoring) entwickelt wurde und auf einer gleichnamigen Onlineplattform (www.levumi.de) für Lehrkräfte angeboten wird, ist der DBR-MIS (Direct Behavior Rating-Multi Item Scale). Diese Ratingskala stellt ein „multidimensionales

Verfahren zur direkten Verhaltensbewertung dar“ (Schurig, Jungjohann & Gebhardt, 2019, S. 4), welches bereits in einer ersten Pilotierungsstudie von Hisker (2018) erprobt und bezüglich verschiedener Gütekriterien überprüft wurde. Auch wenn der DBR-MIS bereits auf der Onlineplattform veröffentlicht wurde, besteht hinsichtlich dessen Weiterentwicklung, Verbesserung und Etablierung in der schulischen Praxis weiterhin Forschungsbedarf. Die vorliegende Arbeit knüpft daher an die bisherigen Forschungsarbeiten zum DBR-MIS im Rahmen des LEVUMI Forschungsprojekts an und hat zum Ziel, durch die Bestimmung kriterialer Bezugsnormen in Form von Schwellenwerten den diagnostischen Wert des Instruments für die schulische Praxis zu steigern.

Insgesamt umfasst die vorliegende Arbeit sieben inhaltliche Kapitel. Anschließend an diese Einleitung erfolgt zunächst die Erläuterung des theoretischen Hintergrunds der Arbeit (Kapitel 2). Dabei werden grundlegende Theorien, Begrifflichkeiten und Forschungsbefunde, die den Rahmen für die nachfolgenden Analysen und Untersuchungen bilden, vorgestellt. In Anlehnung an die Ausführungen des theoretischen Hintergrunds erfolgt dann die Konkretisierung der Zielsetzung dieser Arbeit sowie die Ableitung von Forschungsfragen und darauf bezogenen Hypothesen (Kapitel 3). Das methodische Vorgehen für die Untersuchungen im Rahmen der vorliegenden Arbeit wird in Kapitel 4 beschrieben. In den darauffolgenden zwei Kapiteln werden die Ergebnisse der Untersuchungen zunächst dargestellt (Kapitel 5) und anschließend in Rückbezug zu den Forschungsfragen sowie dem methodischen Vorgehen kritisch diskutiert (Kapitel 6). Abschließend wird ein zusammenfassendes Fazit formuliert und ein Ausblick auf weiterführende Forschungsmöglichkeiten gegeben (Kapitel 7).

2. Theoretischer Hintergrund

Das Forschungsinteresse dieser Arbeit bezieht sich auf ein spezifisches Instrument der schulischen Verhaltensverlaufsdiagnostik. Bevor jedoch konkrete Zielsetzungen und Fragestellungen vorgestellt werden können, ist es zunächst notwendig, die Arbeit in einen theoretischen Kontext einzuordnen. Dafür werden die Entwicklungslinien und -zusammenhänge des fokussierten Instruments in den folgenden Unterkapiteln beschrieben. Das Instrument ist einer Methode der Verhaltensdiagnostik, dem DBR, zuzuordnen. Da diese Methode im Rahmen des Inklusionsbestrebens im deutschsprachigen Raum an Bedeutung gewonnen hat, erfolgt zuerst eine kurze Einführung in den Inklusionsbegriff und inklusionsinduzierte Veränderungen im schulischen Kontext (Kapitel 2.1). Anschließend werden die Bereiche Verhalten und Verhaltensstörung als schulisch relevante Dimensionen betrachtet (Kapitel 2.2 & 2.2.1). Im Zuge dessen wird zusätzlich ein Überblick über Prävalenzen von Verhaltensstörungen und psychischen Auffälligkeiten im Kindes- und Jugendalter gegeben (Kapitel 2.2.2). In Kapitel 2.3 wird dann das Konzept der pädagogischen Verlaufsdiagnostik mit Fokus auf die Bereiche der Lern- und der Verhaltensverlaufsdiagnostik (Kapitel 2.3.1 & 2.3.2) beschrieben. Den Abschluss des theoretischen Rahmens bilden die Vorstellung der grundlegenden Konzeption des DBR (Kapitel 2.4.1) sowie die Darlegung des aktuellen Forschungsstands zu dieser Methode (Kapitel 2.4.2).

2.1 Heterogenität und Inklusion

Unter dem Begriff Heterogenität wird allgemeinhin die „Verschiedenheit, Ungleichartigkeit oder Andersartigkeit bezogen auf Individuen, Gruppen oder pädagogische Organisationen“ (Walgenbach, 2014, S. 13) verstanden. In der Schulpädagogik liegt der Fokus dabei auf der Heterogenität der Schülerschaft hinsichtlich verschiedener Dimensionen. Diese Heterogenitätsdimensionen umfassen sowohl zentrale Strukturmerkmale wie die Differenzlinien Nationalität, Alter, Geschlecht, Sprache, soziale Herkunft und Behinderung als auch individuelle Besonderheiten wie Interessen, Bedürfnisse und Leistungen (Rendtorff, 2014). Bei Schuleintritt bestehen bereits große Unterschiede in den Entwicklungsständen und Voraussetzungen der Kinder, die sich im Laufe der Schulzeit oftmals manifestieren oder vergrößern (Schuchardt, 2015). Eine im Bildungssystem verankerte und lange vorherrschende Strategie im Umgang mit dieser Heterogenität, ist die Homogenisierung der Schülerschaft. Diese Strategie

umfasst Maßnahmen, nach denen SuS hinsichtlich verschiedener Aspekte wie ihrem Leistungs- oder Entwicklungsstand selektiv geordnet werden, um so möglichst homogene Lerngruppen zu erhalten. Beispiele für diese Maßnahmen sind die Einteilung in verschiedene Schulformen und Schulklassen sowie auch die Sonder- bzw. Förderbeschulung (Kutscher, 2008). Dieser Umgang mit Heterogenität in der Schule steht seit der Veröffentlichung verschiedener Bildungsberichte in bildungspolitischen und erziehungswissenschaftlichen Diskursen besonders in der Kritik. So zeigten die Ergebnisse der ersten PISA-Studie 2002, dass das deutsche Schulsystem durch ein hohes Maß an sozialer Selektion geprägt ist und individuelle Lernvoraussetzungen von SuS oftmals nur mangelhaft berücksichtigt werden (Saalfrank & Zierer, 2017).

Im Zuge der Ratifizierung der UN-Behindertenrechtskonvention im Jahr 2009 wurde ein Umdenken in vielen gesellschaftlichen Bereichen wie auch dem Bildungssektor bezüglich des Umgangs mit Behinderung und Heterogenität angestoßen (Mähler & Krüger, 2015). Nach Artikel 24 der Konvention hat sich Deutschland dazu verpflichtet, sicherzustellen, dass Menschen mit Behinderung hinsichtlich der Chancengleichheit ohne Einschränkungen Zugang zum allgemeinen Bildungssystem haben (Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderung, 2017). Das Leitbild, das sich aus diesem Übereinkommen entwickelt hat, ist die „Inklusion“ (Hartke, 2017).

Im Bildungsbereich haben sich zwei Verständnisse des Ziels der Inklusion und dessen Umsetzung herausgebildet. Für Vertreter eines engen Inklusionsbegriffs ist die schulische Inklusion das Recht von Kindern mit Behinderung bzw. sonderpädagogischem Förderbedarf auf einen Platz im allgemeinen Schulsystem. Hinter dem gemäßigten oder auch weiten Inklusionsverständnis steht dagegen ein pädagogisches Konzept, das „neben dem Recht auf Teilhabe auch das Recht auf eine optimale Bildung und Förderung aller Kinder und Jugendlichen“ (Hartke, 2017, S. 13) vorsieht. Inklusionsziele wie Chancengleichheit, Teilhabe sowie Anerkennung und Wertschätzung von Vielfalt stehen in dem Konzept neben dem zentralen Anliegen, optimale Lern- und Förderangebote zu schaffen, und beziehen sich dabei nicht nur auf SuS mit Behinderungen, sondern auf alle Kinder mit Beeinträchtigungen oder anderen besonderen Bedürfnissen (ebd.). Das Inklusionskonzept fordert das Umdenken von einem bisher leistungs- und selektionsorientierten hin zu einem förder- und

präventionsorientierten Schulsystem, in dem die Schul- und Lernangebote an die Bedürfnisse der einzelnen SuS angepasst werden und nicht umgekehrt (Mähler & Krüger, 2015). Die Grundlage für die vorliegende Arbeit soll im Folgenden das zuletzt beschriebene gemäßigte Inklusionsverständnis bilden, welches die Zielvorstellung eines inklusiven Unterrichts für alle SuS umfasst (Hartke, 2017).

Die Umsetzung des Leitbildes Inklusion sowie der damit einhergehenden Anliegen und Ziele bedeutet für das deutsche Schulsystem umfassende Veränderungen und stellt die einzelnen Schulen und Lehrkräfte vor große Herausforderungen. In Deutschland gibt es für die Umsetzung der inklusiven Beschulung bundesweite Empfehlungen der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). Für die konkrete Regelung der Rahmenbedingungen und der Umsetzungsgestaltung sind allerdings die einzelnen Bundesländer selbst verantwortlich (Gebhardt, Sälzer & Tretter, 2014). Die Ergebnisse einer Studie von Gebhardt, Sälzer und Tretter (2014) zeigen, dass bundesweit inklusive Konzepte implementiert werden, sich die Umsetzungsmodelle zwischen den Bundesländern allerdings leicht voneinander unterscheiden. Trotz dieser Unterschiede haben sich in der schulpädagogischen Literatur und auch in der schulischen Praxis spezifische Lehr- und Lernformen für einen inklusiven Unterricht etabliert. Der Grundgedanke dieser inklusiven Lehr- und Lernformen ist, dass SuS mit und ohne offiziellen Förderbedarf in einem gemeinsamen Unterricht im Klassenverbund lernen. Entscheidend bei der Gestaltung dieses gemeinsamen Unterrichts im Sinne der Inklusion ist, dass die individuellen Entwicklungsstände und Bedürfnisse der einzelnen SuS berücksichtigt werden, um so optimale Lernbedingungen zu schaffen (Götz & Hauenschild, 2015).

Neben spezifischen unterrichtlichen Lehr- und Lernformen spielt auch die Etablierung von Förderkonzepten eine tragende Rolle in der Umsetzung schulischer Inklusion (Huber & Rietz, 2015). Ein konkretes Rahmenkonzept für die inklusive Beschulung stellt der sogenannte Response to Intervention (RTI)-Ansatz dar. Dieser hat die Prävention und Früherkennung von Lern- und Entwicklungsstörungen bei SuS zum Ziel und umfasst eine gestuftes Fördersystem, das auf einer Passung zwischen den individuellen Bedürfnissen eines Kindes, dem Unterricht sowie der Förderung basiert. Entscheidend bei diesem Förderansatz ist die kontinuierliche Lern- und

Entwicklungsverlaufsdokumentation, anhand der die jeweiligen Förderentscheidungen für ein Kind getroffen werden (Blumenthal & Hartke, 2015).

In einem inklusiven Unterricht sollten die Lern- und Entwicklungsstände eines Kindes sowie dessen Stärken und Schwächen für die Entscheidungen über Unterrichtsformen und Fördermaßnahmen ausschlaggebend sein. Für Lehrkräfte ist es daher wichtig, Kenntnisse über verschiedene Lern- und Entwicklungsbereiche von Kindern zu haben, die für deren weitere schulische und auch persönliche Entwicklung von Bedeutung sind (Schuchardt, 2015). In der allgemeinen Entwicklungspsychologie wird zwischen sechs zentralen Entwicklungsbereichen eines Kindes unterschieden: Kognition, Emotionalität, Kommunikation/ Sprache, Motorik, Wahrnehmung und Soziabilität (Oerter, 2002). Zusätzlich zu diesen Entwicklungsbereichen benennen Flott-Tönjes et al. (2017) das Lern- und Arbeitsverhalten sowie die Lebensgestaltung eines Kindes als zwei sonderpädagogisch relevante Förderbereiche in schulischen Settings. In der Sonderpädagogik werden für diese Entwicklungs- und Förderbereiche in den übergeordneten Förderschwerpunkten Lernen, emotionale und soziale Entwicklung, Sprache, körperliche und motorische Entwicklung, Sehen, Hören und Kommunikation sowie geistige Entwicklung spezifische Förderteilbereiche formuliert (Flott-Tönjes et al., 2017). Diese Anzahl an unterschiedlichen Schwerpunkten und Bereichen zeigt die Vielfalt der möglichen Förderbedürfnisse von SuS an.

Im Schuljahr 2016/ 2017 wurde bei 7,1 % der SuS in Deutschland ein Förderbedarf diagnostiziert, von denen etwa 39% inklusiv beschult wurden. Insgesamt nimmt die Förderquote in den letzten Jahren leicht zu, wobei es Unterschiede in den jeweiligen Förderschwerpunkten gibt. Insbesondere im Förderschwerpunkt emotionale und soziale Entwicklung (FS ESE) ist die Schüleranzahl steigend und hat sich von 2000 bis 2016 verdoppelt (Autorengruppe Bildungsberichterstattung, 2018). Zu der Schülergruppe, die mit dem Förderbedarf Emotionale und soziale Entwicklung in Verbindung gebracht werden, gehören oftmals Kinder mit auffälligem und herausforderndem Verhalten (Stein & Müller, 2015b). Die ansteigenden Zahlen in diesem Förderschwerpunkt verdeutlichen, dass der Umgang mit Verhaltensauffälligkeiten und -störungen einen besonderen Brennpunkt in den deutschen Schulen darstellt und in Hinblick auf Prävention und Förderung effektiver Instrumente sowie Konzepte bedarf (Stein & Müller, 2015a). Im nachfolgenden Kapitel werden daher die Begriffe Verhalten und

Verhaltensstörungen mit Fokus auf schulrelevante Verhaltensweisen näher erläutert sowie ein Überblick über die Prävalenzen von Verhaltensstörungen bei Kindern und Jugendlichen gegeben.

2.2 Verhalten und Verhaltensstörung

Der Begriff Verhalten wird in der Pädagogik als allgemeine Bezeichnung für alle Aktivitäten von lebenden Organismen, die messbar oder beobachtbar sind, verwendet. Dabei wird zwischen offenem und verdecktem Verhalten unterschieden. Offen bedeutet an dieser Stelle, dass das Verhalten von außen direkt beobachtbar ist, während verdecktes Verhalten physiologische Veränderungen umfasst, die oftmals nur mithilfe gewisser Messverfahren erfasst werden können (Tenorth & Tippelt, 2012).

Menschliches Verhalten vollzieht sich in sozialen Systemen und stellt daher in der Regel eine Reaktion auf verschiedene Reize der Umwelt dar (Tschira, 2005). In verschiedenen Bezugssystemen gelten bestimmte gesellschaftliche Normen und Werte. Die Bewegung in diesen Systemen und Settings wie zum Beispiel der Schule oder der eigenen Familie ist daher mit spezifischen Erwartungen an Verhaltens- und Interaktionsweisen verbunden. Daher hängt die Frage, „wie sich Menschen verhalten und wie andere darauf reagieren, […] im Wesentlichen von den Handlungszusammenhängen ab, in die beide eingeschaltet sind“ (Tschira, 2005, S. 39). Weicht das Verhalten von Kindern und Jugendlichen von geltenden gesellschaftlichen Normen und Werten sowie den damit verbundenen Verhaltenserwartungen ab, wird häufig von Verhaltensauffälligkeiten oder Verhaltensstörungen gesprochen. Die beiden Begriffe werden sowohl in der Literatur als auch in der Praxis oftmals synonym verwendet und verstanden (Myschker & Stein, 2018). Für Vertreter einer interaktionistischen Perspektive, die Verhaltensstörungen als Funktionsgleichgewichtsstörung im Person-Umwelt-Bezug betrachten, stellen Verhaltensauffälligkeiten bei Kindern und Jugendlichen Symptome einer dahinterliegenden Verhaltensstörung dar (Seitz & Stein, 2010). Der Begriff Verhaltensauffälligkeit gilt als wertneutral, steht allerdings aufgrund seiner Unschärfe und Mehrdeutigkeit in Bezug auf Verhaltensweisen in der Kritik. Daher hat sich in wissenschaftlichen und administrativen Bereichen insbesondere der Begriff der Verhaltensstörung durchgesetzt (Myschker & Stein, 2018). Eine in der

deutschen Pädagogik anerkannte Definition zu diesem Begriff liefert Myschker (2002):

> *„Verhaltensstörung ist ein von den zeit- und kulturspezifischen Erwartungsnormen abweichendes maladaptives Verhalten, das organogen und/oder milieureaktiv bedingt ist, wegen der Mehrdimensionalität, der Häufigkeit und des Schweregrades die Entwicklungs-, Lern- und Arbeitsfähigkeit sowie das Interaktionsgeschehen in der Umwelt beeinträchtigt und ohne besondere pädagogische-therapeutische Hilfe nicht oder nur unzureichend überwunden werden kann."* (Myschker, 2002, S. 44)

Nach dieser Definition umfasst eine Verhaltensstörung überdauerndes fehlangepasstes Verhalten, das in verschiedenen Bereichen problematische Auswirkungen mit sich zieht. Dieses Verhalten ist multifaktoriell bedingt und bedarf spezifischer pädagogischer und therapeutischer Maßnahmen. Der Begriff Verhaltensstörung ist damit ein Oberbegriff unter den sich zahlreiche Phänomene und Erscheinungsformen zusammenfassen lassen (Myschker & Stein, 2018). Ein weiterer insbesondere in schulischen Kontexten bedeutender Begriff ist der „Förderschwerpunkt emotionale und soziale Entwicklung" (FS ESE). Der Begriff wurde von der Kultusministerkonferenz (KMK) eingeführt und umschreibt ein Problemfeld, das den Bereich Verhaltensstörungen umfasst:

> *„Sonderpädagogischer Förderbedarf ist bei Kindern und Jugendlichen mit Beeinträchtigungen der emotionalen und sozialen Entwicklung, des Erlebens und der Selbststeuerung anzunehmen, wenn sie in ihren Bildungs-, Lern- und Entwicklungsmöglichkeiten so eingeschränkt sind, dass sie im Unterricht der allgemeinen Schule auch mit Hilfe anderer Dienste nicht hinreichend gefördert werden können."* (KMK, 2000, S. 10)

Ein entscheidender Punkt ist, dass in dieser Umschreibung sowohl Verhaltens- als auch Erlebniskomponenten berücksichtigt werden (Wember, Stein & Heimlich, 2014). Der Begriff Verhaltensstörung umfasst damit nicht nur manifestes Verhalten, sondern auch das psychische Verhalten von Kindern und Jugendlichen und wird dahingehend teilweise durch den Terminus Gefühls- oder Erlebnisstörung erweitert (Seitz & Stein, 2010). Zudem wird deutlich, dass sich die Zuweisung eines sonderpädagogischen Förderbedarfs an den Kriterien des allgemeinbildenden Schulsystems orientiert (Hillenbrand, 2008).

Die Schule stellt ein komplexes soziales Bezugssystem dar, in dessen Rahmen sich für verschiedene schulische Settings bestimmte Verhaltenserwartungen an SuS ergeben. Ob das Verhalten eines Schülers oder einer Schülerin als abweichend oder störend angesehen wird, ist daher abhängig von dem situativen Kontext und der subjektiven Beurteilung einer Bezugsperson. So können Verhaltensweisen in einem schulischen Setting als abweichend empfunden werden, während sie in einem anderen sozialen Rahmen akzeptiert werden (Hillenbrand, 2008). Dementsprechend ist es bei Verhaltensbeurteilungen in der Schule notwendig, die situativen Kontexte, in denen das Verhalten auftritt, und die jeweiligen Bezugsnormen zu berücksichtigen (Tschira, 2005). Zudem sollte sich die Beurteilung an bestimmten Kriterien für auffälliges oder abweichendes Verhalten orientieren. In Anlehnung an die Definition von Myschker (2002) wird Verhalten in schulischen Kontexten als abweichend bezeichnet, wenn es nicht alters- und entwicklungsstandgemäß ist, auf inakzeptable Weise gegen die sozialen Normen eines Settings verstößt, langfristig in verschiedenen Situationen auftritt, die soziale Kontaktaufnahme sowie Lern- und Entwicklungsprozesse eines Kindes verhindert oder einschränkt und außerdem das soziale Umfeld, andere SuS oder den Unterricht beeinträchtig (Schmischke, 2008). Verhaltensstörungen zeigen sich in einer Vielzahl von Erscheinungsformen, die sich wiederum in individuell sehr unterschiedlichen Verhaltensweisen von SuS äußern. Damit stellen Kinder und Jugendliche mit Verhaltensstörungen eine sehr heterogene Gruppe dar, die durch die empirische Klassifikation der Erscheinungsformen in Untergruppen eingeteilt werden kann (Myschker & Stein, 2018).

2.2.1 Störungsklassen und Verhaltensbereiche

Auf Basis der bisherigen empirischen Forschung können nach Myschker und Stein (2018) vier Klassen von Verhaltensstörungen unterschieden werden. Insbesondere die Abgrenzung von externalisierenden, aggressiv-ausagierenden und internalisierenden, ängstlich-gehemmten Störungsklassen ist empirisch fundiert, wohingegen die zwei weiteren Klassen, die zum einen sozial-unreifes und zum anderen sozialisiert-delinquentes Verhalten umfassen, weniger gut belegt sind (Myschker & Stein, 2018).

Zu den externalisierenden Verhaltensstörungen zählen nach außen gerichtete Verhaltensweisen wie Aggressivität, Impulsivität, Aufmerksamkeitsstörungen und Hyperaktivität, die primär problematische Auswirkungen auf das Umfeld nach sich ziehen (Wember et al., 2014). Internalisierende Verhaltensstörungen umfassen innen gerichtete Verhaltensweisen, die nach außen hin weniger auffällig sind, aber für die betreffende Person eine hohe Belastung und Gefährdung darstellen. Dazu gehören Störungsbilder wie Ängstlichkeit, Depressivität und Minderwertigkeitsgefühle (Myschker & Stein, 2018). Die Störungsbilder beider Formklassen sind etwa gleich weit verbreitet, wobei geschlechterspezifische Unterschiede auftreten. So zeigen Jungen öfter externalisierende Verhaltensstörungen, während Mädchen eher von internalisierenden Störungen betroffen sind (Hillenbrand, 2008). Unter sozial-unreifem Verhalten werden altersunangemessene Verhaltensweisen sowie Konzentrations- und Leistungsschwächen gefasst. Verhaltensweisen wie Gewalttätigkeiten, Diebstähle und Regelmissachtungen fallen unter die Klasse der sozialisiert-delinquenten Verhaltensstörungen (Myschker & Stein, 2018).

Das Verhalten von Kindern und Jugendlichen ist im schulischen Bereich für individuelle Lernerfolge, das allgemeine Schul- und Klassenklima sowie für interpersonelle Beziehungen von Bedeutung (Jantzer et al., 2012). In Hinblick auf den Lernerfolg zeigten unterschiedliche Studien, dass Verhaltensschwierigkeiten oftmals gemeinsam mit Lernproblemen oder -störungen auftreten. So kann sich eine Verhaltensstörung wie Hyperaktivität ungünstig auf Lernprozesse auswirken und zu Lernproblemen führen. Andererseits können auch Lernschwierigkeiten und damit einhergehende Misserfolgserfahrungen zur Entwicklung verschiedener Verhaltensauffälligkeiten oder -störungen beitragen (Linderkamp & Grünke, 2007). Häufig ist es allerdings nicht möglich herauszufinden, „welche Störung am Beginn des Fehlentwicklungs-Prozesses stand, oder ob sich nicht beide Störungen in einem gemeinsamen Prozess manifestierten“ (Myschker & Stein, 2018, S. 74).

Für die Untersuchung der Verbreitung von Verhaltensstörungen oder auch psychischen Störungen sowie die Erforschung schulrelevanten Verhaltens werden in verschiedenen Studien bestimmte Verhaltensbereiche unterschieden. Diese spiegeln größtenteils die beschriebenen Störungsklassen nach Myschker und Stein (2018) wider, werden aber durch explizit schulrelevante Verhaltensbereiche erweitert. In der

deutschen Version des Strengths and Difficulties Questionnaire (SDQ; Stärken und Schwächen Fragebogen), der als Erfassungsinstrument für psychosoziale Auffälligkeiten im Kindes- und Jugendalter genutzt wird, werden fünf Verhaltensbereiche unterschieden: die vier Problembereiche emotionale Probleme, Hyperaktivität, Verhaltensprobleme und Probleme mit Gleichaltrigen sowie der Stärkebereich prosoziales Verhalten (Altendorfer-Kling, Ardelt-Gattinger & Thun-Hohenstein, 2007). Der Bereich emotionale Probleme lässt sich der Störungsklasse internalisierende Verhaltensstörungen zuordnen und umfasst damit Angststörungen und Depressivität. Ein starkes Angsterleben zeigt sich häufig in körperlichen Symptomen wie Nervosität, Anspannung und Schwitzen. Depressive Störungen äußern sich dagegen in Freudlosigkeit, Antriebsminderung und auch psychosomatischen Beschwerden (Goodman, Lamping & Ploubidis, 2010; Linderkamp & Grünke, 2007). Die Verhaltensbereiche Hyperaktivität und Verhaltensprobleme beziehen sich auf Erscheinungsformen der externalisierenden Verhaltensstörungen. Verhaltensprobleme umfassen dabei oppositionelle, aggressive und regelmissachtende Verhaltensweisen, die sich im schulischen Bereich gegen andere SuS, Lehrkräfte, die Schuleinrichtung oder auch Schulregeln richten. Der Bereich Hyperaktivität schließt Symptome wie Unaufmerksamkeit, körperliche Überaktivität und Impulsivität mit ein, die oftmals in Verbindung mit einer Aufmerksamkeitsdefizit-/ Hyperaktivitätsstörung auftreten (ebd.). Verhaltensweisen, die diesem Bereich zugeordnet werden, können die sozialen Kontakte eines betroffenen Kindes und den Unterricht in der Schule massiv beeinträchtigen (Wember et al., 2014). Verhaltensprobleme mit Gleichaltrigen zeigen sich in einem gestörten Kontaktverhalten von betroffenen Kindern zu ihren Mitschülerinnen und Mitschülern, das z. B. durch Ablehnung, Isolation und Mobbing geprägt ist. Hinter Verhaltensproblemen dieses Bereiches liegen sowohl externalisierende als auch internalisierende Störungsbilder (Goodman et al., 2010; Linderkamp & Grünke, 2007). In einer Studie von DeVries, Rathmann und Gebhardt (2018) konnten diese aufzeigen, dass Verhaltensprobleme mit Gleichaltrigen die Schulleistungen von betroffenen SuS negativ beeinflussen. Dagegen konnte für prosoziale Verhaltensweisen ein positiver Einfluss verzeichnet werden (DeVries et al., 2018). Damit ist der Bereich des prosozialen Verhaltens, der im Gegensatz zu den anderen Bereichen keinen Problem-, sondern einen Stärkebereich darstellt, für schulische Kontexte relevant. Prosoziale

Verhaltensweisen wie freiwilliges Helfen, Kooperieren und Unterstützen werden mit sozialen Kompetenzen in Verbindung gebracht und stehen im Gegensatz zu aggressivem und selbstbezogenem Verhalten (Bierhoff, 2002). Ein weiterer Verhaltensbereich, der im SDQ nicht explizit auftaucht, aber in anderen Erhebungsinstrumenten Beachtung findet, ist das Lern- und Arbeitsverhalten. Dieser Bereich bezieht sich konkret auf schul- und lernbezogene Verhaltensweisen wie die Bearbeitung von Aufgaben, die Beteiligung am Unterricht oder auch die Befolgung von Anweisungen (Casale, Hennemann, Volpe, Briesch & Grosche, 2015c).

2.2.2 Prävalenzen von Verhaltensstörungen und psychischen Auffälligkeiten

Die Prävalenz von kinder- und jugendpsychiatrischen Auffälligkeiten wird seit vielen Jahren in umfassenden internationalen und nationalen Studien untersucht. Die Ergebnisse der verschiedenen Studien weisen unterschiedliche Prävalenzraten von Kindern und Jugendlichen mit psychischen Auffälligkeiten oder Verhaltensstörungen aus. Die hohe Variabilität in den Verbreitungs- und Häufigkeitsanagaben der Studien ist durch ihre methodische Heterogenität zu erklären. Unterschiede ergeben sich zum Beispiel durch die Nutzung verschiedener Erhebungsinstrumente, die Bezugnahme auf unterschiedliche Altersgruppen sowie durch grundlegend differierende Störungsklassifikationen. Während international bereits zahlreiche Prävalenzstudien durchgeführt wurden, ist die Anzahl nationaler und bevölkerungsrepräsentativer Erhebungen relativ gering (Hölling, Schlack, Petermann, Ravenz-Sieberer & Mauz, 2014). Im Folgenden werden einige Ergebnisse von Studien zur Prävalenz von psychischen und verhaltensbezogenen Auffälligkeiten und Störungen im Kindes- und Jugendalter zusammenfassend vorgestellt.

Angesichts der hohen Heterogenität internationaler und nationaler Studien haben Ihle und Esser (2002) sowie Barkmann und Schulte-Markwort (2004) in Übersichtsarbeiten die Prävalenzangaben verschiedener Studien miteinander verglichen. Trotz Unterschieden bezüglich darin berücksichtigter Studien und der Fokussetzung, ähneln sich die ermittelten Prozentwerte der durchschnittlichen Gesamtprävalenz in beiden Übersichtsarbeiten stark (Kowalewski, 2009). In einem Vergleich von 19 internationalen Studien berechneten Ihle und Esser (2002) für die durchschnittliche Gesamtprävalenz einen Mittelwert von 18%, wobei also etwa dreiviertel der in den

berücksichtigten Studien angegeben Prävalenzraten zwischen 15 und 22% lagen. Für die Ermittlung der Prävalenz spezifischer psychischer Störungen im Kindes- und Jugendalter, darunter depressive Störungen, Angststörungen, hyperkinetische Störungen und dissoziale Störungen, konnten Daten aus insgesamt 15 der 19 Studien verwendet werden. Bei den depressiven Störungen und den Angststörungen, die beide zur Klasse der internalisierenden Verhaltensstörungen zählen, ergaben sich Prävalenzen von durchschnittlich 4,4% und 10,4%. Die dissozialen Störungen bildeten mit einer durchschnittlichen Prävalenz von 7,5% den zweithäufigsten Störungsbereich, während sich für hyperkinetische Störungen mit 4,4% eine ähnliche Prävalenzrate wie bei depressiven Störungen ergab. In Hinblick auf die altersspezifische Verteilung von psychischen Störungen zeigten sich für die eingegrenzte Altersgruppe bis 13 Jahren in den meisten Bereichen nur leicht geringere Prävalenzraten. Eine deutliche Veränderung konnte allerdings für die internalisierenden Störungen festgestellt werden. So verringerte sich die durchschnittliche Prävalenzrate von Angststörungen auf 6,5% und von depressiven Störungen auf 1,5%. Auch bezogen auf geschlechtsspezifische Unterschiede in den Prävalenzen spielt das Alter eine Rolle (Ihle & Esser, 2002). So weisen die Befunde von Ihle und Esser (2002) darauf hin, dass Jungen in der Kindheit insgesamt häufiger psychische und verhaltensbezogene Auffälligkeiten zeigen als Mädchen. Im Jugendalter gleichen sich die durchschnittlichen Gesamtprävalenzraten beider Geschlechter in etwa an. Für Jungen konnten dabei durchgehend höhere Raten für externalisierende Störungen und für Mädchen insbesondere im späten Jugendalter höhere Raten für internalisierende Störungen beobachtet werden (ebd.).

Barkmann und Schulte-Markwort (2004) berücksichtigten in ihrer Übersichtsarbeit 29 nationale epidemiologische Studien und verglichen deren Angaben zur Prävalenz psychischer Auffälligkeiten im Kindes- und Jugendalter in Deutschland. Die Prävalenzangaben der verschiedenen Studien lagen zwischen 10,3% und 29,9%, woraus die Autoren für die durchschnittliche Gesamtprävalenz einen Mittelwert von 17,2% ermittelten. Analysen bezüglich unterschiedlicher Gesamtprävalenzen in spezifischen Altersgruppen zeigten eine generelle Zunahme psychischer Auffälligkeiten von der frühen Kindheit bis zum späten Jugendalter (Barkmann & Schulte-Markwort, 2004). Trotz nationalem Fokus und unterschiedlichem methodischen Vorgehen liegen die

Ergebnisse der Übersichtsarbeit von Barkmann und Schulte-Markwort (2004) insbesondere mit Blick auf die Angabe zur durchschnittlichen Gesamtprävalenz psychischer Auffälligkeiten nah an den Ergebnissen von Ihle und Esser (2002).

Eine aktuelle und sehr umfassende Studie zur Prävalenz von psychischen Auffälligkeiten und psychosozialen Beeinträchtigungen bei Kindern und Jugendlichen in Deutschland ist die KiGGS-Studie (Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland). Die KiGGS-Studie ist eine Langzeitstudie des Robert-Kochs-Instituts mit dem Ziel der Erfassung von Prävalenzen und zeitlichen Trends über mehrere Erhebungszeiträume. Zentrales Erhebungsinstrument ist dabei die Elternversion des oben beschriebenen SDQ für drei- bis siebzehnjährige Kinder und Jugendliche. Bei der KiGGS-Basiserhebung (2003-2006) wurden anhand deutscher Normwerte für den SDQ etwa 20% der Kinder und Jugendlichen als psychisch auffällig identifiziert. Über die erste Folgeerhebung in der KiGGS-Welle 1 (2009-2012) ist diese Prävalenzrate unverändert geblieben (Hölling et al., 2014). In der KiGGS-Welle 2 (2014-2017) wurden die Daten von 13.205 Kindern und Jugendlichen (6.637 Mädchen, 6.568 Jungen) analysiert und darauf basierend Prävalenzen mit Berücksichtigung der Faktoren Geschlecht, Alter und sozioökonomischer Status der Familie berechnet. Für den Zeitraum von 2014 bis 2017 betrug die Prävalenzrate von psychischen Auffälligkeiten bei Kindern und Jugendlichen nach den Ergebnissen der KiGGS-Welle 2 insgesamt etwa 16,9%, verringerte sich also um circa 3% im Vergleich zu den Ergebnissen der Basis- und der ersten Folgeerhebung. Mit 19,1% konnte bei Jungen durchschnittlich eine höhere Prävalenz von psychischen Auffälligkeiten festgestellt werden als bei Mädchen mit 14,5%. In der Altersgruppe von 15 bis 17 Jahren nähert sich die Häufigkeit psychischer Auffälligkeiten zwischen den Geschlechtern an. Eine kontinuierliche Ab- oder Zunahme der Prävalenzen über die Altersgruppen zeigte sich jedoch nicht. Zudem waren Kinder und Jugendliche aus sozioökonomisch schwachen Familien signifikant häufiger von psychischen Auffälligkeiten betroffen als Kinder und Jugendliche mit hohen sozioökonomischen Status (Klipker, Baumgarten, Göbel, Lampert & Hölling, 2018). Informationen über die Ausprägung von Auffälligkeiten in den einzelnen Verhaltensbereichen können in den Ergebnissen der KiGGS-Welle 1 (2009-2012) gefunden werden. Diese zeigen, dass Hyperaktivitätsprobleme insbesondere bis zum Altersbereich der sieben- bis dreizehnjährigen Kinder am stärksten

ausgeprägt waren, während die insgesamt geringste Ausprägung für den Bereich der Peer-Probleme festgestellt werden konnte. Verhaltensprobleme und emotionale Probleme gingen mit vergleichbaren Ausprägungswerten im ebenfalls niedrigen Bereich einher (Hölling et al., 2014). Bei der Betrachtung und Einordnung der Ergebnisse der KiGGS-Studie ist zu berücksichtigen, dass die Studie nicht den Anspruch auf eine konkrete Identifizierung von Kindern und Jugendlichen mit manifesten psychischen Störungen erhebt. Vielmehr soll anhand der ermittelten Prävalenzen eine Risikogruppe im Sinne des Präventionsgedankens definiert werden (Hölling et al., 2014; Klipker et al., 2018).

Konkrete Prävalenzangaben für spezifische Störungen bei Kindern und Jugendlichen anhand einer repräsentativen Stichprobe sind auf nationaler Ebene in Deutschland eher selten. Die Angaben zur Gesamtprävalenz psychischer Störungen im Kindes- und Jugendalter bewegen sich sowohl in internationalen als auch in nationalen Studien größtenteils im Bereich von 10-20%. Die Ergebnisse der beschriebenen Studien verdeutlichen, dass Verhaltensstörungen und psychische Auffälligkeiten bei Kindern und Jugendlichen in jedem Alter und in unterschiedlichsten Formen und Intensitäten auftreten können (Hölling et al., 2014). Diese Störungen und Auffälligkeiten können das Leben der Betroffenen in unterschiedlichen sozialen Bereichen stark beeinträchtigen und die allgemeine Lebensqualität vermindern. Zudem besteht ein hohes Risiko, dass sie sich im Laufe der weiteren Entwicklung manifestieren und so mit Langzeitfolgen bis in das Erwachsenenalter einhergehen (Haller et al., 2016). Um die Entwicklung von Verhaltensstörungen und psychischen Störungen sowie deren Manifestation zu verhindern, ist es entscheidend, diesen möglichst früh entgegenzuwirken. Je fortgeschrittener eine Störungsentwicklung ist, desto schwieriger ist es, diese durch bestimmte Maßnahmen zu behandeln (Petermann & Lehmkuhl, 2010). Im Umgang mit Verhaltensstörungen oder -auffälligkeiten spielen daher Handlungsformen der Prävention und Intervention eine zentrale Rolle. Präventive Maßnahmen umfassen dabei gezielte Handlungen, die „zur Vorbeugung gegen psychische, soziale und emotionale Störungen bei Kindern und Jugendlichen eingesetzt werden“ (Hillenbrand, 2008, S. 133). Die Übergänge zwischen Präventions- und Interventionsmaßnahmen sind teilweise fließend. Letztere werden bei bereits bestehenden Problembelastungen eingesetzt, sind allerdings in der Regel mit umfassenderen Konzepten

und daher auch mit einem größeren Aufwand verbunden (Hillenbrand, 2008). Die Schule als Institution, in der Kinder und Jugendliche einen Großteil ihrer Zeit verbringen, stellt ein bedeutsames Interventions- und Präventionssetting dar. Für den Erfolg von Maßnahmen in schulischen Settings ist es entscheidend, dass sie auf die individuellen Bedürfnisse und Schwierigkeiten der SuS ausgelegt sind (Brezinka, 2003). Dies bedeutet, dass Bedürfnisse und Problemfelder systematisch erfasst sowie Maßnahmen stetig evaluiert und angepasst werden müssen (Hillenbrand, 2008). Dabei ist es zudem notwendig, dass über die Maßnahmen sowie ihre Wirkung eine kontinuierliche Kommunikation zwischen den unterschiedlichen Kontexten, in denen sich ein Schüler oder eine Schülerin bewegt, geführt wird (Brezinka, 2003; Petermann & Lehmkuhl, 2010). Die Grundlage für die Entwicklung, Evaluation und Modifikation von präventiven und interventiven Maßnahmen bilden diagnostische Verfahren und Instrumente, mit denen gezielt Schülerdaten erhoben werden können (Hartmann, 2017). Welche diagnostischen Formen und Verfahren in schulischen Kontexten insbesondere hinsichtlich der konkreten Erfassung von Verhalten und Verhaltensveränderungen verwendet werden, wird im folgenden Kapitel dargelegt.

2.3 Diagnostik in schulischen Kontexten

Diagnostische Prozesse im Schulkontext zählen zu dem Bereich der pädagogischen Diagnostik. Dieser ist durch konkrete Zielsetzungen sowie die Verwendung spezifischer Kriterien und Methoden von anderen Diagnostikbereichen abzugrenzen. Die pädagogische Diagnostik ist mit pädagogischen Zielsetzungen verbunden, die durch die Aufgaben und Ansprüche eines Schulsystems geprägt sind (Ophuysen & Lintorf, 2013). Ein maßgebliches Ziel im Schulkontext ist „die Optimierung des Lernens der Schülerinnen und Schüler“ (ebd., 2013, S. 57). Damit dieses Ziel erreicht werden kann, müssen pädagogische Entscheidungen auf Basis gezielter diagnostischer Prozesse getroffen werden (ebd.). Über viele Jahrzehnte lag der Fokus dabei auf einer platzierungsrelevanten Diagnostik. Im Zuge dieser wurde anhand der Erhebung bestimmter Merkmale über die Platzierung eines Kindes in Regel- oder Sonderschulformen entschieden. Durch die Veränderungen im Schulsystem hinsichtlich des Leitbildes Inklusion hat sich auch die Funktion der pädagogischen Diagnostik im Schulkontext gewandelt. Der Fokus liegt nun mehr auf der Erfassung von individuellen

Unterstützungsbedarfen und der damit einhergehenden Planung, Durchführung und Evaluation von Fördermaßnahmen (Gold, Gawrilow & Hasselhorn, 2016).

Neben den beschriebenen entscheidungsbezogenen Funktionen sind auch die Vergabe von Qualifikationen und die informative Berichterstattung bedeutende pädagogische Handlungsfelder, die auf diagnostischen Prozessen basieren. Zu den zentralen Inhalts- oder Anwendungsbereichen der pädagogischen Diagnostik zählen unter anderem die Diagnostik von schulischen Leistungen, Lernvoraussetzungen und Lernschwierigkeiten sowie auch Verhaltensauffälligkeiten (Gold et al, 2016; Hesse & Latzko, 2017). Mit diagnostischen Prozessen im Schulkontext können also sehr unterschiedliche Bereiche fokussiert und verschiedene Interessen verfolgt werden. An diesen Bereichen und Interessen orientieren sich unterschiedliche diagnostische Methoden, Ansätze und Verfahren, die in der pädagogischen Diagnostik Anwendung finden (Hascher, 2005).

Der im Zuge des Inklusionsbestrebens eingesetzte Wandel von einer platzierungs- zu einer förderorientierten Diagnostik geht auch mit einer Veränderung in den benötigten diagnostischen Ansätzen und Verfahren einher. An dieser Stelle werden in der pädagogischen Diagnostik die Status- und die Prozessdiagnostik unterschieden (Casale et al., 2015b). Die Statusdiagnostik ist eng mit der Zielsetzung der Platzierungsdiagnostik verbunden, sodass die Begriffe in der Literatur teilweise auch synonym verwendet werden (Hartmann, 2017). Ziel bei dieser Form der Diagnostik ist die Erfassung des Ist-Zustandes eines Schülers oder einer Schülerin, um anhand dessen „prognostische Aussagen über die weitere Entwicklung vorzunehmen und damit u.a. auch Schulformzuweisungen […] zu begründen“ (ebd., 2017, S. 76). Weiterhin sind statusdiagnostische Erhebungen dadurch gekennzeichnet, dass sie selten und dafür umso intensiver durchgeführt werden, dabei eher breite Konstrukte messen und die Ergebnisse der SuS an einer sozialen Bezugsnorm beurteilt werden (Casale et al., 2015b). Insbesondere für den Anspruch, anhand der Erhebung des Ist-Zustandes Informationen für eine gezielte Förderung der SuS zu erlangen, steht die Statusdiagnostik in der Kritik, da mit den Ergebnissen lediglich Aussagen über den Lern- und Entwicklungsstand eines Kindes zu einem situativen Zeitpunkt getroffen werden können (Hartmann, 2017).

Da es im Sinne der Inklusion aktuell weniger um die Frage der Platzierung, sondern viel mehr um die Frage der optimalen Förderung geht, treten statt den statusdiagnostischen immer mehr prozess- bzw. verlaufsdiagnostische Verfahren und Ansätze in den Vordergrund. Die Verlaufsdiagnostik unterscheidet sich in einigen zentralen Aspekten von der Statusdiagnostik. So zielt sie auf der Basis kurzer und häufiger Messungen darauf ab, individuelle Veränderungen bei SuS hinsichtlich eines eng definierten Konstruktes zu erfassen. Mit Hilfe kontinuierlicher verlaufsdiagnostischer Erhebungen können Aussagen über die Entwicklung von SuS in verschiedenen Kompetenz- und Entwicklungsbereichen getroffen sowie in Verbindung damit auch Rückmeldungen über die Wirkung unterrichtlicher und förderspezifischer Maßnahmen gewonnen werden (Hartmann, 2017). Die Abbildung der Entwicklungsverläufe ist dabei anders als bei statusdiagnostischen Erhebungen an der individuellen Bezugsnorm orientiert (Casale et al., 2015b). An dieser Stelle soll noch einmal auf den in Kapitel 2.1 bereits kurz beschriebenen RTI-Ansatz eingegangen werden. Dieser Ansatz umfasst ein gestuftes Fördersystem, in dem die Nutzung kontinuierlicher Verlaufsdiagnostik eine entscheidende Rolle spielt. Je nach Entwicklungs- und Lernbereich werden mehrmals in einem Schuljahr, teilweise sogar wöchentlich, Erhebungen mit entsprechenden verlaufsdiagnostischen Verfahren durchgeführt. Anhand der gewonnenen Daten werden Entscheidungen für die weitere Beschulung und Förderung der SuS getroffen (Hartke, 2017). Zeigen die Daten beispielsweise eine Entwicklungsstagnation oder Rückschritte, „erfolgt schnellstmöglich eine zusätzliche, intensivere und individuellere Förderung auf [der nächsten] Förderebene“ (ebd., 2017, S. 23f.).

Der RTI-Ansatz verdeutlicht, dass die Verlaufsdiagnostik im Rahmen der individuellen Förderung in einem inklusiven Unterricht bedeutend ist. Dahingehend ergibt sich eine hohe Nachfrage nach verlaufsdiagnostischen Instrumenten, die jeweils auf bestimmte Inhaltsbereiche wie zum Beispiel akademische Leistungen, kognitive Kompetenzen oder auch das Verhalten von SuS ausgelegt sind (Fuchs, 2004; Voß & Gebhardt, 2017b). Wie genau verlaufsdiagnostische Instrumente und Verfahren aufgebaut sind und angewendet werden, ist abhängig von dem jeweiligen Bereich, auf den sie ausgelegt sind. Während einige Instrumente der Verlaufsdiagnostik des Lernens, auch Lernverlaufsdiagnostik genannt, bereits in Deutschland etabliert sind, gibt es in der relativ jungen Verhaltensverlaufsdiagnostik erste Ansätze, aber noch wenig fundierte

Instrumente (Hartmann, 2017). In den nachfolgenden Unterkapiteln werden die beiden Ansätze kurz dargestellt, wobei im Hinblick auf das Forschungsinteresse dieser Arbeit der Fokus auf der Verhaltensverlaufsdiagnostik liegt.

2.3.1 Lernverlaufsdiagnostik

Die Form der Lernverlaufsdiagnostik hat ihren Ursprung in dem Ansatz des „Curriculum Based Measurements“ (CBM), der in den 1970er Jahren von Stanley Deno entwickelt wurde. Grundgedanke dieses Ansatzes ist die kontinuierliche Erfassung der Beherrschung von Inhalten, die im aktuellen Unterricht behandelt und vermittelt werden. In Deutschland hat sich der Ansatz erst Jahrzehnte später durch einen Artikel von Klauer (2006) verbreitet (Souvignier, Förster & Zeuch, 2016). Klauer prägte zudem den für das CBM-Verfahren gängigen deutschen Begriff der Lernverlaufsdiagnostik (Klauer, 2011). Ziel der Lernverlaufsdiagnostik ist es, individuelle Lernverläufe bezogen auf bestimmte schulische Lernbereiche zu dokumentieren und anhand der somit erhaltenen Informationen, Lehr-, Lern- und Förderprozesse zu optimieren. Sie umfasst damit sowohl Informations- und Rückmeldungs- als auch Planungsfunktionen (Souvignier et al., 2016).

Im Fokus lernverlaufsdiagnostischer Verfahren stehen konkrete akademische Kompetenzen aus bestimmten Lernbereichen wie der Mathematik, dem Lesen oder der Rechtschreibung. Um Lernverläufe abbilden zu können, ist es notwendig, eine bestimmte Kompetenz über einen längeren Zeitraum und in relativ kurzen Abständen anhand von Testinstrumenten wiederholt zu erheben (Klauer, 2011; Gebhardt, Heine, Zeuch & Förster, 2015b). Die genaue Durchführung lernverlaufsdiagnostischer Verfahren ist von den jeweiligen konkreten Zielsetzungen der Erhebungen abhängig. An dieser Stelle lassen sich zwei Varianten der Lernverlaufsdiagnostik unterscheiden, die mit unterschiedlichen vorrangigen Interessen einhergehen. Zum einen kann die Verbesserung einer bereits vorhandenen Kompetenz durch gezielte Übungen im Fokus stehen (Klauer, 2011). Dazu werden Testinstrumente gebraucht, die immer wieder die gleiche Kompetenz erfassen und so „fortlaufend die Verbesserungen in Form abnehmender Fehler oder zunehmender Lösungsgeschwindigkeit dokumentieren“ (ebd., S. 219). Bei der zweiten Variante steht die Erweiterung des Wissens oder einer Kompetenz im Vordergrund. Um diese Erweiterung zu dokumentieren, müssen

Testinstrumente bereits von Beginn der Verlaufsmessung an den Kompetenzumfang prüfen, der am Ende erreicht werden soll. Weitere verlaufsdiagnostische Variationen ergeben sich durch Unterschiede in Testintervallen und organisatorischen Rahmenbedingungen wie z. B. dem Testgruppenumfang oder der Testlänge (ebd.).

Die Konstruktion konkreter lernverlaufsdiagnostischer Testinstrumente steht in Abhängigkeit zu den beschriebenen Faktoren und stellt hinsichtlich der Erfüllung notwendiger Testgütevoraussetzungen von diagnostischen Verfahren eine große Herausforderung dar. Testinstrumente müssen zunächst den klassischen Testgütekriterien Objektivität, Reliabilität und Validität entsprechen, um Zufallsergebnisse sowie Beurteilungs- und Wahrnehmungsfehler zu reduzieren (Bundschuh & Winkler, 2019, Souvignier et al., 2016). Zudem besteht der Anspruch, dass die Testinstrumente ökonomisch einsetzbar sind und gleichzeitig differenzierte Informationen für anknüpfende Fördermaßnahmen liefern (Souvignier et al., 2016). Eine besondere Schwierigkeit in der Testkonstruktion lernverlaufsdiagnostischer Instrumente besteht darin, dass für die wiederholte Messung einer Kompetenz immer neue Tests verwendet werden müssen, „die stets dasselbe Leistungsspektrum abdecken sollen und immer gleich schwierig sein müssen“ (Klauer, 2011, S. 209). Konstrukt- und Schwierigkeitshomogenität sind von zentraler Bedeutung, um Verzerrungen in den Lerneffekten und damit auch in der Abbildung von Lernverläufen zu vermeiden (Souvignier et al., 2016). Ein weiteres entscheidendes Kriterium, das lernverlaufsdiagnostische Testinstrumente erfüllen müssen, ist die Änderungssensibilität. Sie sollten dahingehend in der Lage sein, sowohl Lernstagnationen als auch Veränderungen in Form von Fort- oder Rückschritten zu erfassen (Klauer, 2011).

Der Anspruch an die Erfüllung der beschriebenen Kriterien stellt die Konstruktion von lernverlaufsdiagnostischen Tests vor einige Probleme. Daher gibt es im deutschsprachigen Raum bisher nur relativ wenig fundierte Testinstrumente (Hennemann, 2015). Anerkannte Testverfahren, die an englischsprachige CBM-Konzepte angelehnt sind, gibt es vor allem im Lernbereich Lesen. Darunter sind zum Beispiel die Verlaufsdiagnostik sinnerfassendes Lesen (Walter, 2013) oder das Inventar zur Erfassung der Lesekompetenzen von Erstklässlern (Diehl & Hartke, 2012). Auch im mathematischen Bereich wurden einige Tests wie die Lernverlaufsdiagnostik Mathematik (Strathmann & Klauer, 2012) entwickelt (Souvignier et al., 2016). Zudem treten im

Hinblick auf das Kriterium der Ökonomie immer mehr computerbasierte Testverfahren in den Vordergrund. Forschergruppen von deutschen Universitäten entwickeln Testverfahren, die dann auf unterschiedlichen Internetseiten abrufbar sind (Voß, 2017). Als Beispiel soll an dieser Stelle die Onlineplattform LEVUMI (www.levumi.de) genannt werden, in deren Projektrahmen die vorliegende Arbeit eingebettet ist. Auf der Plattform werden unter anderem empirisch geprüfte lernverlaufsdiagnostische Tests für die Bereiche Lesen und Mathematik kostenlos angeboten (Mühling, Gebhardt & Diehl, 2017). Eine detaillierte Vorstellung der Plattform und des Projekts LEVUMI erfolgt in Kapitel 4.1.

2.3.2 Verhaltensverlaufsdiagnostik

Die Lernverlaufsdiagnostik fokussiert den Entwicklungsverlauf akademischer Kompetenzen. Ein weiterer bedeutender Entwicklungsbereich ist, wie in Kapitel 2.3 beschrieben, die soziale und emotionale Entwicklung von Kindern und Jugendlichen. Die Förderung dieser Entwicklung stellt neben der Förderung akademischer Kompetenzen einen ebenso schulrelevanten Aufgabenbereich dar, in welchem auch ein großer Bedarf für verlaufsdiagnostische Instrumente besteht. Rückstände in der Entwicklung sozialer und emotionaler Kompetenzen äußern sich in negativen Verhaltensweisen und gehen oftmals mit einem problematischen Lernverhalten einher (Linderkamp & Grünke, 2007). Die Diagnostik im Bereich der emotionalen und sozialen Entwicklung ist daher oftmals auf das Verhalten von SuS ausgelegt, weshalb Casale et al. (2015b) die Diagnostik von Verläufen in diesem Bereich als Verhaltensverlaufsdiagnostik bezeichnen.

Abgesehen von dem inhaltlichen Fokus ähneln die Ziele und Funktionen der Verhaltensverlaufsdiagnostik denen der Lernverlaufsdiagnostik. Sie dient der Erfassung individueller Lern- und Entwicklungsstände von SuS im sozial-emotionalen Bereich sowie der damit einhergehenden Identifikation auffälliger und problematischer Verhaltensweisen. Auf der Basis dieser Erhebungen können dann gezielte Fördermaßnahmen für einzelne SuS entwickelt werden. Durch die kontinuierliche Fortführung der Erhebungen mithilfe verlaufsdiagnostischer Instrumente werden zudem individuelle Verhaltens- und Entwicklungsverläufe abgebildet und anhand dessen Fördermaßnahmen evaluiert und ggf. angepasst (Casale, Hennemann & Grosche, 2015a). In

gestuften Fördersystemen wie dem RTI-Ansatz werden verhaltensverlaufsdiagnostische Instrumente beispielsweise möglichst früh im Förderprozess und je nach Förderebene mehrmals täglich eingesetzt, um eine optimale Förderung zu gewährleisten. Die Verhaltensverlaufsdiagnostik spielt daher bei konkreten Verhaltensfördermaßnahmen eine zentrale Rolle und kann somit zu einer evidenzbasierten, also wissenschaftlich und praktisch erprobten, (sonder-)pädagogischen Praxis beitragen (Casale, Grosche, Volpe & Hennemann, 2017).

Die Auswahl konkreter Inhalte der Verhaltensverlaufsdiagnostik erfolgt in Orientierung an den situativen Kontext, in dem die Diagnostik stattfindet. Welche Verhaltensbereiche und Verhaltensweisen fokussiert und beurteilt werden, steht also in Abhängigkeit zu dem jeweiligen schulischen Setting. Dabei ist zu berücksichtigen, dass die Beurteilung des Verhaltensverlaufes eines Schülers oder einer Schülerin anhand der individuellen Bezugsnorm vorgenommen wird, aber die Verhaltensbeurteilung zu den einzelnen Messzeitpunkten an der sozialen Bezugsnorm bzw. an den Normen und Erwartungen in einem bestimmten Setting orientiert ist (Voß & Gebhardt, 2017b). Die Auswahl bestimmter Verhaltensbereiche und Verhaltensweisen ist damit ein entscheidender Aspekt in der Konstruktion entsprechender diagnostischer Instrumente. Die Forschung im Bereich der Verhaltensverlaufsdiagnostik in Deutschland ist noch relativ jung, sodass es, anders als bei der Lernverlaufsdiagnostik, momentan noch wenig geprüfte deutschsprachige Testinstrumente und Verfahren gibt, „die für eine engmaschige Erfassung des Verhaltens geeignet sind“ (Casale, 2017, S. 1). Der Grund für diese Lücke liegt nach Casale et. al. (2015b) unter anderem in der Problematik, dass verhaltensverlaufsdiagnostische Instrumente einige Testgütekriterien erfüllen müssen, um für die Analyse von Verhaltensverläufen geeignet zu sein. Im weiteren Verlauf ihrer Ausführungen erläutern sie insgesamt elf allgemeine und spezifische Kriterien, die sie als zentrale Testgütekriterien der Verlaufsdiagnostik bezeichnen (Casale et al., 2015b). Diese werden im Folgenden kurz einzeln vorgestellt, um einen Überblick über die Anforderungen an verhaltensverlaufsdiagnostische Instrumente zu schaffen.

Das Gütekriterium der *Objektivität* ist dann gegeben, wenn die Ergebnisse eines Tests von der jeweiligen Testleitung und den Messzeitpunkten unabhängig sind. Dies bedeutet, dass die Durchführung, die Auswertung und die Interpretation eines Tests

trotz unterschiedlicher Testleiter und Messzeitpunkte nicht variiert. Da die Wahrnehmung von Verhalten generell sehr subjektiv ist, müssen Instrumente der Verhaltensverlaufsdiagnostik so konstruiert werden, dass sie eine möglichst objektive Messung gewährleisten (Casale et al., 2015b).

Das Gütekriterium der *Reliabilität* gibt an, mit welchem Messgenauigkeitsgrad ein bestimmtes Merkmal erfasst wird. Berechnet werden können die interne Konsistenz, also die Korrelation der Testitems untereinander, die Retest-Reliabilität, das heißt die Korrelation der Messwerte eines Tests über zwei Messzeitpunkte, sowie die Paralleltestreliabilität, bei der die Messwerte zweier paralleler Tests mit unterschiedlichen Items verglichen werden. Eine hohe Reliabilität ist bei verhaltensverlaufsdiagnostischen Instrumenten wichtig, um Zufallsveränderungen im Verhaltensverlauf durch Messunigenauigkeiten zu verhindern (Casale et al., 2015b).

Die *Validität* gibt an, „ob ein Test wirklich das misst, was er zu messen beansprucht“ (Casale et al., 2015b, S. 39). Die Inhaltsvalidität eines Tests wird beispielsweise durch Theorieableitungen oder Expertenbefragungen erfasst, während die Konstruktvalidität, die den Präzisionsgrad der Messung eines bestimmten Merkmals angibt, über die Korrelationsvergleiche von Tests mit gleichen oder unterschiedlichen Konstrukten geprüft wird. Für die Verhaltensverlaufsdiagnostik in der Schule ist vor allem entscheidend, dass Testinstrumente eine soziale Validität aufweisen, also Verhaltensweisen messen, die im schulischen Setting von Bedeutung sind (ebd.).

Mit dem Gütekriterium der *Skalierung* ist die Verrechnungsvorschrift gemeint, mit der ein Testwert gebildet wird. Sie umfasst die Festlegung eines repräsentativen Testwerts für gemessene Merkmalsausprägungen und dient deren Interpretation. Zudem gibt die Skalierung an, ob die festgelegten Testwerte die empirischen Merkmalsunterschiede zwischen verschiedenen Testpersonen angemessen abbilden (Casale et al., 2015b).

Der Anspruch des Gütekriteriums *Ökonomie* geht mit den Forderungen einher, dass Testinstrumente sowohl in der Durchführung als auch in der Auswertung leicht anzuwenden sind und dabei möglichst wenig personelle und zeitliche Ressourcen beanspruchen. Insbesondere hinsichtlich der häufigen Messzeitpunkte im Rahmen der

Verlaufsdiagnostik besteht ein Bedarf an kurzen und schnellen Testverfahren, die bereits an die Planung von Fördermaßnahmen anknüpfen (Casale et al., 2015b).

Für Testinstrumente ist es weiterhin wichtig, dass sie ein *gültiges Messmodell* anwenden. Messmodelle beschreiben eine Beziehung zwischen latenten, nicht beobachtbaren Variablen und manifesten, beobachtbaren Variablen. Während in reflexiven Messmodellen „die Ausprägung der latenten Variablen die Merkmalsausprägung in den manifesten Indikatoren“ (Casale et al., 2015b, S. 19) bestimmt, ist die Wirkrichtung in formativen Messmodellen genau andersherum. In der klassischen Testtheorie (KTT) wird die Ausprägung der latenten Variablen aus den manifesten Itemwerten und einem Messfehler bestimmt. Probabilistische Messmodelle, die für Lernverlaufsdiagnostische Instrumente empfohlen werden, setzen für die Berechnung der Lösungswahrscheinlichkeit eines Items die Personenfähigkeit und die Itemschwierigkeit in Beziehung. In dem Bereich der Verhaltensverlaufsdiagnostik besteht zum einen die Problematik, dass bezüglich der Itemschwierigkeit noch Unklarheiten darüber bestehen, ob entweder bestimmte Verhaltensweisen oder die Situationen, in denen ein spezifisches Verhalten auftritt, die Items darstellen. Zudem wird Schülerverhalten in der Regel nicht als richtig oder falsch, sondern nur als situativ angemessen oder unangemessen eingeschätzt. Diese Einschätzung ist stets subjektiv und kann damit mit Messfehlern einhergehen (Casale et al., 2015b). Casale et al. (2015b) schlagen daher für die Verhaltensverlaufsdiagnostik Messmodelle vor, die auf der Erweiterung der KTT durch die Generalisierbarkeitstheorie (GT) basieren. Die Anwendung der GT ermöglicht es, neben der Analyse klassischer Testgütekriterien auch verschiedene Fehlerquellen, die die Verhaltensbeurteilung beeinflussen können, heranzuziehen. Zu diesen Fehlerquellen in der Verhaltensbeurteilung zählen z. B. verschiedene Rater, verschiedene Kinder und unterschiedliche Situationen (ebd.).

Eine wichtige Grundlage für Testinstrumente der Verlaufsdiagnostik ist die *Eindimensionalität* der zu messenden Konstrukte. Von Eindimensionalität kann gesprochen werden, wenn die Itemausprägung „auf nur eine spezifische Kompetenz, also die latente Variable, zurückzuführen ist“ (Casale et al., 2015b, S. 41). Eine Überprüfung der Eindimensionalität von Items, die für die Erfassung einer bestimmten Verhaltensdimension in der Schule genutzt werden, ist entscheidend, da sonst

Verhaltensveränderungen im Verlauf nicht mehr eindeutig auf die Entwicklung in einem Konstrukt zurückzuführen sind (ebd.).

Das Kriterium der *Änderungssensitivität* umfasst die Anforderung, dass Items verhaltensverlaufsdiagnostischer Tests bereits über kurze Messzeiträume Veränderungen in der latenten Variablen, also einem fokussierten Verhaltensbereich abbilden können. In diesem Sinne sollte von der Auswahl änderungsresistenter Items, die vorwiegend bei der Statusdiagnostik im Fokus stehen, abgesehen werden (Casale et al., 2015b).

Inferenz meint den Aufwand oder auch die Komplexität der aufzubringenden schlussfolgernden Kognition für die Beantwortung bzw. das Ausfüllen eines Tests. Im Bereich der Verhaltensverlaufsdiagnostik wird der Inferenzgrad maßgeblich dadurch bestimmt, wie genau Items und Itemausprägungen operationalisiert sind. Je allgemeiner ein Item formuliert ist, desto höher ist die Inferenz seitens der Lehrkräfte. Verhaltensverlaufsdiagnostische Instrumente sollten daher hinsichtlich der Gütekriterien der Ökonomie und der Objektivität keine hohe Inferenz aufweisen, sodass Einschätzungen eindeutig und ohne großen Aufwand vorgenommen werden können (Casale et al., 2015b).

Unter dem Begriff *Direktheit* fassen Casale et al. (2015b) das Kriterium zusammen, dass Instrumente der Verhaltensverlaufsdiagnostik auf die direkte Beurteilung kurzer Zeiträume ausgelegt sind. Die Registrierung bzw. Messung eines Verhaltens sollte möglichst zeitnah zu dessen Auftreten erfolgen, um so eine qualitative Erfassung zu sichern (ebd.).

In der Verhaltensverlaufsdiagnostik geht es um die Erfassung individueller Verhaltensveränderungen über einen bestimmten Zeitraum. SuS werden mit sich selbst verglichen, was einer *Orientierung an der individuellen Bezugsnorm* entspricht. In der Statusdiagnostik, die in der Regel an der sozialen Bezugsnorm orientiert ist, stellt die Normierung von Verfahren ein weiteres wichtiges Gütekriterium dar (Casale et al., 2015b). Normierung bedeutet, dass für ein Testverfahren bevölkerungsrepräsentative Vergleichswerte vorliegen, anhand deren individuelle Testwerte eingeschätzt werden können. Dabei beziehen sich Testnormierungen immer auf bestimmte Vergleichsgruppen, die nach verschiedenen Merkmalen wie z. B. Alter, Geschlecht,

geographischer Region oder auch spezifischen Behinderungen aufgestellt werden können (Bundschuh & Winkler, 2019). Über die Stichprobe einer großen Anzahl von Personen (n > 300), die hinsichtlich der Zielgruppe eines Tests eine repräsentative Vergleichsgruppe darstellen, werden Testergebnisse gewonnen. Anhand dieser Ergebnisse werden dann Normen bzw. Vergleichswerte berechnet, die sich bezogen auf verschiedene Merkmale für bestimmte Personengruppen unterscheiden können. Beispiele dafür sind unter anderem alters- oder geschlechtsspezifische Normen (Döring & Bortz, 2016). Eine wichtige Basis für den Prozess der Normierung ist das Modell der Normalverteilung. Das Modell beschreibt die Wahrscheinlichkeitsverteilung von Testwerten in Form einer symmetrischen Glockenkurve, die auf beiden Seiten gleichmäßig ansteigt und einen Gipfelpunkt hat. Die Übertragung dieses stochastischen Modells in die Praxis steht insbesondere bei psychischen Bereichen in der Kritik, da viele der erhobenen Daten nur äußerst selten normalverteilt sind. Trotzdem spielt es für die Testnormierung und damit für die Frage, inwieweit individuelle Testwerte vom Durchschnitt abweichen, eine grundlegende Rolle (Bundschuh & Winkler, 2019). Da es bei der Verhaltensverlaufsdiagnostik im Sinne des Inklusions- und Fördergedankens um die individuelle Verhaltensentwicklung der einzelnen SuS geht, ist eine Normierung von Testverfahren in diesem Bereich allerdings nicht zwingend notwendig (Casale et al., 2015b).

Casale et al. (2015b) konstatieren, dass die Liste der Gütekriterien an dieser Stelle noch weitergeführt werden könnte, die beschriebenen elf Kriterien aber für die Verhaltensverlaufsdiagnostik am wichtigsten erscheinen. In ihren weiteren Ausführungen evaluieren sie gängige Verfahren der Verhaltensdiagnostik anhand der elf Testgütekriterien. Dabei stellen sie fest, dass aus den fünf Oberkategorien Interviewverfahren, Beobachtungsverfahren, Beurteilungsverfahren, Dokumentenanalyse und Alltagsbeobachtungen zwei spezifische Methoden mit den meisten Gütekriterien übereinstimmen: die systematische Verhaltensbeobachtung und die Verhaltensbeurteilung in Form von Ratingskalen (ebd.).

Die systematische Verhaltensbeobachtung stellt eine spezifische Form der Verhaltensbeobachtung dar. Verhaltensbeobachtungen werden dazu genutzt, „Daten über die Häufigkeit bzw. die Dauer des Auftretens von konkreten Verhaltensweisen“ (Casale et al., 2015b, S. 44) zu erhalten. Bei der systematischen Verhaltensbeobachtung

wird in der Regel ein bestimmter Verhaltensbereich fokussiert. Dem Beobachter wird also genau vorgegeben, welche Verhaltensweisen zu beobachten sind und wie die Beobachtung dokumentiert werden soll. Die Beobachtungssituation kann direkt oder indirekt sein. Bei direkten systematischen Verhaltensbeobachtungen ist die beobachtende Person aktiv oder passiv in der Situation anwesend, während bei der indirekten Form beispielsweise Videomittschnitte für die Beobachtung genutzt werden (Huber & Rietz, 2015). Casale et al. (2015b) zeigen in ihrer Evaluation, dass die systematische Verhaltensbeobachtung etwa zehn von elf Gütekriterien der Verhaltensverlaufsdiagnostik entspricht, allerdings aufgrund eines hohen Umsetzungsaufwands für den häufigen Einsatz in der Schule nur bedingt geeignet ist.

Verhaltensbeurteilungen sollen wie Verhaltensbeobachtungen ebenfalls Daten über die Auftretenshäufigkeit und -dauer von bestimmten Verhaltensweisen liefern (Casale et al., 2015b). Anders als bei Beobachtungen geht es bei der Verhaltensbeurteilung aber um die retrospektive Einschätzung von Verhalten z. B. über einen umfassenden Beobachtungszeitraum oder auch von einer Beobachtungssituation, die bereits länger zurückliegt. Die Verhaltenseinschätzung erfolgt dabei meistens über standardisierte und mehrstufige Ratingskalen (Huber & Rietz, 2015). In Hinblick auf die beschriebenen Gütekriterien sind Verhaltensbeurteilungen mittels Ratingskalen besonders ökonomisch, weisen aber ein hohes Risiko für Beurteilungsfehler auf. Zudem kann die Veränderungssensitivität durch eine hohe Latenzzeit zwischen dem Auftreten des Verhaltens und der Beurteilung eingeschränkt werden (Casale et al., 2015b).

Casale et al. (2015b) konstatieren, dass beide beschriebenen Verfahren einzeln betrachtet, nicht uneingeschränkt für die Verhaltensverlaufsdiagnostik in schulischen Settings geeignet sind. Die im US-amerikanischen Raum bereits verbreitete Methode des Direct Behavior Rating (DBR) umfasst allerdings eine Kombination aus systematisch-direkter Verhaltensbeobachtung und der Verhaltensbeurteilung mittels Ratingskalen und wird im deutschsprachigen Raum als geeignete Methode für die Verhaltensverlaufsdiagnostik diskutiert (ebd.).

2.4 Direct Behavior Rating

Als potenziell geeignete Methode in der Verhaltensverlaufsdiagnostik wird in diesem Kapitel das DBR näher vorgestellt. Da gewisse Kenntnisse über die Methode für die

später folgenden Ausführungen bezüglich der Forschungsfrage dieser Arbeit grundlegend sind, werden zunächst das allgemeine Konzept des DBR sowie elementare Eigenschaften und Funktionen beschrieben. Im zweiten Unterkapitel wird dann der aktuelle Forschungsstand zur Methode und konkreten Instrumenten dargelegt.

2.4.1 Theoretische und konzeptuelle Grundlegung

Die Ursprünge des DBR als Methode der Verhaltensdiagnostik sind auf ein Interventionsverfahren des US-amerikanischen Schulpsychologen Edlund (1969) zurückzuführen. Dieses Verfahren im schulischen Setting hatte die Steigerung angemessenen Verhaltens von SuS zum Ziel und umfasste eine tägliche Rückmeldung über akzeptables Sozialverhalten in verschiedenen Unterrichtsblöcken. Die Rückmeldung in Form eines von der Lehrkraft geführten Ratingformulars wurde den SuS am Ende jedes Schultages mitgegeben, sodass auch die Eltern der Kinder in die Intervention einbezogen wurden und letztlich für die Belohnung von positiven Ergebnissen verantwortlich waren. Über die Jahre wurden ausgehend von dem Ansatz dieses Interventionsverfahrens weitere Methoden und Verfahren entwickelt, die heutzutage unter der übergeordneten Methode DBR zusammengefasst werden (Briesch, Chafouleas & Riley-Tillman, 2016). Da sich das DBR in der US-amerikanischen Verhaltensdiagnostik in Ansätzen bewährt hat, wurde die Methode unter dem Begriff *Direkte Verhaltensbeurteilung* (DVB) in den deutschen Sprachraum übertragen und bereits in einigen empirischen Studien überprüft (Casale, 2017).

Wie im vorangegangenen Kapitel beschrieben, vereint das DBR verschiedene Elemente der systematisch-direkten Verhaltensbeobachtung und der Verhaltensbeurteilung mittels Ratingskalen. Ähnlich wie bei der systematisch-direkten Verhaltensbeobachtung wird beim DBR ein konkret operationalisierter Verhaltensbereich über einen festgelegten Zeitraum und in einer Situation, in der dieses Verhalten von Bedeutung ist, beobachtet (Casale et al., 2017). Dabei wird das Verhalten allerdings nicht anhand von Strichlisten dokumentiert, sondern „unmittelbar im Anschluss an die Situation anhand einer Ratingskala beurteilt“ (ebd., S. 145). Diese grobe Beschreibung der praktischen Umsetzung von DBR zeigt, dass konkrete Instrumente und Verfahren dieser Methode hinsichtlich der fokussierten Verhaltensbereiche und der Beobachtungsbedingungen (Situation, Zeitraum) differieren können. Im Kern ist die Methode

aber durch drei wesentliche Merkmale charakterisiert, die im Folgenden aufgeführt werden (Christ, Riley-Tillman & Chafouleas, 2009).

Das erste entscheidende Merkmal ist die Direktheit der Beobachtung, die sich auf die zeitliche Nähe zwischen dem Beobachtungszeitraum, in dem ein bestimmtes Verhalten auftritt, und der darauf basierenden Verhaltensbeurteilung bezieht. Demnach sollte beim DBR die Verhaltensbeurteilung bzw. das Rating möglichst unmittelbar nach dem Beobachtungszeitraum stattfinden, um mögliche Beurteilungsfehler aufgrund von Erinnerungslücken zu vermeiden (Briesch et al., 2016). In Bezug auf den Beobachtungszeitraum, der sowohl einen ganzen Tag als auch nur wenige Sekunden umfassen kann, besteht hinsichtlich des Merkmals der Direktheit ein gewisser Interpretationsspielraum (Huber & Rietz, 2015). Oftmals werden relativ kurze Beobachtungszeiträume empfohlen, um eine möglichst direkte Verhaltensbeurteilung sicherzustellen (Briesch et al., 2016). In der Praxis spielen vor allem ökonomische Aspekte und Auftretenswahrscheinlichkeiten von Verhalten bei der Entscheidung über die Länge von Beobachtungszeiträumen eine Rolle (Huber & Rietz, 2015). Während bei zu kurzen Zeiträumen das Risiko einer niedrigen Auftretenswahrscheinlichkeit besteht, können sehr lange Zeiträume zu Verzerrungen in der Beurteilung und Wahrnehmung eines Raters führen. Ein Beispiel für Letzteres ist der sogenannte Halo-Effekt, bei dem der generelle Eindruck über eine Person die Verhaltensbeurteilung beeinflusst. In schulischen Kontexten haben sich in Bezug auf diese Aspekte insbesondere Beobachtungszeiträume, die mit einzelnen Unterrichtsblöcken oder Pausen korrespondieren, bewährt (Briesch et al., 2016).

Ein weiteres Kernmerkmal des DBR ist, dass ein spezifisches Verhalten fokussiert und beobachtet wird. Dabei sollte berücksichtigt werden, dass Verhalten situationsspezifisch ist. Verschiedene Verhaltensbereiche oder Verhaltenstypen, die sich in unterschiedlichsten Formen und Verhaltensweisen äußern, haben gemeinsam, dass sie mehr oder weniger gut von Ratern beobachtet und beurteilt werden können. Beim DBR sollten Verhaltensbereiche und Verhaltensweisen ausgewählt werden, die sowohl beobachtbar als auch messbar sind. Von abstrakten und globalen Konstrukten sollte abgesehen werden (Briesch et al., 2016). Insbesondere in schulischen Kontexten ist eine ausführliche Operationalisierung von Bedeutung, „so dass die

beobachtbaren Verhaltensweisen für alle am Förderprozess des Kindes beteiligten Personen nachvollziehbar und verständlich sind“ (Casale et al., 2015b, S. 47).

Das dritte Kernmerkmal bezieht sich auf die Nutzung von Ratingskalen für die Verhaltensbeurteilung. Ratingskalen ermöglichen die Wahrnehmungsquantifizierung durch eine beobachtende Person und damit die Einschätzung eines ausgewählten Verhaltens bezüglich verschiedener Aspekte wie Häufigkeit, Dauer und/ oder Intensität (Briesch et al., 2016).

Die beschriebenen Merkmale weisen bereits in Ansätzen auf die Vielfältigkeit von Verfahren und Instrumenten des DBR hin, die in schulischen Settings unterschiedliche Funktionen ausführen können. Zum einen können DBR als konkrete Interventionsinstrumente für die Förderung von SuS genutzt werden (Briesch et al., 2016). Diese Einsatzmöglichkeit geht auf die Anfänge der Methode bzw. auf das von Edlund (1969) entwickelte Interventionsverfahren zurück. Die ausgefüllten Ratingskalen werden für die kontinuierliche Verhaltensrückmeldung an SuS und/ oder auch deren Eltern genutzt und können ggf. mit Verstärkersystemen wie z. B. ausgewählten Belohnungen für positives Verhalten verknüpft werden. Ein Beispielverfahren in diesem Einsatzbereich ist die Daily Report Card (DRC). Eine Karte soll jeweils die individuellen Verhaltensprobleme eines Schülers oder einer Schülerin fokussieren. Über den Tag wird dann das jeweilige Verhalten anhand der DRC von einer Lehrkraft bewertet und den SuS rückgemeldet. Am Ende des Schultages werden die DRC für die Eltern mitgenommen, die dann angehalten sind, tägliche Verhaltensziele zu verstärken (Briesch et al., 2016).

Über die letzten 20 Jahre hat sich eine weitere Einsatzmöglichkeit von DBR im Bereich des Assessment (Einschätzung, Beurteilung) etabliert. Diese bezieht sich auf die Nutzung von DBR als Instrumente zur Erfassung von Verhaltensverläufen sowie der damit einhergehenden Evaluation von Fördermaßnahmen und lässt sich daher dem Funktionsbereich der Verhaltensverlaufsdiagnostik zuordnen (Briesch et al., 2016). Briesch et al. (2016) benennen zudem, dass das DBR im Bereich des Assessments auch für Screenings, also systematische Testverfahren, die im schulischen Kontext genutzt werden, um aus einer ausgewählten Schülergruppe Individuen mit

bestimmten Merkmalen wie beispielsweise Verhaltensauffälligkeiten zu identifizieren, eingesetzt werden können.

Eine dritte Funktion, die DBR in schulischen Kontexten ausführen können, ist die Förderung von Kommunikation. Individuen bewegen sich in verschiedenen sozialen Systemen, mit unterschiedlichen Normen und Werten (Tschira, 2005). Die Systemtheorie nach Bronfenbrenner basiert auf dieser Annahme und postuliert zudem, dass diese Systeme miteinander in Verbindung stehen und z. B. das Verhalten eines Kindes direkt oder indirekt beeinflussen. Um bestimmte Ergebnisse und Ziele hinsichtlich einer Verhaltensveränderung bzw. -förderung zu erreichen, kann eine gezielte Kommunikation über verschiedene Systeme und Personen notwendig sein. In Bezug darauf können DBR für den Informationsaustausch zwischen Akteuren, die an der Förderung beteiligt sind, genutzt werden. Bei dem oben vorgestellten Verfahren der DRC ist die Kommunikation zwischen Lehrkräften, SchülerInnen und Eltern bereits ein wichtiger Bestandteil der Intervention. Ein Vorteil von DBR ist dahingehend, dass nicht nur negative Verhaltensweisen zurückgemeldet werden, sondern auch positives Verhalten anerkannt wird. Dies kann zu einer konstruktiven Kommunikation beitragen. Zudem können Informationen über das Verhalten von SuS in gestuften Fördersystemen schnell und einfach ausgetauscht werden, da die Nutzung und Interpretation von DBR oftmals sehr simpel und ohne großen Zeitaufwand zu bewältigen ist. Durch eine kontinuierliche und einfache Kommunikation über verschiedene Kontexte ist es außerdem möglich, Fördereffekte stärker zu generalisieren, indem Förderziele und -methoden sowie spezifische Verhaltenserwartungen in den unterschiedlichen Settings angeglichen werden (Briesch et al., 2016).

Nach Briesch et al. (2016) beschreiben die unterschiedlichen Einsatzmöglichkeiten in schulischen Settings einen besonderen Vorteil von DBR. Die Nutzung in verschiedenen Bereichen geht allerdings mit vielfältigen Ansprüchen einher. Diesen könne die Methode des DBR angesichts weiterführender Eigenschaften wie Flexibilität, Effizienz, Vertretbarkeit und Wiederholbarkeit aber entsprechen. Eine hohe Flexibilität ist hinsichtlich des Einsatzes in verschiedenen Settings und Situationen konstitutiv (ebd.). Da das DBR nicht durch ein spezifisches Instrument oder einen inhaltlichen Fokus definiert ist, kann die Methode an unterschiedliche Bedingungen und Interessen angepasst werden (Christ et al., 2009). Variationsmöglichkeiten bestehen z. B.

bezüglich des fokussierten Verhaltens, dem genutzten Ratingsystem (Checkliste oder Skalen), der Ratingfrequenz (einmal oder mehrmals täglich), dem Rater (Lehrkräfte, SuS) sowie auch der Häufigkeit, mit der die Informationen und Ratingdaten über die an der Förderung beteiligten Personen ausgetauscht werden. Zudem kann das DBR an unterschiedliche Settings und Populationen hinsichtlich Schulformen, Jahrgängen und Gruppengrößen angepasst werden (Briesch et al., 2016).

Als weitere Eigenschaft wird dem DBR eine gewisse Effizienz zugeschrieben. Effizienz kann an dieser Stelle mit dem Gütekriterium der Ökonomie gleichgesetzt werden. Im Schulalltag gilt ein Verfahren als ökonomisch und effizient, wenn es möglichst wenig personelle, materielle und finanzielle Ressourcen in Anspruch nimmt (ebd.). Briesch et al. (2016) beschreiben, dass das DBR eine besonders effiziente Methode darstellt, da ein Rating, vor allem wenn die ausführende Person einmal mit dem Instrument vertraut ist, sehr schnell durchgeführt werden kann. Außerdem könne das Verfahren ohne externe Hilfen und Mittel in bestehende Interventionssysteme eingebaut werden. Neben dem Kriterium der Ökonomie sollten DBR-Verfahren zudem valide und reliabel sein, um akkurate Daten für den jeweils intendierten Einsatzzweck erfassen zu können. Diese empirische Vertretbarkeit konnte bereits für spezifische DBR-Instrumente in Studien belegt werden. Generell stellt die Flexibilität von Verfahren und Instrumenten des DBR allerdings eine große Herausforderung für die Überprüfung von psychometrischen Testgütekriterien dar. Es besteht dahingehend noch ein großer Bedarf an weiterer Forschung mit unterschiedlichen Instrumenten unter verschiedenen Testbedingungen (Briesch et al., 2016). Im Hinblick auf den Einsatz von DBR-Verfahren in gestuften Fördersystemen im Sinne einer Verlaufsdiagnostik ist es des Weiteren entscheidend, dass diese wiederholbar sind. Instrumente sollten also auch über längere Zeiträume kontinuierliche Messungen ermöglichen. Diese Wiederholbarkeit ist eng mit dem Anspruch der Effizienz verbunden und zielt auf die Abbildung von Verhaltensverläufen ab (Briesch et al., 2016, Christ et al., 2009).

Angesichts der beschriebenen Flexibilität sind der Aufbau und die konkrete Gestaltung von DBR-Instrumenten sowohl vom Zweck der Erhebung als auch von den Rahmenbedingungen (Ressourcen) abhängig. Grundsätzlich werden aber zwei Formen von DBR unterschieden: Instrumente mit Single-Item Skalen (SIS) und Instrumente mit Multi-Item Skalen (MIS). Bei DBR-SIS erfolgt die Erfassung einer Dimension des

Schülerverhaltens mit lediglich einem Item (Casale et al., 2017). Diese Form von DBR ist insbesondere dann ökonomisch, wenn sehr globale Verhaltensausschnitte bzw. übergeordnete Verhaltensdimensionen wie das Arbeits- oder Sozialverhalten erfasst werden, „um bei einer Schülerin/ einem Schüler ein eher breit gefächertes Verhaltensproblem zu beurteilen“ (ebd., S. 145). Weniger geeignet sind DBR-SIS allerdings für die Erfassung von konkreten Verhaltensweisen. Dies ist allerdings für die Sammlung von Informationen über individuelle Erfolge und Verhaltensveränderungen durch spezifische Fördermaßnahmen besonders entscheidend. Für diesen Zweck wird daher der Einsatz von DBR-MIS empfohlen. Bei MIS wird eine übergeordnete Verhaltensdimension über mehrere Items erfasst. In der Regel werden dafür drei bis fünf spezifische Verhaltensweisen operationalisiert, die als Indikatoren für eine übergeordnete Verhaltensdimension gelten. Wie mit den Ergebnissen der einzelnen Items weiterverfahren wird, ist abhängig von der Bedarfs- und Interessenlage der Messung. So können die Items einerseits einzeln analysiert oder auch bezogen auf die übergeordnete Verhaltensdimension aufsummiert betrachtet werden (ebd.).

Die konkrete Auswahl von Items für DBR ist ein sehr komplexes Verfahren, bei dem unterschiedliche Aspekte berücksichtigt werden sollten. Ebenso wie bei der Entscheidung für eine Ratingform, sollte die Auswahl von Items in Abhängigkeit zum Einsatzinteresse getroffen werden. Ein relevanter Punkt ist dabei, ob ein DBR-Instrument universal für die Verhaltensbeurteilung von mehreren SuS genutzt werden soll oder individuell auf nur ein Kind ausgelegt ist (Briesch et al., 2016). Wie oben bereits beschrieben, können für SIS und MIS sowohl globale Verhaltensdimensionen (z. B. „Störendes Verhalten“, „Arbeitsverhalten“) als auch spezifische Verhaltensweisen (z. B. „Ruft dazwischen“, „Arbeitet konzentriert“) als Items ausgewählt werden (Casale et al., 2017). Briesch et al. (2016) konstatieren, dass global formulierte Items besser für DBR-SIS und universale Erfassungen geeignet sind, während spezifisch formulierte Items für DBR-MIS und umfassende Messungen auf einem intensiven Level angemessen erscheinen. Die Generierung von globalen oder spezifischen Items folgt zwei unterschiedlichen Ansätzen. Bei nomothetischen Ansätzen werden Items ausgewählt, „die sich in Studien mit großen Stichproben psychometrisch am besten bewährt haben“ (Casale et al., 2015b, S. 48). Diese nomothetisch entwickelten Items werden dann für die Verhaltensverlaufsdiagnostik bei Individuen genutzt. Während

die Testgüte dieser Items sehr gut ist, können sie aber für den Einzelfall aufgrund ihres globalen Auswahlhintergrunds unpassend sein. Die Itemgenerierung bei ideographischen Ansätzen geht dagegen immer von dem einzelnen Schüler oder der einzelnen Schülerin aus. Für jedes Kind wird also ein eigenes Instrument mit spezifisch ausgewählten Items erstellt. Auf diese Weise entwickelte DBR weisen eine sehr hohe Passung zwischen Item und Kind auf, wohingegen aber die Evaluation der Testgüte aufgrund des geringen Stichprobenumfangs (N = 1) deutlich erschwert ist (ebd.). Im Hinblick auf die Testgüte ist es generell entscheidend, dass Verhaltensweisen ausgewählt werden, die direkt beobachtbar sind. Diese sollten zudem so als Items formuliert und operationalisiert sein, dass sie mit einer möglichst niedrigen Inferenz einhergehen. Ein weiterer, bei der Itemauswahl zu berücksichtigender Aspekt ist zudem, dass diese bezüglich des Einsatzinteresses und des spezifischen Kontextes einer Messung sozial valide, also situations- und förderrelevant sein sollten. So sollten bei der Nutzung eines DBR als Interventionsinstrument hinsichtlich des Ziels, negatives Verhalten zu reduzieren und positives Verhalten zu fördern, nicht nur negative Verhaltensweisen, sondern auch positive Verhaltensweisen als Items im Rating inkludiert sein (Briesch et al., 2016).

Neben der Auswahl der Items spielt auch die Form der Ratingskala bei DBR eine Rolle. Bei DBR-Verfahren werden, ähnlich wie bei der Verhaltensbeurteilung, mehrstufige Skalen verwendet, um den Häufigkeits- oder Intensitätsgrad eines fokussierten Verhaltens einschätzen zu können (Briesch et al., 2016; Huber & Rietz, 2015). Für DBR-SIS hat sich eine 11-Punkte Skala, die numerisch von 0-10 kodiert ist, als Standardform etabliert. Zudem sind in den letzten Jahren aber auch andere Skalen wie z. B. mehrstufige Likert-Skalen, bei denen Items über fünf bis sieben Skalenpunkte eingeschätzt werden, in den Forschungsfokus gerückt (Briesch et al., 2016; Casale et al., 2015b). Generell werden für den Einsatz in der Verhaltensverlaufsdiagnostik mindestens sechsstufige Skalen empfohlen, um intraindividuelle Verhaltensveränderungen über die Zeit abbilden zu können (Christ et al., 2009). Über die Rohwerte von MIS oder SIS, die durch die Itemeinschätzung einer Lehrkraft anhand der genutzten Skala über verschiedene Messzeitpunkte gewonnen werden, ist es möglich, einen Verlauf in Form eines Liniendiagramms darzustellen, „wobei die x-Achse die Zeit und die y-Achse die Rohwerte abbildet“ (Casale et al., 2015b, S. 48). Bei MIS

ist dabei zu beachten, dass als Rohwert meist der Summenscore der Items eines übergeordneten Verhaltensbereichs dient. Anhand der erhaltenen Kurve können Lehrkräfte oder andere Beteiligte Informationen über die Entwicklung eines Kindes und die Wirkung von Fördermaßnahmen erhalten (ebd.).

In diesem Unterkapitel konnten wichtige Grundmerkmale und Funktionen von DBR-Verfahren vorgestellt werden. Einige der genannten Vorteile und Empfehlungen, insbesondere zur Gestaltung von DBR, basieren auf Forschungsergebnissen aus unterschiedlichen Studien. Im Folgenden soll daher ein kurzer Überblick über die Forschung zur Beobachtungsgüte von DBR sowie zu konkreten DBR-Instrumenten gegeben werden.

2.4.2 Forschungsstand im Bereich DBR

In einer umfassenden systematischen Review haben Huber und Rietz (2015) insgesamt 17 zentrale Methodenstudien zum DBR mit Fokus auf die Gütekriterien Validität und Reliabilität analysiert. Dabei haben die Autoren zunächst die Ergebnisse der Studien, die sich überwiegend auf DBR-SIS beziehen, zusammengefasst und anschließend mit Blick auf den praktischen Einsatz von DBR in der Verlaufsdiagnostik diskutiert. Für die Bestimmung der allgemeinen Beobachtungsgüte von DBR berücksichtigten Huber und Rietz (2015) vorwiegend grundlegende Studien, die den Varianzanteil von Messdaten durch den Einfluss verschiedener Facetten (z. B. unterschiedliche Rater) und die Stabilität von DBR-Beurteilungen über die Zeit untersuchten. Einige grundlegenden Ergebnisse werden nachfolgend kurz dargestellt.

Briesch, Chafouleas und Riley-Tillman (2010) stellten in einer Studie fest, dass bei Lehrkräften, die den DBR als Methode nutzten, etwa 20% der Beurteilungsvarianz durch die Interaktion dieser beobachtenden Lehrkräfte mit den zu beurteilenden SuS erklärbar war. Im Vergleich dazu betrug der Varianzanteil durch den Interaktionseffekt bei der Nutzung eines Instruments der direkten systematischen Verhaltensbeobachtung unter 1% (Briesch et al., 2010). Demnach scheinen DBR in einem hohen Maße von der beobachtenden Person geprägt zu sein (Huber & Rietz, 2015). Im Gegensatz dazu weisen die Befunde einer kleinen Studie von Steege, Davin und Hathaway (2001) auf „eine insgesamt gute Kriteriumsvalidität und eine vergleichsweise hohe Interraterübereinstimmung von DBR-Messungen“ (Huber & Rietz, 2015, S. 84). Eine

nähere Betrachtung von Urteilsfehlern zeigte zudem, dass die Beobachtungsgüte stark von dem beobachteten Verhaltensaspekt abhängig ist (Christ, Riley-Tillman, Chafouleas & Jaffery, 2011).

Im Hinblick auf die Stabilität von DBR-Beurteilungen untersuchten Riley-Tillman, Christ, Chafouleas, Boice-Mallach und Briesch (2011) die Übereinstimmung von Beurteilungen gleicher Probanden über zwei Messzeitpunkte. Für die Test-Retest-Reliabilität ergaben sich relativ hohe Übereinstimmungen, was darauf hindeutet, dass DBR-Beurteilungen nicht willkürlich erfolgen (Huber & Rietz, 2015).

Neben Studien zur grundlegenden Beobachtungsgüte von DBR beschäftigen sich einige Studien außerdem mit unterschiedlichen Faktoren bzw. methodischen Aspekten, die einen Einfluss auf die Beobachtungsgüte haben können. Für die konkrete Gestaltung des Skalendesigns konnten bisher keine empirisch fundierten und signifikanten Einflüsse von Skalenbreite oder -länge auf die Beobachtungsgüte von DBR nachgewiesen werden (Huber & Rietz, 2015). Empfohlen werden allerdings mindestens sechsstufige Skalen, um stabile Ergebnisse und Entwicklungsverläufe abbilden zu können (Christ et al., 2009).

Eine Studie, die sich mit dem Einfluss der Itemanzahl auf die Beobachtungsgüte von DBR beschäftigt, liegt von Volpe und Briesch (2012) vor. Diese untersuchten unter anderem die Anzahl notwendiger Messwiederholungen bei SIS und MIS für die Erreichung einer angemessenen Messgenauigkeit. Aus den Ergebnissen leiten die Autoren eine insgesamt höhere Messgenauigkeit für MIS ab, wobei unklar ist, ob die konkretere Operationalisierung der MIS-Items oder ihre Zusammenfassung zu einem stabileren Mittelwert positiven Einfluss auf die Messgenauigkeit haben (Huber & Rietz, 2015).

Die Frage, wie viele Messzeitpunkte notwendig sind, um möglichst reliable Messergebnisse zu erhalten, ist anhand der Befunde verschiedener Studien nicht eindeutig zu klären, da deren Ergebnisse aufgrund unterschiedlicher Studiendesigns nicht vergleichbar sind (ebd.). Christ et al. (2009) konstatieren in einer Zusammenfassung der bisherigen Forschungsergebnisse aber, dass die notwendige Anzahl von Messzeitpunkten in Abhängigkeit zu den jeweiligen Beobachtungszielen steht.

In Bezug auf die Auswahl und Formulierung von Beobachtungszielen zeigen Studienergebnisse, dass einzelne Beobachtungsziele im Sinne verschiedener Verhaltensbereiche bzw. Verhaltensweisen eine unterschiedliche Validität und Interrater-Reliabilität aufweisen können (Huber & Rietz, 2015). Für die Formulierung von Beobachtungszielen finden sich divergente Befunde. So konnten einige Autoren aus den Ergebnissen ihrer Studien eine höhere Messgenauigkeit für global formulierte Items ableiten, während andere Studien auf eine höhere Interrater-Reliabilität bei spezifisch formulierten Items hinweisen (Casale et al., 2015b). Auch die Valenz von Zielformulierungen, die den Fokus von Beobachtungszielen auf positiv, erwünschtes oder negativ, unerwünschtes Verhalten bestimmt, kann einen Einfluss auf die Beobachtungsgüte haben (Huber & Rietz, 2015). Studien von Riley-Tillman, Chafouleas, Christ, Briesch und LeBel (2009) sowie von Chafouleas, Jaffery, Riley-Tillman, Christ und Sen (2013) zeigten, dass eine Verbesserung der Messgenauigkeit durch eine bestimmte Formulierung in Abhängigkeit zum beobachteten Verhaltensbereich steht. So führte eine positive Zielformulierung in dem Bereich „Unterrichtsteilnahme“ zu einer höheren Beobachtungsgenauigkeit, während sich für den Bereich „störendes Verhalten“ eine negative Zielformulierung bewährte (Huber & Rietz, 2015).

Im Hinblick auf den Einfluss der Länge von Verhaltensstichproben auf die Beobachtungsgüte, gibt es bisher wenig fundierte Befunde. Einige Studienergebnisse weisen allerdings darauf, dass längere Beobachtungszeiträume eher zu einer höheren diagnostischen Güte beitragen als kurze Sequenzen (Christ et al., 2009).

Ein weiterer Faktor, der in verschiedenen Studien untersucht wurde, ist die Auswirkung von Beobachtertraining auf die Messgenauigkeit. Auch in diesem Bereich sind die Befunde nicht eindeutig (Huber & rietz, 2015). So stellten Schlientz, Riley-Tillmann, Briesch, Walcott und Chafouleas (2009) bei einer Trainingsgruppe im Vergleich zu einer Kontrollgruppe eine signifikant höhere Messgenauigkeit fest. Ergebnisse einer Studie von LeBel, Kilgus, Briesch und Chafouleas (2009) zeigten dagegen keine signifikanten Unterschiede zwischen Gruppen, die entweder intensives oder nur kurzes Beobachtertraining erhielten (Huber & Rietz, 2015).

Neben Huber und Rietz (2015) beziehen sich auch Casale et al. (2015b) bei der Überprüfung, der von ihnen beschriebenen Testgütekriterien für Instrumente der

Verlaufsdiagnostik (vgl. Kapitel 2.3.2), auf einige der oben genannten Studien. Sie kommen zu dem Ergebnis, dass die Methode des DBR sowohl den psychometrischen Gütekriterien als auch anderen Kriterien wie z. B. ökonomischen Ansprüchen entspricht und damit für den Einsatz in der Verhaltensverlaufsdiagnostik geeignet scheint. Dis bisherigen Befunde weisen darauf hin, dass Ergebnisse von DBR-Verfahren als Grundlage für individuelle und normorientierte Interpretationen in Bezug auf Zielverhaltensweisen geeignet sind (Casale et al., 2015b). Gleichzeitig konstatieren Casale et al. (2015b) ebenso wie Huber und Rietz (2015), dass der allgemeine Forschungsstand zum DBR zu gering ist, „um die Methode abschließend und fraglos für die Verlaufsdiagnostik empfehlen zu können“ (Casale et al., 2015b, S. 50). Viele Fragen und Unsicherheiten, z. B. über den Einfluss einer diagnostischen Situation auf die Testergebnisse oder über die Notwendigkeit einer Normierung von DBR-Instrumenten, bleiben bestehen (Casale et al., 2015b; Huber & Rietz, 2015). Da die bisherigen empirischen Befunde größtenteils von nur einer Forschergruppe aus dem nordamerikanischen Raum stammen, sind sie nur eingeschränkt interpretierbar. Zudem stehen die Güte und Qualität eines Testinstruments immer in Abhängigkeit zu dem jeweiligen Testdesign und den Durchführungsbedingungen, weshalb Ergebnisse zur DBR-Testgüte nicht auf alle Instrumente übertragen werden können. Dahingehend besteht im deutschsprachigen Raum weiterhin ein großer Forschungsbedarf zu der DBR-Methode und konkreten DBR-Instrumenten (Casale et al., 2015b).

Im Rahmen des Projekts LEVUMI, das in Kapitel 4.1 näher vorgestellt wird, wurden zwei Ratingskalen zur Verhaltensverlaufsmessung in Anlehnung an die DBR-Methode entwickelt und bereits hinsichtlich spezifischer Testgütekriterien untersucht. So wurde die von Gebhardt, Casale, Jungjohann und DeVries (2017) entwickelte Prototypversion des DBR-MIS in einer Masterarbeit von Sauerland (2018) mit Bezug auf die deutschsprachige Fremdbeurteilungsversion des SDQ sowie anhand von Experteninterviews überarbeitet. Erste Ergebnisse zur Testgüte des Instruments konnten dann in einer Pilotierungsstudie ebenfalls im Rahmen einer Masterarbeit gewonnen werden. Die in der Studie verwendeten Daten, welche über eine quantitative Erhebung mittels des DBR-MIS sowie über qualitative halbstrukturierte Experteninterviews gesammelt wurden, wurden anschließend von Hisker (2018) im Hinblick auf verschiedene Gütekriterien analysiert und ausgewertet. Basierend auf den

Studienergebnissen, konnte für den DBR-MIS insgesamt eine gute Testgüte insbesondere bezüglich der Kriterien Reliabilität, Änderungssensitivität und Ökonomie abgeleitet werden (ebd.). Hisker (2018) konstatierte jedoch, dass die Ergebnisse der Studie z. B. aufgrund von Durchführungsfehlern oder fehlenden Daten nur eingeschränkt betrachtet werden können und weiterführende Forschungsarbeiten notwendig sind. Ein wichtiger Punkt in Bezug auf die Ökonomie des Instruments sei dahingehend auch die Entwicklung von Auswertungshilfen, beispielsweise in Form eines computergestützten Auswertungstools. Ein solches lag zum Zeitpunkt der Arbeit noch nicht vor, weshalb der Auswertungsprozess im Rahmen der Studie nur geringfügig berücksichtigt wurde (ebd.).

Neben dem DBR-MIS ist auch das Ratinginstrument PUTSIE entwickelt und erprobt worden (Schurig et al., 2019). Zu dem Instrument liegt im Rahmen des Projekts LEVUMI ebenfalls eine Masterarbeit vor. In dieser untersuchte Krause (2019) die Anwendbarkeit des Instruments in der Praxis. Die qualitative Datengewinnung erfolgte anhand von mehreren Leitfrageninterviews mit Lehrkräften, die den PUTSIE praktisch erprobt hatten. Die Aussagen der interviewten Lehrkräfte zu gezielten Fragen, die der Erhebung ihrer Einstellungen und Erfahrungen bezüglich Anwendung, Auswertung und Handhabbarkeit des PUTSIE dienten, wurden in Kategorien zusammengefasst und ausgewertet. Die Ergebnisse zeigten, dass die Anwendung des Instruments in der Papierversion mit Unsicherheiten sowie einem gewissem zeitlichen und personellen Aufwand seitens der Lehrkräfte einhergeht. Probleme oder Herausforderungen konnten z. B. bei der Auswertung und Interpretation von Daten und Verlaufskurven sowie im Umgang mit Item- und Verhaltensbereichen festgestellt werden. Für die Weiterentwicklung und Überarbeitung des PUTSIE erscheinen daher ebenso weitere Forschungsarbeiten notwendig (Krause, 2019).

Zusammenfassend lässt sich also festhalten, dass sowohl internationale als auch erste nationale Forschungsarbeiten auf ein hohes Potenzial von DBR-Instrumenten für den Einsatz in der Verlaufsdiagnostik weisen. Im deutschsprachigen Raum wurden bereits konkrete Instrumente entwickelt und in ersten Studien untersucht. Die Ergebnisse zeigen, dass umfassendere Forschungen und Weiterentwicklungen der Instrumente notwendig sind, um Unklarheiten in der Anwendung sowie Messfehler zu reduzieren und so eine Etablierung der Instrumente in der Praxis voranzutreiben.

3. Fragestellung und Zielsetzung

Im Hinblick auf die Konzeption und den praktischen Einsatz von DBR stellt die Frage, inwiefern die Bereitstellung einer Vergleichsnorm für verhaltensverlaufsdiagnostische Instrumente zur intersubjektiven Einordnung von Testergebnissen sinnvoll und notwendig ist, einen bisher unzureichend diskutierten Forschungsbereich dar (Huber & Rietz, 2015). Die Durchführung einer Normierung ist ein wesentliches Gütekriterium für Testinstrumente in der Statusdiagnostik. Für diese Instrumente werden über die Berechnung der durchschnittlichen Testergebnisse einer repräsentativen Vergleichsstichprobe Normwerte bereitgestellt, sodass individuelle Testwerte anhand dieser Testnormen eingeordnet und eingeschätzt werden können (Döring & Bortz, 2016). So wurden für die verschiedenen Versionen des deutschen SDQ, der als Screening-Instrument für die Erfassung von Verhaltensauffälligkeiten und -stärken von Kindern und Jugendlichen genutzt wird, über die Jahre einige Normierungen durchgeführt (Lohbeck, Schultheiß, Petermann & Petermann, 2015; Woerner et al., 2002). Dabei wurden anhand der Skalenrohwertverteilung einer jeweils umfassenden Feldstichprobe Grenzwerte vorgeschlagen, die in Anlehnung „an die von Goodman (1997) empfohlenen Zielvorgaben von 80–10–10% für die Bestimmung von Grenzwerten in die drei Kategorien unauffällig, grenzwertig und auffällig“ (Lohbeck et al., 2015, S. 222) berechnet wurden. Zusätzlich wurden auch die Effekte der Variablen Alter, Geschlecht und Schichtzugehörigkeit betrachtet, was teilweise zur Angabe von geschlechts- und altersspezifischen Grenzwerten führte. Über die erhobenen Normdaten und Grenzwertbestimmungen im Rahmen der SDQ-Normierungen können individuelle Testergebnisse zu einer Kategorie (*auffällig*, *grenzwertig* oder *unauffällig*) zugeordnet werden, was eine Interpretationserleichterung darstellt und für Förderentscheidungen in schulischen Kontexten hilfreich sein kann (Lohbeck et al., 2015, Woerner et al, 2002). Zudem ermöglicht die Bereitstellung der SDQ-Normwerte in Forschungsstudien wie der KiGGS-Studie die Angabe von Prävalenzen bzw. von Risikogruppen für psychische Auffälligkeiten (Hölling et al., 2014).

Für Testinstrumente der Statusdiagnostik stellt die Normierung ein wichtiges Gütekriterium dar, das mit einer Erhöhung des diagnostischen Werts und der praktischen Brauchbarkeit eines Instruments einhergeht (Woerner et al., 2002). Individuelle Testergebnisse von SuS können so anhand einer kriterialen Bezugsnorm z. B. in Form

von Grenzwerten eingeordnet und interpretiert werden. Für den Bereich der Verhaltensverlaufsdiagnostik konstatieren Casale et al. (2015b), dass eine Normierung von Testinstrumenten nicht unbedingt wichtig sei, da das diagnostische Interesse dabei vielmehr auf der Erfassung von individuellen Entwicklungsverläufen und nicht auf der Einschätzung einer absoluten Ausprägung von Verhalten liege. Verhaltensverläufe sollten daher anhand der individuellen Bezugsnorm überprüft werden. Weitergehend erläutern die Autoren aber, dass für den Einsatz eines DBR-Instruments je nach diagnostischer Zielsetzung die Bereitstellung einer Vergleichsnorm hilfreich sein kann (ebd.). Im Rahmen von Fördermodellen wie dem RTI-Ansatz werden pädagogische Entscheidungen basierend auf den individuellen Ergebnissen der SuS in Screenings und Verlaufsdokumentationen getroffen (Blumenthal & Hartke, 2015). Hinsichtlich des Kriteriums der Ökonomie könnten DBR-Instrumente in einem solchen gestuften Fördersystem neben der Erhebung von Verhaltensverläufen als primäres Einsatzziel auch als Screening-Instrument genutzt werden. Angesichts dieser Einsatzmöglichkeiten von verhaltensverlaufsdiagnostischen Instrumenten in der Praxis kann eine Normierung also sinnvoll sein und den diagnostischen Wert des Instruments erhöhen. Zudem weisen die Forschungsergebnisse zu konkreten DBR-Instrumenten im Rahmen des Projekts LEVUMI auf generelle Auswertungs- und Interpretationsunsicherheiten seitens der Lehrkräfte im Umgang mit den Testergebnissen (vgl. Kapitel 2.4.2). Die Bereitstellung einer Vergleichsnorm könnte an diese Problematik anknüpfen und bei der Verlaufsdokumentation als strukturgebende Orientierungslinie für Lehrkräfte dienen sowie die Ergebnisinterpretation hinsichtlich der Identifikation und Ausprägung von auffälligem Verhalten erleichtern.

Das Forschungsinteresse der vorliegenden Arbeit bezieht sich daher auf die Bestimmung einer Vergleichsnorm für das Instrument DBR-MIS, welches auf der Onlineplattform LEVUMI als verhaltensverlaufsdiagnostisches Instrument für Lehrkräfte angeboten wird. Die Normierung eines solchen verhaltensverlaufsdiagnostischen Instruments könnte in Form der Angabe einer normierten Entwicklungsverlaufslinie oder, wie bei der deutschen Version des SDQ, durch die Bereitstellung von Schwellenwertangaben für die Eingrenzung von Kategorien vorgenommen werden. Da eine Normierung von Entwicklungsverläufen im Bereich der schulischen Verhaltensverlaufsdiagnostik allerdings eng mit bestimmten Interventions- und

Präventionsmaßnahmen verbunden ist, bezieht sich die Zielsetzung dieser Arbeit auf die Berechnung von Schwellenwerten für den DBR-MIS, die als Orientierungspunkte und Interpretationshilfen eine vergleichende Einordnung von individuellen Testergebnissen ermöglichen.

Für die Bearbeitung dieses übergeordneten Forschungsinteresses ist zunächst eine Erhebung umfassender Stichprobendaten mit der Ratingskala DBR-MIS erforderlich. Mit diesen Daten kann zunächst die faktorielle Struktur des Instruments überprüft werden. Der DBR-MIS umfasst zwei Konstruktebenen, über die das Rating ausgewertet werden kann. Damit liegen zwei Messmodelle (Sechs-Faktoren-Modell und Drei-Faktoren-Modell) vor, die als Grundlage für die Schwellenwertberechnung dienen können. Eine Überprüfung der Messmodelle bezüglich ihrer Passung mit den empirisch erhobenen Daten ist notwendig, um zu entscheiden, auf welcher Ebene die Berechnung von Schwellenwerten für das Instrument sinnvoll ist (Döring & Bortz, 2016). Anschließend können anhand der erhobenen Stichprobendaten allgemeine Schwellenwerte auf der ausgewählten Ebene berechnet werden. Für diese Berechnung werden zudem die in Kapitel 2.2.2 aufgeführten Prävalenzraten von verschiedenen Verhaltensauffälligkeiten und psychischen Störungen herangezogen und berücksichtigt. Da sowohl in der Literatur als auch in umfassenden Prävalenzstudien auf Verhaltensunterschiede bezüglich Heterogenitätsdimensionen wie Alter und Geschlecht hingewiesen wird (vgl. Kapitel 2.3), werden in einem nächsten Schritt auch mögliche Alters- und Geschlechtseffekte untersucht. Die konkreten Forschungsfragen, die die verschiedenen Schritte zur Umsetzung des übergeordneten Forschungsinteresses dieser Arbeit markieren, lauten daher:

1) Auf welcher Konstruktebene des DBR-MIS ist eine Schwellenwertberechnung sinnvoll?

   a) Welches Messmodell ist hinsichtlich der vorliegenden Stichprobengröße als Grundlage für die Schwellenwertberechnung am geeignetsten?

   b) Wie ist die interne Konsistenz des ausgewählten Faktormodells?

2) Welche allgemeinen Schwellenwerte können unter Berücksichtigung der Ergebnisse psychologischer Prävalenzforschung anhand der Mittelwertverteilung der vorliegenden Stichprobendaten berechnet werden?

3) Welchen Effekt haben die Variablen Alter und Geschlecht auf die Konstrukte des DBR-MIS? Ist eine Angabe von alters- und geschlechtsspezifischen Schwellenwerten notwendig?

Für die erste und dritte Forschungsfragen können aus der Theorie sowie aus bisherigen Forschungsergebnissen Hypothesen abgeleitet werden, die dann im Rahmen der nachfolgenden Analysen überprüft werden. Angesichts der auf der Lernplattform LEVUMI angebotenen computergestützten Datenauswertung des DBR-MIS, die auf dem Sechs-Faktoren-Modell basiert, kann angenommen werden, dass dieses Messmodell als Grundlage für die Berechnung von Schwellenwerten am geeignetsten ist. Die Hypothese (H1), die die Untersuchungen zur ersten Forschungsfrage begleitet, lautet daher: „*Das Sechs-Faktoren-Modell passt besser zu den empirischen Stichprobendaten als das Drei-Faktoren-Modell“*. Für die dritte Forschungsfrage wird aufgrund der bisherigen Studienergebnisse im Bereich der Verhaltens- und Prävalenzforschung, die auf einen Einfluss der Variablen Alter und Geschlecht auf das Verhalten von Kindern und Jugendlichen hinweisen (vgl. Kapitel 2.3.2), die Hypothese (H2) *„Die Variablen Alter und Geschlecht haben spezifische Einflüsse auf die Mittelwerte der Konstrukte des DBR-MIS“* formuliert. Der Bestimmung von konkreten Schwellenwerten im Rahmen der zweiten Forschungsfrage wird keine Hypothese vorangestellt, da bisher keine geeigneten Vergleichswerte aus vorangegangen Studien zum DBR-MIS vorliegen und die Fragestellung zudem ein eher exploratives Vorgehen impliziert.

4. Methodisches Vorgehen

In den folgenden Unterkapiteln wird das methodische Vorgehen zur Beantwortung der vorgestellten Forschungsfragen dargelegt. Zunächst erfolgt eine kurze Beschreibung des übergeordneten Forschungsprojekts LEVUMI, in dessen Rahmen die vorliegende Arbeit eingeordnet werden kann. Anschließend wird die Erhebungsmethode in Verbindung mit dem konkreten Erhebungsinstrument vorgestellt. Da aufgrund der Corona-Pandemie eine eigenständige Erhebung von Daten für die Untersuchungen im Rahmen dieser Arbeit nicht möglich war, wurde auf einen bereits bestehenden Datensatz zurückgegriffen. Infolge fehlender Informationen bezüglich der originalen Datenerhebung kann eine genaue Beschreibung der Erhebungsdurchführung im Folgenden nicht gewährleistet werden. Bekannte Eckpunkte der Datenerhebung werden allerdings im Zuge der Stichprobendarstellung in Kapitel 4.3 aufgegriffen und benannt. Den Abschluss des Methodenteils der vorliegenden Arbeit bilden die Vorstellung und Erläuterung der ausgewählten Analyse- und Auswertungsmethoden, die zur Beantwortung der einzelnen Fragestellungen herangezogen werden.

4.1 Projekt LEVUMI

Die Lernplattform LEVUMI ist ein kooperatives Forschungsprojekt mehrerer Universitäten (u. a. Technische Universität Dortmund, Europa-Universität Flensburg und Christian-Albrechts-Universität zu Kiel), das im Jahr 2015 gemeinschaftlich von den Wissenschaftler_innen Prof. Dr. Markus Gebhardt, Prof. Dr. Kirsten Diehl sowie Prof. Dr. Andreas Mühling initiiert und entwickelt wurde (Schurig et al., 2019). Zu den ausgeschriebenen Zielen der Lernplattform zählen zum einen die Vertiefung der Forschung zur Lernverlaufsdiagnostik sowie zum anderen die Entwicklung und Bereitstellung eines Onlineinstruments zur Lernverlaufsdiagnostik, das für den Einsatz in schulischen Kontexten praktikabel ist (Gebhardt, Diehl & Mühling, 2015). Das aktuelle Angebot der Lernplattform umfasst neben unterschiedlichen Tests zur Verlaufsmessung unter anderem auch informative Handbücher sowie Kopiervorlagen für den Unterricht und ist in die vier zentralen Lernbereiche Lesen, Rechtschreiben, Zahlen und Operationen sowie Verhalten eingeteilt (Schurig et al., 2019). Für jeden Lernbereich werden verschiedene Testarten und -instrumente zur Verfügung gestellt. Zudem werden im Rahmen des Onlineformats computergestützte Auswertungs- und Darstellungsmöglichkeiten für den Umgang mit Testergebnissen angeboten. Die gesamten

Materialien der Lernplattform sind für Lehrkräfte kostenfrei zugänglich. Seit der Gründung des Forschungsprojekts werden die Plattform und die dort angebotenen kostenlosen Tests stetig weiterentwickelt und durch Nutzerfeedback, Datenanalysen sowie gezielte Forschungsarbeiten optimiert. Auf diese Art und Weise soll in einem offenen und engen Austausch mit Lehrkräften, im Sinne einer bestmöglichen Förderung aller SuS, ein wichtiger Beitrag zur Forschung im Bereich der (Lern-)Verlaufsmessung geleistet werden (Gebhardt et al., 2015).

Bereits seit Anfang des Jahres 2018 beinhaltet die Onlineplattform Lernverlaufstests zum Lesen, der Rechtschreibung oder auch basalen mathematischen Fähigkeiten. Im Herbst 2018 wurde das Angebot um zwei Ratingskalen zur Verhaltensverlaufsmessung und einen Fragebogen zur sozialen Partizipation erweitert. Wie bereits in Kapitel 2.4.2 skizziert, sind die zwei im Rahmen des Projekts LEVUMI entwickelten und angebotenen Ratingskalen zum einen der DBR-MIS und zum anderen der PUTSIE. Unterschiede zwischen den beiden Ratingskalen ergeben sich durch die Orientierung an verschiedenen Klassifikationssystemen für psychische Störungen. Während der DBR-MIS an der „Internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme“ (ICD-10) orientiert ist, wurde der PUTSIE in Anlehnung an das „Diagnostic and Statistical Manual of Mental Disorders V“ (DSM-V) der American Psychiatric Association (2000) entwickelt (Schurig et al., 2019). Da sich das Forschungsinteresse der vorliegenden Arbeit auf den DBR-MIS bezieht, wird im Nachfolgenden der Aufbau und die Operationalisierung dieser Ratingskala beschrieben.

4.2 Erhebungsmethode und Operationalisierung

Die Erhebung der Daten, welche die Grundlage für die nachfolgenden Untersuchungen und Analysen bilden, erfolgte in einer quantitativen Feldstudie über die Ratingskala DBR-MIS, mittels der verschiedene Lehrkräfte das Verhalten von SuS im Schulalltag einschätzten. Damit lässt sich das Vorgehen in dieser Arbeit dem quantitativen Forschungsansatz zuordnen (Döring & Bortz, 2016). Ratingskalen werden in der Forschungspraxis oftmals „zur Messung intervallskalierter psychologischer und sozialer Variablen“ (ebd., S. 244) eingesetzt. Die Messung bei Ratingskalen erfolgt dabei über Rater, die den Ausprägungsgrad eines fokussierten Merkmals auf einer

gestuften Skala subjektiv einschätzen. In der Regel wird davon ausgegangen, dass Ratingskalen intervallskaliert sind, was beinhaltet, dass die Abstände zwischen den Skalenstufen als gleich groß bewertet werden sollten. Wie Skalenstufen gesetzt und benannt werden ist abhängig von den zu beurteilenden Merkmalen und dem konkreten Einsatzziel der Ratingskala (ebd.).

Mit dem DBR-MIS wird im Rahmen des Projekts LEVUMI ein multidimensionales Verfahren angeboten, das für den Einsatz auf der Lernplattform stetig weiterentwickelt und überprüft wird. Ein erster Prototyp des Instruments wurde von Gebhardt et al. (2017) in Anlehnung an die deutschsprachige Fremdbeurteilungsversion des SDQ entwickelt. Der SDQ ist ein etabliertes und evaluiertes Screening-Instrument, das zur statusdiagnostischen Erfassung verhaltensbezogener Auffälligkeiten und Stärken eingesetzt wird (Altendorfer-Kling et al., 2007). Für die Erstellung des Prototyps wurden die im SDQ vorhandenen Verhaltensbereiche „Verhaltensprobleme“, „Emotionale Probleme“ und „Hyperaktivität“ adaptiert sowie die zusätzliche Dimension „Schulbezogenes Verhalten“ ausgearbeitet (Sauerland, 2018). Die Begründung für die Erweiterung um den letztgenannten Bereich liegt in der Möglichkeit, mit dieser Dimension auch das Arbeitsverhalten von SuS zu erfassen, welches insbesondere im schulischen Kontext eine relevante Rolle spielt (Voß & Gebhardt, 2017). Im Rahmen einer empirischen Masterarbeit wurde die prototypische Version des DBR-MIS von Sauerland (2018) weiterentwickelt und im schulischen Setting überprüft. Dabei wurde die Ursprungsversion von Gebhardt et al. (2017) in Anlehnung an zwei weitere Verhaltensbereiche des SDQ um die Dimensionen „Prosoziales Verhalten“ und „Verhaltensprobleme mit Gleichaltrigen“ erweitert (Sauerland, 2018). Die Auswahl bzw. Generierung der konkreten Items für den DBR-MIS erfolgte nach dem nomothetischen Ansatz, da sich sowohl Gebhardt et al. (2017) als auch Sauerland (2018) bei der Konstruktion und Erweiterung der Ratingskala überwiegend an den evaluierten sowie bewährten Dimensionen und Items des SDQ orientierten. Items stellen beobachtbare Variablen dar und sind damit die Indikatoren, mittels derer die Ausprägung von übergeordneten theoretischen Konstrukten, die im Falle des DBR-MIS in Form von Verhaltensdimensionen operationalisiert sind, festgestellt werden soll (Döring & Bortz, 2016). Die Verhaltensdimensionen stellen damit die zu messenden latenten Variablen dar. Im Rahmen der Untersuchungen von Sauerland (2018) wurde

die Auswahl der insgesamt 19 Items für die Verhaltensdimensionen des DBR-MIS sowie auch deren Formulierung weiterführend geprüft und überarbeitet. Die aus der Forschungsarbeit von Sauerland (2018) herausgehende Version des DBR-MIS, deren Testgüte anhand einer Pilotierungsstudie von Hisker (2018) untersucht wurde, umfasst die sechs Verhaltensdimensionen „Schulbezogenes Verhalten“ (SV), „Verhaltensprobleme“ (VP), „Hyperaktivität“ (HY), „Emotionale Probleme“ (EP), „Prosoziales Verhalten“ (PS) und „Verhaltensprobleme mit Gleichaltrigen“ (VPG). Für die Veröffentlichung des Instruments auf der Lernplattform LEVUMI im Herbst 2018 wurde die äußere Struktur des DBR-MIS hinsichtlich der Zuordnung der sechs Verhaltensdimensionen zu den übergeordneten Dimensionen „Externalisierendes Verhalten“ (EXT), „Internalisierendes Verhalten“ (INT) und „Positives Schulverhalten“ (PSV), die eine höhere theoretische Konstruktebene repräsentieren, verändert. Diese Einteilung in zwei Dimensionsebenen erfolgte ebenfalls in Anlehnung an den SDQ, der sowohl auf einer höheren Ebene über drei Dimension als auch auf niedriger Ebene über fünf Subdimensionen ausgewertet werden kann. Zudem erfolgten einige begriffliche Anpassungen in Orientierung an die Begrifflichkeiten und Definitionen des Klassifikationssystems ICD-10, welches von der Weltgesundheitsorganisation (WHO) herausgegeben wird (Schurig et al., 2019).

Beim DBR-MIS handelt es sich um eine Multi-Item Skala. Die einzelnen Verhaltensdimensionen werden also über mehrere Items erfasst. Dabei wurden für jede Verhaltensdimension drei bis vier beobachtbare Verhaltensweisen operationalisiert, die als Indikatoren für die Feststellung der jeweiligen Ausprägung der latenten Variablen dienen (Schurig et al., 2019). Angesichts des Ziels, „über multiple Indikatoren (Skalen-Items) ein theoretisches Konzept präziser zu erfassen als dies mit einem Einzelindikator […] möglich wäre“ (Döring & Bortz, 2016, S. 267), entspricht die dimensionsbezogene Gruppierung der Items des DBR-MIS einer psychometrischen Skala (ebd.). Die aktuelle Version der Ratingskala, die die Grundlage für diese Forschungsarbeit bildet, umfasst die sechs psychometrischen Subskalen „Störendes und auflehnendes Verhalten“ (SAV), „Verhaltensprobleme beim Lernen“ (VPL), „Depressives und ängstliches Verhalten“ (DAV), „Probleme in sozialen Interaktionen“ (PSI), „Schulbezogenes Verhalten“ (SV) und „Prosoziales Verhalten“ (PS). Die Skalen SAV, VPL, DAV, PSI und PS bilden die Verhaltensdimensionen ab, die in Anlehnung an die im SDQ

vorhandenen Verhaltensbereiche, welche in Kapitel 2.2.1 näher erläutert wurden, von Gebhardt et al. (2017) und Sauerland (2018) für die Vorgängerversion des DBR-MIS adaptiert wurden. Im Folgenden werden die Skalen kurz näher vorgestellt, um die konkrete Operationalisierung der Verhaltensdimensionen sowie die Verbindungen zu den Verhaltensbereichen des SDQ bzw. den begrifflich differierenden Skalen der Vorgängerversion zu erläutern.

Die Skala SAV erfasst oppositionell aufsässige Verhaltensprobleme und entspricht damit hinsichtlich der Vorgängerversion des DBR-MIS von Sauerland (2018) der Dimension bzw. dem Verhaltensbereich VP (Verhaltensprobleme). Für diese Skala wurden drei Items operationalisiert, welche wutbezogene, regelbrechende und provozierende Verhaltensweisen umfassen. Die Skala VPL ist mit der ursprünglichen Dimension HY (Hyperaktivität) gleichzusetzen und erfasst Verhaltensprobleme, die sich in konkreten Lernsituationen zeigen. Für diese latente Variable wurden drei beobachtbare Verhaltensweisen als Indikatoren ausgewählt, die insbesondere Symptome wie Unaufmerksamkeit und Hyperaktivität aufgreifen. Anhand der drei Items der Skala DAV wird die Ausprägung von ängstlichem und depressivem Verhalten abgefragt. Die Skala entspricht damit der Dimension EP (Emotionale Probleme) der Vorgängerversion. Die ursprüngliche Dimension VPG (Verhaltensprobleme mit Gleichaltrigen) ist mit der begrifflich angepassten und aktuellen Skala PSI gleichzusetzen. Für diese Skala wurden Verhaltensweisen als Indikatoren ausgewählt, die Probleme im sozialen Kontaktverhalten aufzeigen. Konkret wird abgefragt, inwiefern ein Kind oft allein oder nur mit Erwachsenen spielt und arbeitet sowie ob es von anderen Kindern abgelehnt und gehänselt wird. Während die bisher vorgestellten Skalen bzw. Verhaltensdimensionen eher problemorientiert sind, stellen die Skalen PS und SV Stärkedimensionen dar. Im Vergleich zur Vorgängerversion von Sauerland (2018) wurden die beiden Skalen auch begrifflich nicht verändert. Die Skala SV, die als einzige Skala nicht in Anlehnung an einen Verhaltensbereich des SDQ ausgewählt wurde, erfasst positives Verhalten bezogen auf Lern- und Unterrichtssituationen. Die vier Verhaltensweisen, die jeweils als Indikator dienen, sind das Melden im Unterricht, die Einhaltung von Gesprächsregeln, die konzentrierte Aufgabenbearbeitung und eine ruhige Mitarbeit. Für die letzte Skala PS wurden drei Items operationalisiert, mittels derer die Ausprägung sozialer Kompetenzen festgestellt werden soll. Dabei werden

positive Verhaltensweisen wie das Anbieten von Hilfe und das Kooperieren mit anderen Kindern abgefragt, die in konkreten sozialen Interaktionen beobachtet werden können (Sauerland, 2018; Schurig et al., 2019). In den später folgenden statistischen Analysen dieser Arbeit werden die jeweiligen Items der Skalen durch Kürzel repräsentiert. Diese Kürzel umfassen zum einen den Namen der Skala, zu der ein Item zugeordnet ist, sowie eine Zahl, die der Nummer eines Items im DBR-MIS entspricht. Die Items der Skala SAV werden demnach durch die Kürzel SAV01, SAV02 und SAV03 repräsentiert.

Wie zuvor angerissen, sind die sechs beschriebenen Verhaltensdimensionen des DBR-MIS den drei übergeordneten Dimensionen EXT, INT sowie PSV zugeordnet und können demnach als Subskalen oder auch als Dimensionen zweiter Ebene betrachtet werden. Die Dimensionen SAV und VPL bilden die Subskalen für die Dimension EXT, da sie vor allem ausagierende Verhaltensweisen fokussieren, welche nach Myschker und Stein (2018) der Klasse der externalisierenden Verhaltensstörungen zuzuordnen sind. Die Verhaltensdimensionen DAV und PSI erfassen dagegen emotionale Probleme wie Depressionen und Ängste, die mit einer sozialen Isolation oder sozial unreifem Verhalten einhergehen können. Derartige Verhaltensweisen werden als internalisierend bezeichnet, weshalb die Skalen DAV und PSI der Dimension INT auf erster Ebene zugeordnet sind (ebd.). Die beiden Stärkedimensionen SV und PS bilden die Subskalen für die übergeordnete Dimension PSV.

Die Skalierung des DBR-MIS entspricht einer siebenstufigen Häufigkeitsskala. Die Abstufung der Ratingskala wird durch die Ziffern 1 bis 7 als numerische Marken dargestellt. Zusätzlich sind für die Skalenendpunkte 1 und 7 die verbalen Marken „Nie“ (1) und „Immer“ (7) angegeben. Die Ratingskala bildet also in einer graduellen Abstufung die Auftretenshäufigkeit einer Verhaltensweise ab. Alle Items sind unidirektional formuliert, d. h. sie sind als positive Items alle in dieselbe Richtung gepolt und sprechen für eine starke Ausprägung des übergeordneten Konstrukts (Döring & Bortz, 2016). Im Hinblick auf die Einteilung in Problem- und Stärkedimensionen ist zu beachten, dass hohe Werte auf den Skalen SAV, VPL, DAV und PSI für eine hohe Auftretenshäufigkeit von unangemessenen bzw. problematischen Verhaltensweisen stehen, während hohe Werte auf den Skalen SV und PS einem hohen Ausprägungsgrad positiver Verhaltensweisen entsprechen. Die Antwortkategorien sind durch die

Skalierung des DBR-MIS klar festgelegt. Die Möglichkeit eigene Antworten und weiterführende Kommentare einzubringen, ist nicht vorgesehen. Es liegt also ein gebundenes Antwortformat vor (Bühner, 2011). Eine aktuelle Version des DBR-MIS, auf die sich die vorangegangenen Beschreibungen beziehen, ist dem Anhang dieser Arbeit beigefügt (Anhang A).

Der DBR-MIS kann zum einen für die Überprüfung eines Verdachts von Verhaltensproblemen und zum anderen für die Überprüfung von Interventionswirkungen eingesetzt werden. Eine Altersbeschränkung für die Verwendung des Instruments liegt nicht vor. In Bezug auf den Umgang mit Daten, die mit dem DBR-MIS erfasst werden, wird auf der Onlineplattform LEVUMI eine computergestützte Auswertung angeboten. Dabei werden in Form einer Auswertungstabelle für jeden Messzeitpunkt, die vergebenen Punkte für die jeweiligen Items innerhalb der sechs Verhaltensdimensionen angezeigt. Zudem werden pro Skala die Mittelwerte berechnet und ebenfalls in der Tabelle dargelegt. Anhand dieser Auswertungstabelle können Lehrkräfte sowohl die Werte der verschiedenen Dimensionen innerhalb eines Messzeitpunktes vergleichen und herausfinden, ob ein Schüler oder eine Schülerin in einer bestimmten Dimension Probleme aufweist, als auch Veränderungen über die Messzeitpunkte beobachten (Schurig et al., 2019). Die Auswertung wie sie im Rahmen von LEVUMI konkret angeboten wird, findet demnach auf der zweiten Dimensionsebene, d. h. auf der Ebene der sechs Subskalen, statt.

4.3 Stichprobenbeschreibung

Da eine Erhebung von aktuellen Daten für die vorliegende Arbeit aufgrund der Corona-Pandemie nicht durchgeführt werden konnte, wird für die Bearbeitung der Forschungsfragen auf einen bereits bestehenden Datensatz zurückgegriffen. Dieser Datensatz wurde 2018 im Rahmen einer Erhebung mit dem DBR-MIS in einem Zeitraum von etwa vier Monaten (Januar bis April) zusammengetragen. Die ursprüngliche Stichprobe umfasst Daten von 219 SuS unterschiedlicher Schulformen, die über fünf Messzeitpunkte erhoben wurden. Da das Forschungsinteresse dieser Arbeit allerdings nicht auf der Betrachtung des Verhaltensverlaufs der SuS über die Messzeitpunkte, sondern auf der Berechnung von Schwellenwerten im Sinne einer statusdiagnostischen Normierung liegt, werden nur die Daten von einem Messzeitpunkt

benötigt. Dahingehend werden für die nachfolgenden Analysen und Untersuchungen im Rahmen der vorliegenden Arbeit, lediglich die Daten des ersten Messzeitpunkts von 209 SuS verwendet. Die reduzierte Stichprobenzahl (*N* = 209) ergibt sich durch den Ausschluss einiger Daten im Zuge der Datenbereinigung. Für drei SuS lagen für den ersten Messzeitpunkt keine Daten vor, da sie an diesem Erhebungstag nicht anwesend waren. Außerdem wurde der Datensatz um die Daten von sieben SuS mit dem FS ESE gekürzt. Bei SuS mit dem FS ESE liegen bereits diagnostizierte Beeinträchtigungen in den Bereichen des emotionalen und sozialen Erlebens und Verhaltens vor. Im Vergleich zu SuS ohne diesen Förderschwerpunkt, können für betroffene Kinder und Jugendliche höhere Werte in den Problemdimensionen des DBR-MIS angenommen werden. Eine Angabe von Schwellenwerten, anhand derer die Werte von Kindern und Jugendlichen mit und ohne den FS ESE gleichermaßen eingeschätzt werden, erscheint dahingehend nicht tragbar. Da aufgrund der geringen Stichprobenanzahl von Kindern und Jugendlichen mit dem FS ESE (*n* = 7) allerdings keine förderschwerpunktspezifischen Effekte untersucht werden können und damit auch eine Berechnung förderschwerpunktspezifischer Schwellenwerte nicht möglich ist, werden die Daten von der Untersuchung ausgeschlossen. Insgesamt weisen 26 weitere SuS einen sonderpädagogischen Förderbedarf auf. Davon hat ein Schüler den Förderschwerpunkt Sprache, zehn SuS haben den Förderschwerpunkt Lernen und dreizehn SuS den Förderschwerpunkt geistige Entwicklung. Zwei weitere SuS haben sowohl im Förderschwerpunkt Lernen als auch im Förderschwerpunkt geistige Entwicklung einen zugewiesenen Förderbedarf. Die erhobenen Daten dieser Kinder und Jugendlichen werden nicht ausgeschlossen, da die angegebenen Förderschwerpunkte nicht zwingend mit ausgeprägten Verhaltensauffälligkeiten einhergehen.

Alle Daten wurden über 41 Lehrkräfte von insgesamt zehn Schulen der unterschiedlichen Schulformen Primarschule, Sekundarschule und der Schule für Kranke erhoben. Für die Schule für Kranke liegen nur Daten von 14 SuS vor, während SuS aus der Primarstufe (*n* = 101) und der Sekundarstufe (*n* = 94) deutlich höher vertreten sind. In der Betrachtung der allgemeinen Geschlechtsverteilung fällt auf, dass die Anzahl der Jungen (*n* = 148) in der Gesamtstichprobe (*N* = 209) deutlich höher ist als die Anzahl der Mädchen (*n*= 61). Das Alter der beobachteten SuS liegt zwischen sechs und 18 Jahren. Die Stichrobe umfasst damit eine Alterspanne von etwa zwölf

Jahren. In Tabelle 1 wird die genaue Alters- und Geschlechtsverteilung für die Gesamtstichprobe nach Schulformen getrennt aufgeschlüsselt. Für die übersichtliche altersspezifische Darstellung wurden drei Altersgruppen eingeteilt, die jeweils etwa eine gleichgroße Alterspanne umfassen.

Tabelle 1: Altersgruppen- und Geschlechtsverteilung für die Gesamtstichprobe getrennt nach Schulformen

Schulform	**Geschlecht**		**Altersgruppe**				**Gesamtsumme**
			6-9	10-13	14-18	Unbekannt	
Primarstufe	Jungen	Anzahl	49	23	0	0	72
		%	68	32	0	0	100
	Mädchen	Anzahl	26	3	0	0	29
		%	90	10	0	0	100
	Gesamt	Anzahl	75	26	0	0	101
		%	74	27	0	0	100
Sekundarschulen	Jungen	Anzahl	0	34	31	0	65
		%	0	52	48	0	100
	Mädchen	Anzahl	0	16	12	1	29
		%	0	55	42	3	100
	Gesamt	Anzahl	0	50	43	1	94
		%	0	53	46	1	100
Schule für Kranke	Jungen	Anzahl	0	1	10	0	11
		%	0	9	91	0	100
	Mädchen	Anzahl	0	0	2	1	3
		%	0	0	67	33	100
	Gesamt	Anzahl	0	1	12	1	14
		%	0	7	86	7	100
Alle Schulformen	Jungen	Anzahl	49	58	41	0	148
		%	33	39	28	0	100
	Mädchen	Anzahl	26	19	14	2	61
		%	42	31	23	3	100
	Gesamt	Anzahl	75	77	55	2	209
		%	36	37	26	1	100

4.4 Auswertungsmethode

Für die Auswertung der erhobenen Daten wird das Statistikprogramm *jamovi* verwendet. Dieses ermöglicht eine automatisierte Auswertung der Daten in Form von verschiedenen statistischen Analysen sowie die Visualisierung der Daten und Analyseergebnisse in unterschiedlichen Darstellungsformen. Die jeweilig durchzuführenden Analyseschritte und Testverfahren werden in Hinblick auf die Forschungsfragen der vorliegenden Arbeit ausgewählt.

Die erste Forschungsfrage bezieht sich auf die beiden vorgegebenen Konstruktebenen des multidimensionalen DBR-MIS, die jeweils über ein spezifisches empirisches Messmodell operationalisiert sind. Zum einen liegt über die Einteilung der sechs

Verhaltensdimensionen SAV, VPL, DAV, PSI, PS und SV, die jeweils über verschiedene Indikatoren operationalisiert sind, eine sechs-faktorielle Struktur vor. Dieses Sechs-Faktoren-Modell bildet die Grundlage für die Auswertung des DBR-MIS auf der Onlineplattform LEVUMI und kann daher als bisher vorherrschendes Messmodell bezeichnet werden. Zum anderen bilden jeweils zwei der sechs Verhaltensdimensionen, die übergeordneten Dimensionen EXT, INT und PSV ab. Diese Dimensionen erster Ebene sind also über die Kombination der Indikatoren zweier Dimensionen der zweiten Ebene operationalisiert, was einer drei-faktoriellen Struktur entspricht. Sowohl das Drei-Faktoren-Modell als auch das Sechs-Faktoren-Modell stellen als Grundlage für die Bestimmung von Schwellenwerten potenziell sinnvolle Messmodelle dar. Die Auswahl eines der Messmodelle bestimmt, auf welcher der zwei möglichen Konstruktebenen die Schwellenwerte für den DBR-MIS im Rahmen der vorliegenden Arbeit berechnet werden. Dahingehend muss festgestellt werden, welches Messmodell am besten mit den erhobenen Stichprobendaten vereinbar ist (Döring & Bortz, 2016). Um diese Frage zu beantworten, wurden die beiden Messmodelle jeweils mit Hilfe konfirmatorischer Faktorenanalysen überprüft.

Die konfirmatorische Faktorenanalyse wird zur Untersuchung der Operationalisierung theoretischer Konstrukte in Form von empirischen Messmodellen eingesetzt. Dabei wird über die Schätzung der Faktorladungen überprüft, inwiefern die jeweiligen Indikatoren durch den übergeordneten Faktor erklärt werden. Wenn der Wert eines Faktors um 1 steigt, dann steigt der Wert des Indikators um die Faktorladung. Betraglich hohe Faktorladungen deuten dabei auf einen engeren Messzusammenhang zwischen Indikator und Faktor hin als betraglich niedrigere Faktorladungen. Je näher die Faktorladung an den Werten 1 oder -1 liegt, desto stärker ist der Einfluss des Faktors auf den Indikator. Ladungen nahe 0 stehen daher für einen geringen Einfluss. Weiterhin wird geprüft, ob der durch die Faktorladung angegebene Zusammenhang signifikant ist. Bei einem p-Wert kleiner .05 wird i. d. R. angenommen, dass eine Korrelation besteht und ein Indikator damit für die Abbildung der latenten Variablen geeignet ist. Im Anschluss an die generelle Überprüfung der Faktorenstruktur der beiden vorgegebenen Messmodelle wird zudem deren Modellgüte untersucht. Diese Untersuchung erfolgt ebenfalls im Rahmen der konfirmatorischen Faktorenanalyse mittels der Maximum-Likelihood (ML)-Schätzung (Döring & Bortz, 2016). Die Modellgüte,

auch Modell-Fit genannt, „gibt an, wie gut ein aufgestelltes Modell den empirisch beobachteten Daten entspricht“ (ebd., S. 967). Als Grundlage für die Beurteilung der Modelle anhand verschiedener Modellgütekriterien wurde dabei die ML-Schätzmethode angenommen. Diese Methode ist ein Verfahren zur Schätzung von Parametern der Grundgesamtheit aus der Stichprobe. Dabei werden die Schätzwerte ausgewählt, unter denen die Wahrscheinlichkeit der beobachteten Stichprobenrealisation am höchsten ist. Voraussetzungen für die Annahme der ML-Schätzung sind z. B. eine multivariate Normalverteilung der Indikatorvariablen und ein angemessener Stichprobenumfang. Sind diese Voraussetzungen bei einer Stichprobe nicht gegeben, kann es bei der Überprüfung der Modellgüte zu Verzerrungen bei verschiedenen Tests kommen (ebd.). Die vergleichende Beurteilung der Modellgüte im Rahmen dieser Arbeit erfolgt anhand des x^2-Tests, den klassischen Fit-Indizes CFI (Comparative Fit Index), TLI (Tucker-Lewis Index) und RMSEA (Root Mean Square Error of Approximation) sowie den zusätzlichen Vergleichskriterien AIC (Aikake-Information-Criterion) und BIC (Bayesian-Information-Criterion).

Der x^2-Test ist ein inferenzstatistischer Test, mit dem geprüft wird, ob sich die modelltheoretische Kovarianzmatrix, die aus den berechneten Modellparametern reproduziert wird, signifikant von der empirischen Kovarianzmatrix, die auf der Basis der Stichprobendaten berechnet wird, unterscheidet. Ein nicht-signifikanter x^2-Wert ($p > .05$) bezogen auf die Freiheitsgrade (*df*), die die Anzahl frei veränderbarer Werte angeben, weist darauf hin, dass das analysierte Messmodell zu den erhobenen Daten passt (Döring & Bortz, 2016). Der Schwellenwert für einen guten Modell-Fit liegt für den x^2-Test bei $x^2/df \leq 3$ (Weiber & Mühlhaus, 2014). Bei der Betrachtung von x^2-Werten sollte beachtet werden, dass diese bei nicht normalverteilten Daten größer ausfallen, was dazu führen kann, dass auch wahre Modelle abgelehnt werden. Insgesamt wird der x^2-Test stark vom Stichprobenumfang beeinflusst, weshalb zusätzlich deskriptive Gütekriterien wie der CFI und der RMSEA berücksichtigt werden, die „eine graduelle Beurteilung der Abweichung zwischen Modell und Daten“ (Döring & Bortz, 2016, S.967) ermöglichen. Der RMSEA, der als Gütemaß die Passung zwischen Modell und Daten hinsichtlich der Freiheitsgrade und der Stichprobengröße bestimmt, sollte möglichst gering sein. Werte ≤ .05 sprechen für einen guten Modell-Fit. Sowohl der CFI als auch der TLI beruhen auf dem Vergleich des untersuchten

Modells mit der Annahme, dass keine Korrelationen zwischen den Variablen bestehen (Unabhängigkeitsmodell). Dabei bestimmen sie, inwieweit das untersuchte Modell für die Erklärung der erhobenen Daten besser geeignet ist als das Unabhängigkeitsmodell (ebd.). Hu und Bentler (1999) geben an, dass für beide Gütemaße Werte von ≥ .95 auf einen guten Modell-Fit weisen. Im Hinblick auf das Ziel der Modellselektion werden zudem die Werte des AIC und des BIC herangezogen, die dazu dienen verschiedene Modellkandidaten zu vergleichen. Für diese beiden Kriterien gibt es keine Schwellenwerte, da sie keine absoluten Gütemaße darstellen. Beim Modellvergleich gilt vielmehr, dass das Modell mit den niedrigsten AIC und dem niedrigsten BIC bevorzugt wird (Gehrke, 2019).

Nach der Auswahl des Messmodells, das angesichts der beschriebenen Gütekriterien am besten zu den erhobenen Stichprobendaten passt, wird daran anschließend die Reliabilität der Ratingskala überprüft. Im Hinblick auf das ausgewählte Modell und die damit einhergehende Skalenstruktur wird also die Messgenauigkeit bestimmt, mit der die Ausprägungen der Verhaltensdimensionen gemessen wird. Dafür werden die internen Konsistenzen der entweder drei oder sechs Skalen des DBR-MIS nach Cronbachs α berechnet. Die berechneten Werte ermöglichen Aussagen darüber, inwiefern die Items, die eine Skala abbilden, widerspruchsfreie und eindimensionale Indikatoren für die jeweilige Skala bzw. Verhaltensdimension darstellen. Cronbachs α ist damit ein Maß für die Korrelation von Items untereinander und beantwortet die Frage, inwieweit diese Items dasselbe messen (Döring & Bortz, 2016). Der Wert α ≥ .70 gilt dabei als Mindestmaß für eine akzeptable Reliabilität. Damit geben Werte von α ≥ .80 eine gute Reliabilität an, während Werte von α < .50 auf eine nicht akzeptable Reliabilität hinweisen (Himme, 2007). Im Zuge der Prüfung der internen Konsistenz wird zudem die Trennschärfe der Items einer Skala betrachtet, die angibt, „wie gut ein einzelnes Item das Zielkonstrukt des Tests misst“ (Döring & Bortz, 2016, S. 478). Berechnet wird die Trennschärfe eines Items über die Korrelation des Messwertes dieses Items mit dem Summenwert der übrigen Skalenitems. Kennwerte von r_{it} > .50 weisen auf eine hohe Trennschärfe. Dagegen werden Werte zwischen .30 und .50 als mittelmäßig eingeordnet (ebd.). Ähneln sich die Trennschärfen der Items einer Skala stark, ist das ein Hinweis darauf, dass das Testinstrument eine geeignete Skalierung aufweist und die damit berechneten empirischen Testwerte für die

Interpretation einer Merkmalsausprägung herangezogen werden können (Bühner, 2011; Casale et al., 2015b).

Mit der Auswahl eines Messmodells und dessen psychometrischer Überprüfung wird die Frage, auf welcher Konstruktebene des DBR-MIS eine Schwellenwertberechnung sinnvoll ist, beantwortet. Aufbauend darauf, erfolgt dann in Bezug zur zweiten Forschungsfrage, die konkrete Bestimmung von zunächst geschlechts- und altersunspezifischen Schwellenwerten. Für diese Bestimmung konnte sich im Rahmen dieser Arbeit an der Normierung der deutschen Version des SDQ von Lohbeck et al. (2015) orientiert werden. Die Grundlage für diese Normierung des SDQ bildete die Verteilung der Rohwerte pro Skala und die damit einhergehende Zuweisung der einzelnen Ausprägungsgrade zu einem Prozentrang (Lohbeck et al., 2015). In Anlehnung an die von Goodman (1997) empfohlenen Zielvorgaben für die Bestimmung von SDQ-Grenzwerten, wurde für die Einteilung der Werte in die Kategorien *unauffällig*, *grenzwertig* und *auffällig* ein Anteil von etwa 80–10–10% angenommen. In Annäherung an diese Zielvorgaben wurden dann Grenzwerte für die einzelnen Skalen empfohlen, die dem Ziel einer hohen Sensitivität bei der Identifikation von Problemfällen entsprechen (Lohbeck et al., 2015). Für die Bestimmung von Schwellenwerten für die Skalen des DBR-MIS wurden im Hinblick auf die von LEVUMI vorgegebene Auswertungsmethode für den gesamten Datensatz die Skalenmittelwerte pro Verhaltensdimension und pro Schüler oder Schülerin berechnet. Mittelwerte der Skalen SAV, VPL, DAV, PSI und PS, die mehr als zwei Nachkommastellen umfassten, wurden bis auf eine Nachkommastelle abgeschnitten. Angesichts fehlender Itemwerte wurde nur dann eine Mittelwertberechnung durchgeführt, wenn für mehr als die Hälfte der Items einer Skala Werte vorlagen. Anschließend wurde im Sinne einer statusdiagnostischen Normierung die Verteilung aller Mittelwerte pro Skala betrachtet, wobei jedem Ausprägungsgrad ein Prozentrang, der die Auftretenshäufigkeit in der Gesamtstichprobe widerspiegelt, zugeordnet wurde (Bundschuh & Winkler, 2019). Für die Interpretation der Prozentränge in Bezug auf die Festlegung von Schwellenwerten für die drei Kategorien *auffällig*, *grenzwertig* und *unauffällig,* wird sich an den in Kapitel 2.2.2 beschrieben Prävalenzraten von Verhaltensauffälligkeiten und psychischen Störungen bei Kindern und Jugendlichen orientiert. In den Übersichtsarbeiten von Ihle und Esser (2002) sowie Barkmann und Schulte-Markwort (2004) berechneten die Autoren

durchschnittliche Gesamtprävalenzen von 17-18%. Auch in der aktuellen Erhebung der KiGGS-Studie (KiGGS-Welle 2) wurde für die Prävalenz von psychischen Störungen insgesamt ein Wert von etwa 17% festgestellt (Klipker et al., 2018). Für einzelne Verhaltensbereiche konnten Ihle und Esser (2002), im Vergleich der Ergebnisse mehrerer Studien, durchschnittliche Prävalenzen von 1-10% bestimmen. Zusammenfassend betrachtet, entsprechen diese Prävalenzangaben annäherungsweise den von Goodman (1997) vorgeschlagenen Zielvorgaben von 80-10-10%. Die Bestimmung der Schwellenwerte anhand der Mittelwertverteilung der Gesamtstichprobe pro Skala erfolgt in dieser Arbeit daher in Annäherung an diese Zielvorgaben. Den Stichprobendaten wird also ein Anteil von 10% *auffälligen* Werten und ein Anteil von 80% *unauffälligen* Werten pro Skala zugrunde gelegt, während die dazwischenliegenden Mittelwerte als *grenzwertig* eingestuft werden. Im Hinblick auf die Verteilung der Mittelwerte einer Stichprobe kann es vorkommen, dass die empirisch beobachtbaren Anteile entweder über oder unter einer Zielvorgabe liegen (Lohbeck et al., 2015). In solchen Fällen wird für die Bestimmung der Schwellenwerte für die Kategorien *auffällig* und *grenzwertig* in dieser Arbeit der Anteil ausgewählt, der näher an der vorgegebenen Zielvorgabe von jeweils 10% liegt. Dieses Vorgehen ist durch das Forschungsinteresse in Bezug auf das primäre Einsatzziel des DBR-MIS als verlaufsdiagnostisches Instrument begründet, das auf der Bestimmung von Schwellenwerten als Interpretationsorientierung für individuelle Testwerte und nicht als sensitive Identifikationskriterien für auffälliges Verhalten liegt. Entscheidend ist außerdem, dass bei der Bestimmung der Schwellenwerte die Ausrichtungen der Skalen bezogen auf das übergeordnete theoretische Konstrukt berücksichtigt werden. So setzt die Bestimmung von *auffälligen* Anteilen bei den Problemskalen SAV, VPL, DAV und PSI an den hohen Mittelwerten an, während bei den Stärkeskalen SV und PS niedrige Mittelwerte als problematisch gelten und damit den Ansatzpunkt für die Kategorie *auffällig* bilden.

Die Überprüfung von Geschlechts- und Altersunterschieden in den Skalenmittelwerten erfolgte im Rahmen der dritten Forschungsfrage. Für die Untersuchung, ob das Geschlecht einen Effekt auf die Konstrukte des DBR-MIS hat, wurde der Mann Whitney U-Test für unabhängige Stichproben durchgeführt. Dieser stellt eine Alternative zum t-Test dar, wenn die untersuchten Daten nicht normalverteilt sind, was bei der vorliegenden Stichprobe der Fall war (Rasch, Friese, Hofmann & Naumann, 2006b).

Mit dem U-Test kann ähnlich wie bei einem t-Test analysiert werden, „ob zwischen zwei unabhängigen Gruppen ein signifikanter Unterschied besteht“ (ebd., S.144). Der Test gibt dabei an, mit welcher Wahrscheinlichkeit die gefundenen Mittelwertsunterschiede auftreten. Werte von $p < .05$ gelten als signifikant, was bedeutet, dass die Unterschiede in den Daten zwischen den beiden Stichprobengruppen mit großer Wahrscheinlichkeit nicht zufällig zustande kommen. Hinsichtlich weiterer gängiger Signifikanzniveaus gilt $p < .01$ als sehr signifikant und $p < .001$ als höchst signifikant (Rasch et al., 2006a). Zur Beantwortung der Frage, wie groß der Unterschiedseffekt bezüglich der Variablen Geschlecht ist, wird zusätzlich der *Cohen‘s d* als Maß für die Effektgröße herangezogen. Dabei steht $d = 0.2$ für einen kleinen Effekt, $d = 0.5$ für einen mittleren Effekt und $d = 0.8$ für einen großen Effekt (Döring & Bortz, 2016).

Für die Überprüfung des Effekts der Variablen Alter auf die Konstrukte des DBR-MIS wurde aufgrund der hohen Altersheterogenität die Einteilung von drei altershomogeneren Gruppen, wie sie zur Beschreibung der Stichprobe in Kapitel 4.3 vorgenommen wurde, beibehalten. Ob signifikante Unterschiede zwischen den Skalenmittelwerten der verschiedenen Altersgruppen vorliegen, wurde mittels einer einfaktoriellen Varianzanalyse („one-way ANOVA“) für alle sechs Skalen untersucht. Die Varianzanalyse ermöglicht den simultanen Vergleich der Mittelwerte mehrerer Gruppen, indem die Varianzen der Gruppen betrachtet werden (Rasch et al., 2006b). Damit kann bezüglich der drei Altersgruppen geprüft werden, „ob sich mindestens zwei der drei Gruppen überzufällig voneinander unterscheiden“ (Döring & Bortz, 2016, S. 709). Die Varianzanalyse fokussiert neben den Varianzen zwischen den Gruppen auch die Varianzen innerhalb der Gruppen, weshalb zunächst überprüft wird, inwiefern diese in allen Gruppen gleich sind (Rasch et al., 2006b). Dafür wird der Levene-Test verwendet. Zeigt dieser signifikante Unterschiede ($p < .05$) zwischen den Varianzen der Gruppen, so erfolgt die Varianzanalyse in *jamovi* über den Welch-Test, der als Korrekturverfahren bei ungleichen Varianzen herangezogen wird (Dorsch, Wirtz & Strohmer, 2014). Bestehen keine signifikanten Unterschiede in den Varianzen der Gruppen, kann von einer Varianzhomogenität ausgegangen werden (Rasch et al., 2006b). In *jamovi* erfolgt die einfaktorielle Varianzanalyse in diesem Fall mittels des Fisher-Tests. Im nächsten Schritt werden dann die Ergebnisse des jeweilig verwendeten Tests pro Skala betrachtet. Ergibt die Varianzanalyse für eine der Skalen ein

signifikantes Ergebnis ($p < .005$), besteht bei den Mittelwerten mindestens einer Gruppe ein statistisch bedeutsamer Unterschied zu den anderen (ebd.).

Konnten anhand der Ergebnisse des U-Tests und der Varianzanalyse signifikante Unterschiede jeweils zwischen den Skalenmittelwerten von Mädchen und Jungen sowie den Skalenmittelwerten der drei Altersgruppen festgestellt werden, weist dies auf einen Einfluss der Variablen Geschlecht und Alter auf die Konstrukte des DBR-MIS. In diesem Fall werden in Anlehnung an das Vorgehen der allgemeinen Schwellenwertbestimmung zusätzlich geschlechts- und/ oder altersspezifische Schwellenwerte ermittelt.

5. Ergebnisse

In den folgenden Unterkapiteln werden die Forschungsergebnisse vorgestellt, die im Rahmen der zuvor beschriebenen statistischen Analyseschritte gewonnen wurden. Dafür werden sowohl von *jamovi* ausgegebene Grafiken als auch Ergebnistabellen verwendet und weitergehend beschrieben. Die Struktur der Ergebnisdarstellung erfolgt in Orientierung an die Reihenfolge der übergeordneten Forschungsfragen dieser Arbeit, wie sie in Kapitel 3 aufgeführt ist.

5.1 Faktorenstruktur und psychometrische Skalenkennwerte

Die Beantwortung der ersten Forschungsfrage erfolgt über die Auswahl eines der zwei möglichen empirischen Messmodelle, die der Bestimmung von Schwellenwerten für den DBR-MIS zugrunde gelegt werden können. Dafür wird zunächst jeweils die Faktorenstruktur der Messmodelle sowie deren Modellgüte überprüft und verglichen.

Tabelle 2: Faktorladung des Sechs-Faktoren-Modells des DBR-MIS

Faktor	Indikator	Standardisierte Faktorladung	*p*
SAV	SAV01	1,494	< .001
	SAV02	1,526	< .001
	SAV03	1,778	< .001
VPL	VPL04	1,306	< .001
	VPL05	1,552	< .001
	VPL06	1,604	< .001
DAV	DAV07	1,529	< .001
	DAV08	1,288	< .001
	DAV09	1,392	< .001
PSI	PSI10	0,859	< .001
	PSI11	1,330	< .001
	PSI12	0,664	< .001
SV	SV13	0,598	< .001
	SV14	1,365	< .001
	SV15	1,389	< .001
	SV16	1,539	< .001
PS	PS17	1,632	< .001
	PS18	1,659	< .001
	PS19	1,612	< .001

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, p = Signifikanzwert.

Tabelle 2 gibt einen Überblick über die Faktorladungen des Sechs-Faktoren-Modells, die mittels einer konfirmatorischen Faktorenanalyse berechnet wurden. Lediglich die Ladungen der Indikatoren PSI10 und PSI12 auf den Faktor PSI sowie die Ladung des Indikators SV13 auf den Faktor SV liegen unter dem Wert 1, was für einen geringeren Messzusammenhang dieser Indikatoren mit dem zugeordneten Faktor spricht. Die Faktorladungen der anderen Indikatoren liegen über dem Wert 1 und weisen damit auf einen starken Einfluss der jeweiligen Faktoren auf die einzelnen Indikatoren hin. Die Signifikanzprüfung sichert die über die Faktorladungen angegebenen Zusammenhänge zwischen Faktoren und Indikatoren zufallskritisch ab. Alle ermittelten p-Werte ($p < .001$) sprechen für eine hohe Signifikanz der untersuchten Faktor-Indikator-Zusammenhänge im Rahmen der sechs-faktoriellen Struktur.

Bei dem Drei-Faktoren-Modell sind die Faktorladungen der Indikatoren trotz unterschiedlicher Faktorenstruktur nahezu vergleichbar mit denen des Sechs-Faktoren-Modells (siehe Tabelle 3). So liegt der Großteil der Ladungen über dem Wert 1, was für einen starken Messzusammenhang zwischen den jeweiligen Indikatoren und den latenten Faktoren spricht. Die manifeste Variable SV13 lädt allerdings ähnlich wie beim Sechs-Faktoren-Modell mit einem Wert von nur 0,552 auf den latenten Faktor PSV. Dieser Wert ist deutlich geringer als die Ladungen der anderen fünf Indikatoren auf diesen Faktor und weist auf eine niedrigere Korrelation hin. Für die Indikatoren PSI10 und PSI11 zeigen die relativ niedrigen Faktorladungen von 0,646 und 0,638 ebenfalls einen schlechteren Messzusammenhang zwischen den beiden manifesten und der latenten Variablen an. Die insgesamt niedrigste Ladung mit einem Wert von 0,254 kann für den Indikator PSI12 festgestellt werden. Dieser Wert gibt an, dass der Einfluss des Faktors INT auf den Indikator PSI12 sehr schwach ist und damit nur ein geringer Messzusammenhang besteht. Für diesen Zusammenhang liegt allerdings der p-Wert ($p = 0.057$) im Rahmen der Signifikanzprüfung nur knapp über dem Signifikanzniveau. Die p-Werte ($< .001$) für die anderen untersuchten Messzusammenhänge weisen dagegen durchgehend auf eine hohe Signifikanz der angegebenen Korrelationen hin.

Tabelle 3: Faktorladungen des Drei-Faktoren-Modells des DBR-MIS

Faktor	Indikator	Standardisierte Faktorladung	p
EXT	SAV01	1,443	< .001
	SAV02	1,568	< .001
	SAV03	1,682	< .001
	VPL04	1,270	< .001
	VPL05	1,270	< .001
	VPL06	1,282	< .001
INT	DAV07	1,543	< .001
	DAV08	1,275	< .001
	DAV09	1,364	< .001
	PSI10	0,646	< .001
	PSI11	0,638	< .001
	PSI12	0,254	0.057
PSV	SV13	0,552	< .001
	SV14	1,159	< .001
	SV15	0,993	< .001
	SV16	1,165	< .001
	PS17	1,618	< .001
	PS18	1,587	< .001
	PS19	1,591	< .001

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, EXT = Externalisierendes Verhalten, INT = Internalisierendes Verhalten, PSV = Positives Schulverhalten, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, p = Signifikanzwert.

Die Überprüfung der faktoriellen Struktur des Sechs-Faktoren-Modells und des Drei-Faktoren-Modells anhand konfirmatorischer Faktorenanalysen zeigt, dass bei beiden Messmodellen überwiegend hohe Messzusammenhänge zwischen den Faktoren und den jeweiligen Indikatoren bestehen. Für die Auswahl eines Messmodells als Grundlage für die Schwellenwertbestimmung wird daher auch die Güte der beiden Modelle als Auswahlkriterium hinzugezogen. Dabei wird über den Vergleich von Werten verschiedener Gütemaße und Tests zur Bestimmung der Modellgüte beurteilt, welches Messmodell am besten zu den vorliegenden empirischen Stichprobendaten passt.

In Tabelle 4 sind die Ergebnisse des x^2-Tests für beide Messmodelle aufgeführt. Diese zeigen, dass das Sechs-Faktoren-Modell (x^2/df = 3,6) hinsichtlich des Schwellenwerts für einen guten Modell-Fit ($x^2/df \leq 3$) besser abschneidet als das Drei-Faktoren-Modell (x^2/df = 5,8). Die Signifikanzprüfung zeigt allerdings bei beiden Modellen einen signifikanten x^2-Wert (p < .001), was auf eine jeweilig schlechte Passung zwischen den Modellen und den erhobenen Daten hinweist.

Tabelle 4: Ergebnisse des x^2-Tests für das Sechs-Faktoren-Modell und das Drei-Faktoren-Modell des DBR-MIS

Messmodell	***x²***	***df***	***p***
Sechs-Faktoren-Modell	492	137	< .001
Drei-Faktoren-Modell	875	149	< .001

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, x^2 = x^2-Wert, p = Signifikanzwert.

Weitere Kriterien, die für den Vergleich der Güte beider Modelle herangezogen werden, sind die Fit-Indizes CFI, TLI und RSMEA sowie die Vergleichskriterien AIC und der BIC. Die Werte des Sechs-Faktoren-Modells (CFI = 0.846, TLI = 0.807, RMSEA = 0.111) liegen bei den Gütemaßen CFI und TLI knapp unter dem Schwellenwert für einen guten Modell-Fit (CFI/ TLI ≥ .95). Auch der RMSEA-Wert des Modells spricht angesichts einer perfekten Passung bei einem Wert von 0 nicht für einen guten Fit. Im Vergleich der Werte beider Messmodelle, die in Tabelle 5 dargestellt sind, liegen die Werte des Sechs-Faktoren-Modells für die Gütemaße CFI, TLI und RMSEA allerdings insgesamt näher an den Schwellenwerten für einen gute Modell-Fit als die Werte des Drei-Faktoren-Modells.

Tabelle 5: Fit-Indizes der Messmodelle des DBR-MIS

Messmodell	**CFI**	**TLI**	**RMSEA**	**AIC**	**BIC**
Sechs-Faktoren-Modell	0.846	0.807	0.111	12756	12997
Drei-Faktoren-Modell	0.684	0.637	0.153	13116	13316

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, CFI = Comparative, Fit Index, TLI = Tucker Lewis Index, RMSEA = Root Mean Square Error of Approximation, AIC = Aikake-Information-Criterion, BIS = Bayesian-Information-Criterion.

Zusammenfassend betrachtet, kann für keines der analysierten Messmodelle hinsichtlich der berechneten Werte in den Gütemaßen eine gute Passung mit den empirischen Daten nachgewiesen werden. Dies kann verschiedene Gründe haben, auf die in der Ergebnisdiskussion in Kapitel 6 eingegangen wird. Im Vergleich der beiden Messmodelle zeigt sich allerdings, dass die ermittelten Werte des Sechs-Faktoren-Modells insgesamt näher an den vorgegebenen Schwellenwerten für einen guten Modell-Fit liegen. Das Modell passt demnach besser zu den empirischen Daten als das Drei-Faktoren-Modell. Dies kann auch anhand der Betrachtung der AIC und BIC

Werte bestätigt werden. Diese fallen beim Sechs-Faktoren-Modell niedriger aus, was für eine Bevorzugung dieses Modells für die vorliegenden Daten im Modellvergleich spricht. Die Hypothese H1, die bezogen auf die erste Forschungsfrage dieser Arbeit formuliert wurde, kann damit anhand der vorgestellten Ergebnisse der konfirmatorischen Faktorenanalyse gestärkt werden. Als Grundlage für die Bestimmung der Schwellenwerte wird demnach das Sechs-Faktoren-Modell ausgewählt. Das bedeutet, dass jeweils Schwellenwerte für die sechs Skalen SAV, VPL, DAV, PSI, SV und PS bestimmt werden.

Bevor jedoch die Bestimmung der Schwellenwerte erfolgt, wird in Anlehnung an das ausgewählte Sechs-Faktoren-Modell zunächst die Reliabilität der Ratingskala eingeschätzt. Dafür werden die internen Konsistenzen der Skalen nach Cronbachs α betrachtet. Der Tabelle 6 ist zu entnehmen, dass die Werte der Skalen SAV, VPL, DAV und PS eine gute Reliabilität (α ≥ .80) und damit eine hohe interne Konsistenz dieser Skalen angeben. Auch für die Skala SV kann eine zufriedenstellende Reliabilität (α ≥ .70) festgestellt werden. Lediglich die Skala PSI weist einen fraglichen Reliabilitätswert auf, der nur knapp oberhalb des Niveaus für eine nicht akzeptable Reliabilität (α ≤ .50) liegt.

Tabelle 6: Interne Konsistenzen der Skalen des DBR-MIS nach Cronbachs α

Skala	**Cronbachs α**
SAV	0.887
VPL	0.806
DAV	0.841
PSI	0.593
SV	0.765
PS	0.921

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten.

Für differenzierte Aussagen darüber, wie gut ein einzelnes Item das übergeordnete Zielkonstrukt misst, wird im Zuge der Reliabilitätsanalyse auch die Trennschärfe (r_{it}) der einzelnen Items pro Skala betrachtet. Tabelle 7 bietet einen Überblick über die Itemtrennschärfen pro Skala. Insgesamt konnten für alle Items der Skalen SAV, VPL, DAV und PS als auch für die Items SV14, SV15 sowie SV16 hohe Trennschärfen (r_{it} > .50) festgestellt werden. Das Item SV13 weißt dagegen eine geringe Trennschärfe

(r_{it} < .30) auf. Für die Items der Skala PSI konnten durchgehend mittelmäßige Trennschärfen (.30 < r_{it} > .50) festgestellt werden.

Tabelle 7: Trennschärfen der Items des DBR-MIS nach Cronbachs α

Skala	Item	Itemtrennschärfe (r_{it})
SAV	SAV01	0.771
	SAV02	0.765
	SAV03	0.808
VPL	VPL04	0.573
	VPL05	0.649
	VPL06	0.749
DAV	DAV07	0.729
	DAV08	0.703
	DAV09	0.689
PSI	PSI10	0.426
	PSI11	0.338
	PSI12	0.454
SV	SV13	0.291
	SV14	0.523
	SV15	0.770
	SV16	0.741
PS	PS17	0.829
	PS18	0.863
	PS19	0.827

Anmerkungen. DBR-MIS = Direct Behavior Rating-Multi Item Scale, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten.

Abschließend betrachtet, weisen die ermittelten internen Konsistenzen und Itemtrennschärfen auf eine gute Messgenauigkeit für den Großteil der sechs Skalen des DBR-MIS hin. Der Messzusammenhang zwischen der latenten Variablen PSI und den dazugehörigen Items als Indikatoren ist allerdings sowohl anhand der Faktorladungen als auch anhand der Reliabilitäts- und Trennschärfewerte kritisch anzusehen. Gleiches gilt für den Zusammenhang der Variablen SV mit dem Item SV13.

5.2 Bestimmung allgemeiner Schwellenwerte

Die Ergebnisse der Bestimmung von Schwellenwerten für die drei Kategorien *auffällig*, *grenzwertig* und *unauffällig* anhand der Mittelwertverteilung pro Skala werden im Folgenden für die einzelnen Verhaltensdimensionen dargestellt. Die Reihenfolge der Darstellung ist an der Skalenstruktur des DBR-MIS orientiert. Pro Verhaltensdimension variiert die in der Analyse berücksichtigte Stichprobenanzahl, da für einige Datensätze aufgrund fehlender Itemwerte keine Mittelwerte berechnet werden konnten.

Dimension SAV

Für die Bestimmung der Schwellenwerte für die Dimension SAV konnte auf die Skalenmittelwerte von 200 Kindern und Jugendlichen (*N* = 200, *M* = 2,77, *SD* = 1,68) zurückgegriffen werden. Im Hinblick auf die Gesamtstichprobe lagen also für neun SuS keine verwendbaren Daten vor. Die Abbildung 1 zeigt die Verteilung der Mittelwerte in der Stichprobe in Form eines Säulendiagramms. Die x-Achse bildet dabei die verschiedenen in der Stichprobe aufgetretenen Skalenmittelwerte ab, während die y-Achse die Anzahl der SuS angibt. Insgesamt umfassen die berechneten Mittelwerte eine Spanne von 0,6 bis 7, wobei *7* den höchstmöglichen Ausprägungsgrad der Verhaltensdimension darstellt. Anhand der Abbildung 1 ist zu erkennen, dass die Mittelwerte nicht normalverteilt sind. Die farblichen Linien wurden der Abbildung nachträglich hinzugefügt und markieren bereits die Schwellenwerte für die Kategorien *grenzwertig* (gelbe Linie) und *auffällig* (rote Linie).

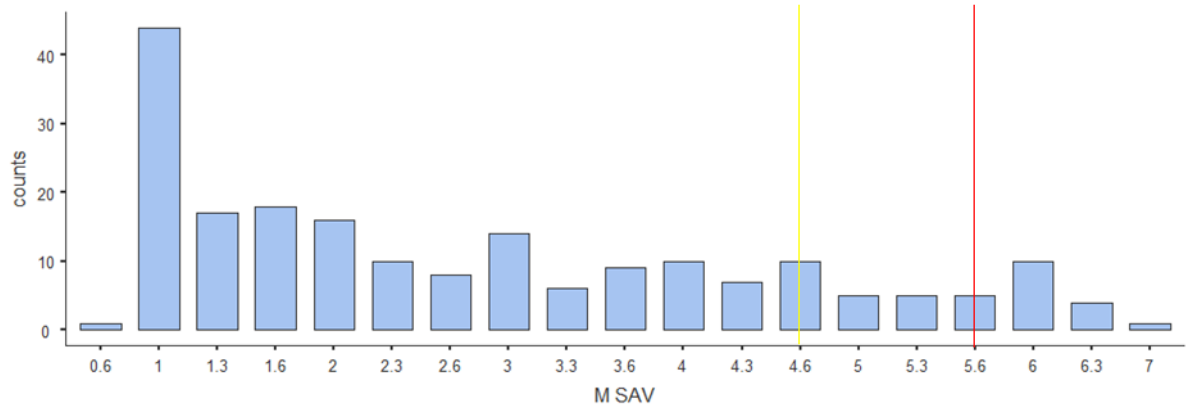

Abbildung 1: Mittelwertverteilung der Skala SAV

Anmerkungen. N = 200, Counts = Anzahl der SuS, SAV = Störendes und auflehnendes Verhalten, M SAV = Mittelwerte der Skala SAV, Linien: gelb = Schwellenwert (SW) Kategorie *grenzwertig*, rot = SW Kategorie *auffällig.*

Die Tabelle 10 zeigt die Zuordnung der Skalenmittelwerte zu Prozenträngen, die über die Verteilung der Werte in der berücksichtigten Stichprobe berechnet wurden. Für die Bestimmung der Schwellenwerte für die Kategorien *unauffällig*, *grenzwertig* und *auffällig* können in der Verteilung der Mittelwerte empirische Anteile gefunden werden, die den Zielvorgaben von 80–10–10% genau entsprechen. Die Dimension SAV stellt ebenso wie die Dimensionen VPL, DAV und PSI eine Problemdimension dar, wobei für die Bestimmung von *auffälligen* und *grenzwertigen* Anteilen also bei den hohen Mittelwerten angesetzt wird. Dahingehend können die Werte von 4,6 für die Kategorie *grenzwertig* und 5,6 für die Kategorie *auffällig* als Schwellenwerte bestimmt

werden. Die farbliche Unterteilung der Tabelle 10 gibt an, welche Mittelwerte für die Verhaltensdimension SAV angesichts der Schwellenwerte in die Kategorien *unauffällig* (grün), *grenzwertig* (gelb) und *auffällig* (rot) eingeordnet werden können.

Tabelle 8: Zuordnung von Mittelwerten der Skala SAV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
0,6	1	0,5%	0,5%
1	44	22,0%	22,5%
1,3	17	8,5%	31,0%
1,6	18	9,0%	40,0%
2	16	8,0%	48,0%
2,3	10	5,0%	53,0%
2,6	8	4,0%	57,0%
3	14	7,0%	64,0%
3,3	6	3,0%	67,0%
3,6	9	4,5%	71,5%
4	10	5,0%	76,5%
4,3	7	3,5%	80,0%
4,6	**10**	**5,0%**	**85,0%**
5	5	2,5%	87,5%
5,3	5	2,5%	90,0%
5,6	**5**	**2,5%**	**92,5%**
6	10	5,0%	97,5%
6,3	4	2,0%	99,5%
7	1	0,5%	100,0%

Anmerkungen. N = 200, Anzahl = Anzahl der SuS, SAV = Störendes und auflehnendes Verhalten, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Dimension VPL

Für die Dimension VPL konnten anhand der vorliegenden empirischen Daten Mittelwerte von insgesamt 201 SuS (*N* = 201, *M* = 3,67, *SD* = 1,66) berechnet werden. Die Abbildung 2 zeigt, dass bei der Verteilung der Mittelwerte mit den Extremwerten 1 und 7 keine Normalverteilung vorliegt.

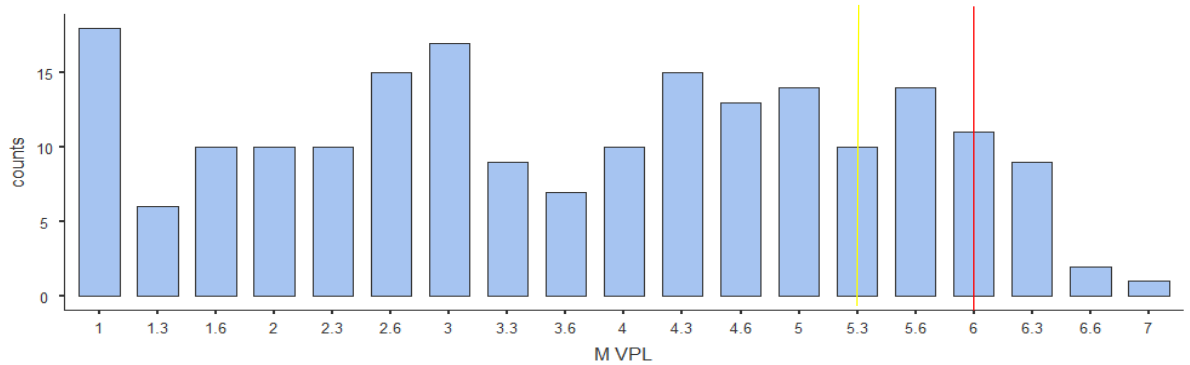

Abbildung 2: Mittelwertverteilung der Skala VPL

Anmerkungen. N = 201, Counts = Anzahl der SuS, VPL = Verhaltensprobleme beim Lernen, M VPL = Mittelwerte der Skala VPL, Linien: gelb = Schwellenwert (SW) Kategorie *grenzwertig*, rot = SW Kategorie *auffällig.*

An der Zuordnung der Mittelwerte zu Prozenträngen in Tabelle 11 ist erkennbar, dass bei dieser Mittelwertverteilung nur empirische beobachtbare Anteile vorliegen, die jeweils über oder unter die Zielvorgabe von jeweils 10% für die Kategorien *auffällig* und *grenzwertig* fallen. Demnach werden als Schwellenwerte die Mittelwerte ausgewählt, bei denen die Prozentanteile näher an der Zielvorgabe liegen. Für die Kategorie *auffällig* liegt der wahre Schwellenwert zwischen den möglichen Mittelwerten 6 und 6,3. Da der Prozentanteil des Mittelwertes 6 (11,5%) der Zielvorgabe 10% allerdings mehr entspricht als der Anteil des Wertes 6,3 (7%), wird dieser als Schwellenwert bestimmt. In Anlehnung an dieses Vorgehen wird für die Kategorie *grenzwertig* der Schwellenwert bei 5,3 gesetzt, was einem Prozentanteil von 12% für diese Kategorie entspricht. Insgesamt fallen damit 76,6% der SuS der Stichprobe in die Kategorie *unauffällig*.

Tabelle 9: Zuordnung von Mittelwerten der Skala VPL zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	18	9,0%	9,0%
1,3	6	3,0%	11,9%
1,6	10	5,0%	16,9%
2	10	5,0%	21,9%
2,3	10	5,0%	26,9%
2,6	15	7,5%	34,3%
3	17	8,5%	42,8%
3,3	9	4,5%	47,3%
3,6	7	3,5%	50,7%
4	10	5,0%	55,7%
4,3	15	7,5%	63,2%
4,6	13	6,5%	69,7%
5	14	7,0%	76,6%
5,3	**10**	**5,0%**	**81,6%**
5,6	14	7,0%	88,6%
6	**11**	**5,5%**	**94,0%**
6,3	9	4,5%	98,5%
6,6	2	1,0%	99,5%
7	1	0,5%	100,0%

Anmerkungen. N = 201, Anzahl = Anzahl der SuS, VPL = Verhaltensprobleme beim Lernen, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Dimension DAV

Wie bei den bereits beschriebenen Dimensionen sind auch die Mittelwerte in der Dimension DAV nicht normalverteilt (siehe Abbildung 3). Im Vergleich lag aber für die Bestimmung der Schwellenwerte eine geringere Anzahl von Stichprobendaten (*N* = 194, *M* = 2,58, *SD* = 1,52) vor, da bei 15 SuS keine Itemwerte für die Skala angegeben wurden.

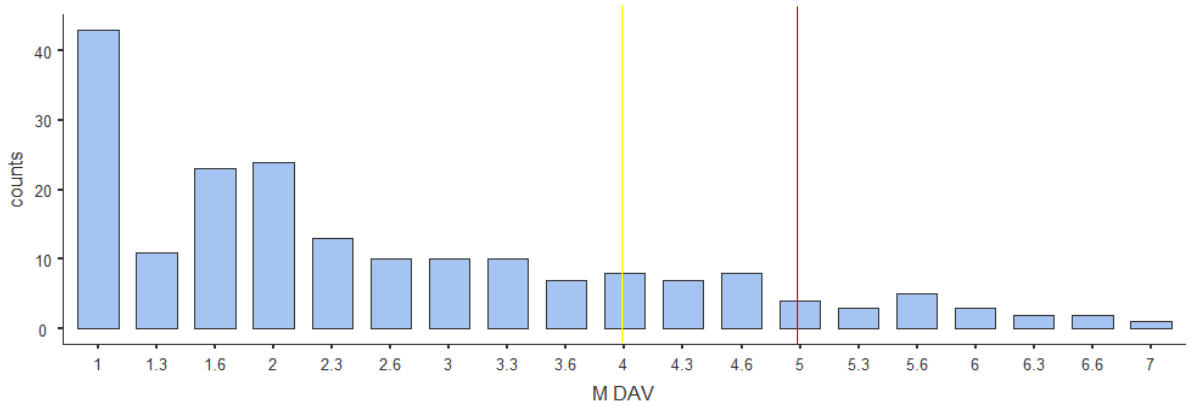

Abbildung 3: Mittelwertverteilung der Skala DAV

Anmerkungen. *N* = 194, Counts = Anzahl der SuS, DAV = Depressives und ängstliches Verhalten, M DAV = Mittelwerte der Skala DAV, Linien: gelb = Schwellenwert (SW) Kategorie *grenzwertig*, rot = SW Kategorie *auffällig.*

Angesichts der Prozentrangzuordnung in Tabelle 10 wird die Schwelle für die Kategorie *auffällig* bei dem Mittelwert 5 gesetzt, der mit einem Prozentanteil von 10,2% nah an der Zielvorgabe liegt. Als Schwellenwert für die Kategorie *grenzwertig* wird der Mittelwert 4 ausgewählt, dessen Prozentanteil mit 11,8% näher an der 10%-Vorgabe liegt als der des Mittelwertes 4,3 (7,7%). Zusammenfassend fallen damit in die Kategorien *auffällig* und *grenzwertig* insgesamt 22%.

Tabelle 10: Zuordnung von Mittelwerten der Skala DAV zu Prozenträngen

Mittelwert	Anzahl (N)	Anzahl in %	Prozentrang
1	43	22,2%	22,2%
1,3	11	5,7%	27,8%
1,6	23	11,9%	39,7%
2	24	12,4%	52,1%
2,3	13	6,7%	58,8%
2,6	10	5,2%	63,9%
3	10	5,2%	69,1%
3,3	10	5,2%	74,2%
3,6	7	3,6%	77,8%
4	**8**	**4,1%**	**82,0%**
4,3	7	3,6%	85,6%
4,6	8	4,1%	89,7%
5	**4**	**2,1%**	**91,8%**
5,3	3	1,5%	93,3%
5,6	5	2,6%	95,9%
6	3	1,5%	97,4%
6,3	2	1,0%	98,5%
6,6	2	1,0%	99,5%
7	1	0,5%	100,0%

Anmerkungen. N = 194, Anzahl = Anzahl der SuS, DAV = Depressives und ängstliches Verhalten, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Dimension PSI

Für die Stichprobenanzahl von 179 SuS (*N* = 179, *M* = 2,73, *SD* = 1,33) ergibt sich die in der Abbildung 4 angezeigte Mittelwertverteilung der Verhaltensdimension PSI. Auffällig sind die hohen Auftretenshäufigkeiten der Mittelwerte 1 und 3,3, die verhindern, dass die Verteilung als Approximation einer Normalverteilung beschrieben werden kann. Insgesamt treten 18 verschiedene Mittelwerte zwischen den Extremwerten 0,6 und 7 auf.

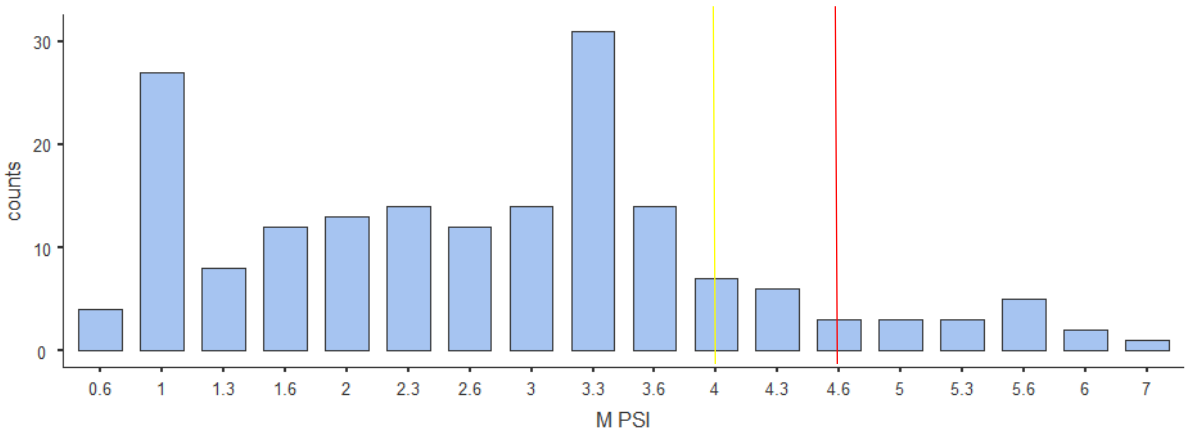

Abbildung 4: Mittelwertverteilung der Skala PSI

Anmerkungen. N = 179, Counts = Anzahl der SuS, PSI = Probleme in sozialen Interaktionen, M PSI = Mittelwerte der Skala PSI, Linien: gelb = Schwellenwert (SW) Kategorie *grenzwertig*, rot = SW Kategorie *auffällig.*

Die farblichen Markierungen in der Tabelle 11 zeigen, dass anhand der Bestimmung der Schwellenwerte für diese Dimension erstmalig ein Anteil von über 80% in die Kategorie *unauffällig* (grün) fällt. Dies liegt daran, dass bei der Schwellenwertbestimmung in Annäherung an die Zielvorgabe für die Kategorien *grenzwertig* und *auffällig* jeweils Prozentanteile ausgewählt wurden, die unter 10% lagen. Mit den Schwellenwerten 4 und 4,6 fallen damit 7,3% in den *grenzwertigen* Bereich (gelb) und 9,6% in den *auffälligen* Bereich (rot).

Tabelle 11: Zuordnung von Mittelwerten der Skala PSI zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
0,6	4	2,2%	2,2%
1	27	15,1%	17,3%
1,3	8	4,5%	21,8%
1,6	12	6,7%	28,5%
2	13	7,3%	35,8%
2,3	14	7,8%	43,6%
2,6	12	6,7%	50,3%
3	14	7,8%	58,1%
3,3	31	17,3%	75,4%
3,6	14	7,8%	83,2%
4	**7**	**3,9%**	**87,2%**
4,3	6	3,4%	90,5%
4,6	**3**	**1,7%**	**92,2%**
5	3	1,7%	93,9%
5,3	3	1,7%	95,5%
5,6	5	2,8%	98,3%
6	2	1,1%	99,4%
7	1	0,6%	100,0%

Anmerkungen. N = 179, Anzahl = Anzahl der SuS, PSI = Probleme in sozialen Interaktionen, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Dimension SV

Die Verteilung der berechneten Mittelwerte der Dimension SV in der Abbildung 5 kann als Approximation einer Normalverteilung bezeichnet werden, da die Werte annähernd glockenförmig zwischen den vorliegenden Extremwerten 1,5 und 7 verteilt sind. Insgesamt zeigt die Abbildung die Mittelwertverteilung einer Stichprobengröße von 175 SuS (*N* = 175, *M* = 4,07, *SD* = 1,33). Bei 34 SuS lagen keine Einschätzungen der Lehrkräfte für die Items dieser Dimension vor.

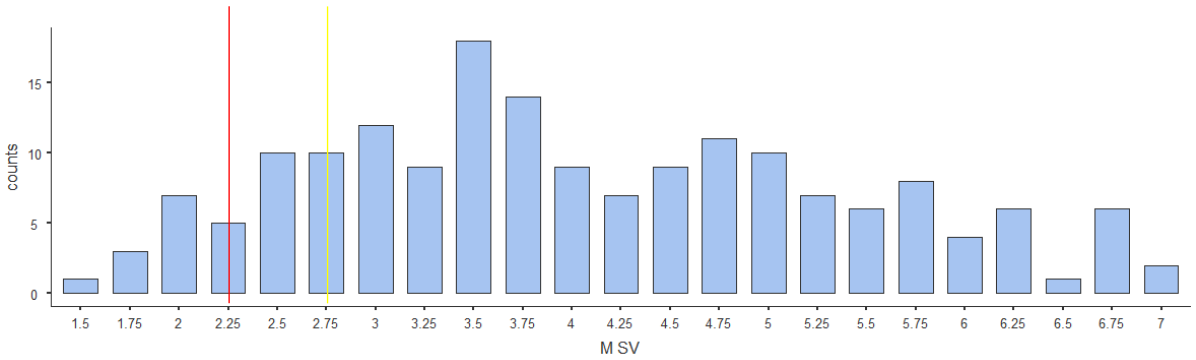

Abbildung 5: Mittelwertverteilung der Skala SV

Anmerkungen. *N* = 175, Counts = Anzahl der SuS, SV = Schulbezogenes Verhalten, M SV = Mittelwerte der Skala SV, Linien: gelb = Schwellenwert Kategorie *grenzwertig*, rot = Schwellenwert Kategorie *auffällig.*

Da die Dimension SV ebenso wie die Dimension PS eine Stärkedimension darstellt, wird bei der Bestimmung der Schwellenwerte für die Kategorie *auffällig* im Gegensatz zu den anderen Dimensionen bei den niedrigen Mittelwerten angesetzt. Dies spiegelt sich auch in der Tabelle 12 wider. Im Hinblick auf die Zielvorgabe können die Schwellenwerte von 2,25 für die Kategorie *auffällig* und 2,75 für die Kategorie *grenzwertig* bestimmt werden. Damit entsprechen die Mittelwerte, die als *auffällig* eingestuft werden, einem Prozentanteil von 9,1%, während 11,4% unter die Kategorie *grenzwertig* fallen.

Tabelle 12: Zuordnung von Mittelwerten der Skala SV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1,5	1	0,6%	0,6%
1,75	3	1,7%	2,3%
2	7	4,0%	6,3%
2,25	**5**	**2,9%**	**9,1%**
2,5	10	5,7%	14,9%
2,75	**10**	**5,7%**	**20,6%**
3	12	6,9%	27,4%
3,25	9	5,1%	32,6%
3,5	18	10,3%	42,9%
3,75	14	8,0%	50,9%
4	9	5,1%	56,0%
4,25	7	4,0%	60,0%
4,5	9	5,1%	65,1%
4,75	11	6,3%	71,4%
5	10	5,7%	77,1%
5,25	7	4,0%	81,1%
5,5	6	3,4%	84,6%
5,75	8	4,6%	89,1%
6	4	2,3%	91,4%
6,23	6	3,4%	94,9%
6,5	1	0,6%	95,4%
6,75	6	3,4%	98,9%
7	2	1,1%	100,0%

Anmerkungen. *N* = 175, Anzahl = Anzahl der SuS, SV = Schulbezogenes Verhalten, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Dimension PS

Für die Bestimmung der Schwellenwerte der Dimension PS konnte auf die Mittelwerte von 194 SuS (*N* = 194, *M* = 4,38, *SD* = 1,74) zurückgegriffen werden. Die Abbildung 6 zeigt, dass in der Stichprobe insgesamt Mittelwerte zwischen den Extremwerten 1 und 7 auftreten, diese aber nicht normalverteilt sind.

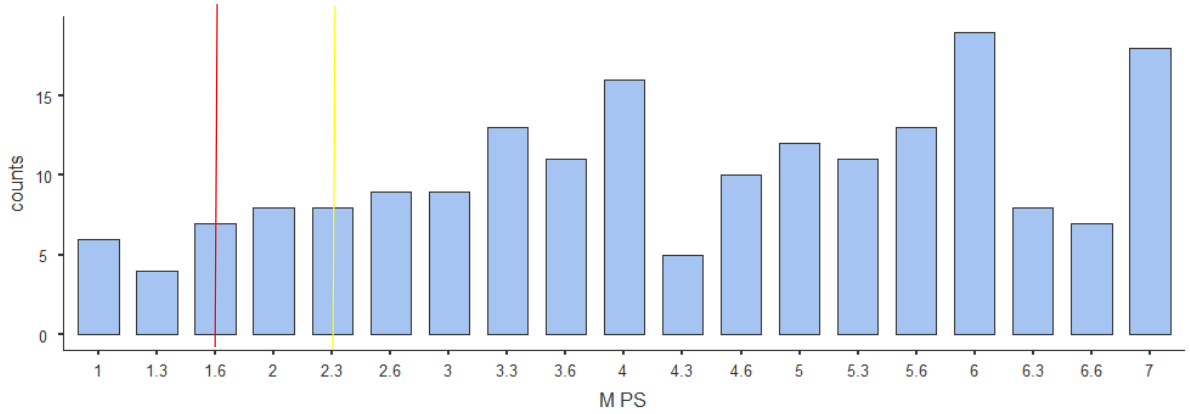

Abbildung 6: Mittelwertverteilung der Skala PS

Anmerkungen. N = 194, Counts = Anzahl der SuS, PS = Prosoziales Verhalten, M PS = Mittelwerte der Skala PS. Linien: gelb = Schwellenwert Kategorie *grenzwertig*, rot = Schwellenwert Kategorie *auffällig.*

Da die Dimension PS ebenfalls eine Stärkedimension darstellt, befinden sich die ausgewählten Schwellenwerte im niedrigen Wertebereich. Mit einem empirischen Prozentanteil von 8,8% für die Kategorie *auffällig*, konnte der Schwellenwert 1,6 bestimmt werden. Der Schwellenwert für die Kategorie *grenzwertig* liegt nur knapp darüber bei 2,3, womit ein Anteil von 8,2% unter diese Kategorie fällt. In der vorliegenden Stichprobe sind angesichts dieser Schwellen, 83% der Mittelwerte dem *unauffälligen* Bereich zuzuordnen.

Tabelle 13: Zuordnung von Mittelwerten der Skala PS zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	6	3,1%	3,1%
1,3	4	2,1%	5,2%
1,6	**7**	**3,6%**	**8,8%**
2	8	4,1%	12,9%
2,3	**8**	**4,1%**	**17,0%**
2,6	9	4,6%	21,6%
3	9	4,6%	26,3%
3,3	13	6,7%	33,0%
3,6	11	5,7%	38,7%
4	16	8,2%	46,9%
4,3	5	8,2%	46,9%
4,6	10	5,2%	54,6%
5	12	6,2%	60,8%
5,3	11	5,7%	66,5%
5,6	13	6,7%	73,2%
6	19	9,8%	83,0%
6,3	8	4,1%	87,1%
6,6	7	3,6%	90,7%
7	18	9,3%	100,0%

Anmerkungen. N = 194, Anzahl = Anzahl der SuS, PS = Prosoziales Verhalten, Farbliche Markierungen: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

5.3 Effekte von Geschlecht und Alter

Für die Überprüfung des Effekts der Variablen Geschlecht auf die Skalen des DBR-MIS gibt die Tabelle 14 zunächst einen Überblick über die durchschnittlichen Mittelwerte und Standardabweichungen von Jungen und Mädchen pro Skala. Dabei fällt auf, dass sich die Stichprobenanzahl zwischen den Geschlechtern unterscheidet und deutlich mehr Daten von Jungen vorliegen. Eine erste vergleichende Betrachtung der Mittelwerte beider Gruppen zeigt, dass insbesondere in den Skalen SAV, VPL, SV und PS Unterschiede vorliegen. In den Skalen DAV und PSI liegen die durchschnittlichen Mittelwerte beider Gruppen dagegen nah beieinander.

Tabelle 14: Durchschnittliche Mittelwerte und Standardabweichungen pro Skala getrennt nach Geschlecht

Skala	Geschlecht	*n*	*M*	*SD*
SAV	Jungen	140	3,08	1,72
	Mädchen	60	203	1,30
VPL	Jungen	143	4,07	1,55
	Mädchen	58	2,69	1,51
DAV	Jungen	136	2,59	1,49
	Mädchen	58	2,55	1,59
PSI	Jungen	127	2,77	1,32
	Mädchen	52	2,62	1,36
SV	Jungen	122	3,82	1,30
	Mädchen	53	4,64	1,24
PS	Jungen	138	4,05	1,74
	Mädchen	56	5,17	1,45

Anmerkungen. n = Teilstichprobe, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten.

Im Hinblick auf die Unterschiede im Stichprobenumfang der beiden Gruppen gibt der Levene-Test an, dass lediglich bei den Skalen SAV ($p = 0.001$) und PS ($p = 0.041$) signifikante Unterschiede ($p < .05$) in den Varianzen der beiden Gruppen bestehen. Da die Prüfung der Mittelwertverteilung anhand des Shapiro-Wilk-Test für die einzelnen Skalen zeigt, dass die Mittelwerte der jeweiligen Teilstichproben nicht normalverteilt sind ($p < .05$), erfolgt die Analyse von Unterschieden zwischen den Mittelwerten der beiden Geschlechtsgruppen mittels des Mann-Whitney U-Tests. Die Ergebnisse des Tests bezogen auf die Signifikanzprüfung (p-Wert) und die Effektstärke (Cohen's

d) sind der Tabelle 15 zu entnehmen. Die p-Werte ($p < .001$) für die Skalen SAV, VPL, SV und PS zeigen an, dass die Unterschiede in den Mittelwerten der zwei Gruppen signifikant sind. Während dabei für die Skalen SAV, SV und PS eine mittlere Effektstärke ($|d| > 0.5$) der Variablen Geschlecht festgestellt werden kann, liegt bei der Skala VPL eine große Effektstärke ($|d| > 0.8$) vor. Bei den Mittelwerten der Skalen DAV und PSI weist der p-Wert auf keine signifikanten Unterschiede zwischen den untersuchten Gruppen.

Tabelle 15: Ergebnisse des Mann-Whitney U-Tests für die zwei unabhängigen Geschlechtsgruppen pro Skala

Skala	**p**	**Cohen's *d***
SAV	< .001	-0.6511
VPL	< .001	-0.8970
DAV	0.733	-0.0246
PSI	0.390	-0.1168
SV	< .001	0.6364
PS	< .001	0.6721

Anmerkungen. SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, p = Signifikanzwert für die Mittelwertsunterschiede der zwei untersuchten Gruppen, Cohen's *d* = Effektgröße.

Für die Untersuchung des Effekts der Variablen Alter auf die Skalen des DBR-MIS wurde im Zuge der Varianzanalyse zunächst die Varianzhomogenität zwischen den drei eingeteilten Altersgruppen (Altersgruppe 1 = 6-9 Jahre; Altersgruppe 2 = 10-13 Jahre; Altersgruppe 3 = 14-18 Jahre) überprüft. Die Tabelle 16 zeigt die vorliegenden Stichprobenanzahlen sowie die durchschnittlichen Mittelwerte der drei Altersgruppen pro Skala.

Die anhand des Levene-Tests berechneten p-Werte der Skalen SAV ($p = 0.481$), VPL ($p = 0.251$), DAV ($p = 0.624$), PSI ($p = 0.570$), SV ($p = 0.147$) und PS ($p = 0.134$) liegen alle über dem Signifikanzniveau ($p < .05$), womit keine signifikanten Unterschiede zwischen den Varianzen der drei Altersgruppen festgestellt werden können. Daher wird für die Varianzanalyse der Fisher-Test herangezogen

Tabelle 16: Durchschnittliche Mittelwerte und Standardabweichungen der drei Altersgruppen pro Skala

Skala	Altersgruppe	*N*	*M*	*SD*
SAV	1	73	2,75	1,78
	2	70	2,76	1,57
	3	55	1,71	1,71
VPL	1	72	3,66	1,70
	2	73	3,90	1,72
	3	54	3,37	1,51
DAV	1	71	2,64	1,60
	2	68	2,59	1,51
	3	53	2,45	1,45
PSI	1	70	2,50	1,35
	2	58	2,77	1,35
	3	50	2,99	1,26
SV	1	70	4,17	1,43
	2	71	4,05	1,30
	3	33	3,83	1,17
PS	1	70	4,47	1,87
	2	69	4,33	1,72
	3	53	4,28	1,61

Anmerkungen. SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, Altersgruppe 1 = 6-9 Jahre, Altersgruppe 2 = 10-13 Jahre, Altersgruppe 3 = 14-18 Jahre.

Die Tabelle 17 zeigt die mittels des Fisher-Tests berechneten p-Werte pro Skala. Bei keiner Skala liegt ein signifikanter p-Wert vor, sodass angenommen werden kann, dass zwischen den Mittelwerten der drei Altersgruppen keine überzufälligen Unterschiede bestehen.

Tabelle 17: Ergebnisse der einfaktoriellen Varianzanalyse („one-way ANOVA") für die drei Altersgruppen pro Skala

Skala	*F*	p
SAV	0.0175	0.983
VPL	1.5623	0.212
DAV	0.2232	0.800
PSI	2.0341	0.134
SV	0.7249	0.486
PS	0.2086	0.812

Anmerkungen. SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, *F* = Kennwert der Varianzanalyse; p = Signifikanzwert.

Zusammenfassend betrachtet, zeigen die Ergebnisse der Analysen im Rahmen der dritten Forschungsfrage, dass die Variable Geschlecht einen mittleren bis hohen Effekt auf vier der sechs Skalen des DBR-MIS hat. Dagegen konnte für die Variable Alter mittels der Varianzanalyse zwischen den drei eingeteilten Altersgruppen kein Effekt festgestellt werden. Die Hypothese H2 kann dahingehend nur in Bezug auf die Variable Geschlecht bestätigt werden, weshalb nachfolgend lediglich eine zusätzliche Bestimmung geschlechtsspezifischer Schwellenwerte vorgenommen wurde.

Die Schwellenwertbestimmung pro Skala für die beiden Geschlechtsgruppen erfolgte ebenfalls in Orientierung an die Zielvorgaben von 80–10–10% für die Kategorien *unauffällig*, *grenzwertig* und *auffällig*. Die jeweiligen tabellarischen Prozentrangzuordnungen pro Skala, anhand derer die Schwellenwerte in Annäherung an die Zielvorgaben bestimmt wurden, sind für die Jungen im Anhang B und für die Mädchen im Anhang C dieser Arbeit zu finden. In der Tabelle 18 sind neben den in Kapitel 5.2 bestimmten allgemeinen Schwellenwerten die ermittelten geschlechtsspezifischen Schwellenwerte für die Kategorien *auffällig* und *grenzwertig* pro Skala des DBR-MIS aufgeführt.

Tabelle 18: Allgemeine und geschlechtsspezifische Schwellenwerte für die Skalen des DBR-MIS und Prozentanteile der SuS mit den Werten in den Kategorien grenzwertig und auffällig in der vorliegenden Stichprobe

Skalen	**Gesamt**					**Jungen**					**Mädchen**				
	n	G	G%	A	A%	*n*	G	G%	A	A%	*n*	G	G%	A	A%
SAV	200	4,6	10	5,6	10	140	5	7,9	6	10,7	60	3,3	10	4,6	10
VPL	201	5,3	12	6	11,5	143	6	7	6,3	7	58	4,3	10,2	5,6	8,5
DAV	194	4	11,8	5	10,2	136	4	12,4	5	9,4	58	4,3	10,3	5,3	10,2
PSI	179	4	7,3	4,6	9,6	127	4	6,3	4,6	11,1	52	3,6	7,6	4,3	11,5
SV	175	2,75	11,4	2,25	9,1	122	2,5	10,7	2	9	53	3,5	11,5	2,75	11,3
PS	194	2,3	8,2	1,6	8,8	138	2,3	10,9	1,6	11,6	56	3,6	10,7	3	8,9

Anmerkungen. DBR-MIS bezeichnet die Direct Behavior Rating-Multi Item Scale der Onlineplattform LEVUMI, SAV = Störendes und auflehnendes Verhalten, VPL = Verhaltensprobleme beim Lernen, DAV = Depressives und ängstliches Verhalten, PSI = Probleme in sozialen Interaktionen, SV = Schulbezogenes Verhalten, PS = Prosoziales Verhalten, *n* = Größe der Teilstichprobe; G = Schwellenwerte für die Kategorie *grenzwertig*; G% = Prozentanteil in der Kategorie *grenzwertig*; A = Schwellenwerte für die Kategorie *auffällig*; A% = Prozentanteil in der Kategorie *auffällig*.

6. Diskussion der Ergebnisse

In der vorliegenden Arbeit wurden sowohl alters- und geschlechtsübergreifende als auch geschlechtsspezifische Schwellenwerte für die Skalen des verhaltensverlaufsdiagnostischen Instruments DBR-MIS anhand einer Stichprobe von SuS in Deutschland (*N* = 209) bestimmt. Die Ergebnisse der Analysen und das methodische Vorgehen im Rahmen der drei übergeordneten Forschungsfragen werden in diesem Kapitel zusammenfassend diskutiert.

Im Hinblick auf das Forschungsinteresse dieser Arbeit, eine kriteriale Bezugsnorm für den DBR-MIS in Form von Schwellenwerten zu bestimmen, zielten die Analysen im Rahmen der ersten Forschungsfrage darauf ab, zu untersuchen, auf welcher Konstruktebene des DBR-MIS eine Schwellenwertbestimmung sinnvoll ist. Für die Beantwortung dieser Frage wurde zunächst die dimensionale Struktur des Instruments betrachtet. Die Struktur des DBR-MIS umfasst zwei Dimensionsebenen, womit zwei mögliche Messmodelle als Grundlage für die Schwellenwertbestimmung vorliegen. Die Entscheidung für ein Messmodell erfolgte über die Überprüfung der Faktorenstrukturen beider Modelle sowie über den Vergleich der jeweiligen Modellgüte. Anhand konfirmatorischer Faktorenanalysen konnten sowohl für das Sechs-Faktoren-Modell als auch für das Drei-Faktoren-Modell größtenteils hohe Messzusammenhänge zwischen den jeweiligen latenten und manifesten Variablen festgestellt werden. Geringere Faktor-Indikator-Zusammenhänge zeigten bei dem Sechs-Faktoren-Modell lediglich die Faktorladungen der Items PSI10, PSI12 und SV13 an. Auch beim Drei-Faktoren-Modell wurden für diese Items geringere Faktorladungen gefunden. Insgesamt bestätigten die ermittelten Faktorladungen weitgehend die Faktorenstruktur beider Messmodelle.

In einem nächsten Schritt wurde daher die Güte der Modelle anhand verschiedener Gütekriterien und Fit-Indizes verglichen. In diesem Vergleich konnte festgestellt werden, dass das Sechs-Faktoren-Modell besser zu den erhobenen empirischen Daten passt als das Drei-Faktoren-Modell, womit die aufgestellte Hypothese H1 bestätigt wurde. Als Grundlage für die nachfolgenden Schwellenwertbestimmungen wurde daher das Sechs-Faktoren-Modell ausgewählt. Dieses umfasst die sechs Subskalen

SAV, VPL, DAV, PSI, SV und PS, die angesichts der Struktur des DBR-MIS, Dimensionen auf zweiter Ebene darstellen.

Hinsichtlich der übergeordneten ersten Forschungsfrage zeigen die Untersuchungsergebnisse also, dass eine Schwellenwertbestimmung auf der zweiten Konstruktebene (sechs Teilkonstrukte) sinnvoller ist als auf der ersten Ebene (drei Teilkonstrukte), welche dem Drei-Faktoren-Modell mit den Skalen EXT, INT und PSV entspricht. Obwohl das Sechs-Faktoren-Modell im Vergleich der Modellgüte besser abschneidet als das Drei-Faktoren-Modell, ist allerdings kritisch zu betrachten, dass die Fit-Indizes dieses Modells nicht zufriedenstellend sind. So weist der signifikante x^2-Wert darauf hin, dass das aufgestellte Modell nicht gut zu den Daten passt. Eine mögliche Erklärung für diesen x^2-Wert könnte die nicht vorliegende Normalverteilung der Indikatorvariablen in der Stichprobe sein. Diese Normalverteilung gilt als Voraussetzung der ML-Schätzmethode, welche im Rahmen des x^2-Tests in *jamovi* angenommen wurde. Die Verletzung dieser Voraussetzung kann zur Berechnung von zu großen x^2-Werten führen, die dann einen schlechten Model-Fit ausweisen, obwohl eigentlich ein guter Model-Fit besteht (Curran, West & Finch, 1996, zitiert nach Döring & Bortz, 2016). Zudem ist der x^2-Wert abhängig von dem Stichprobenumfang (Döring & Bortz, 2016). Da in der vorliegenden Arbeit der Stichprobenumfang relativ gering ist, kann an dieser Stelle von einem Einfluss auf den x^2-Test ausgegangen werden. Allerdings zeigen auch die anderen Fit-Indizes wie der RMSEA, der von der Stichprobengröße relativ unabhängig ist, dass keine perfekte Passung zwischen dem Modell und den Daten vorliegt (ebd.). Über die Fit-Indizes konnte die Konstruktvalidität des Sechs-Faktoren-Modells damit nicht bestätigt werden. Da aber anzunehmen ist, dass die Stichprobengröße und die Verletzung der Voraussetzung einer Normalverteilung einen Einfluss auf die Fit-Indizes haben, kann das Messmodell anhand dieser auch nicht grundlegend abgelehnt werden. Eine weitere Überprüfung der Güte des Modells anhand einer umfassenderen Stichprobe könnte dahingehend weiterführende Erkenntnisse liefern.

In einem nächsten Schritt wurden in Anlehnung an die Auswahl des Sechs-Faktoren-Modells die internen Konsistenzen der sechs Subskalen des DBR-MIS nach Cronbachs α berechnet. Bis auf die Skala PSI konnten für alle Skalen akzeptable bis hohe Werte ermittelt werden, die auf eine gute Reliabilität der Ratingskala hinweisen.

Kritisch zu betrachten, ist allerdings die niedrige interne Konsistenz der Skala PSI. Bei dieser Skala ist also fraglich, ob die Items PSI10, PSI11 und PSI12 eindimensionale und widerspruchsfreie Indikatoren für die Messung der latenten Variablen PSI darstellen. Die Trennschärfen der drei PSI-Items liegen jedoch in einem mittelmäßigen Wertebereich, weshalb davon auszugehen ist, dass sie, wie auch die Items der Skalen SAV (SAV01, SAV02, SAV03), VPL (VPL04, VPL05, VPL06), DAV (DAV07, DAV08, DAV09), SV (SV14, SV15, SV16) und PS (PS17, PS18, PS19), für die durchgehend hohe Trennschärfen ermittelt wurden, Messungen der jeweilig zugeordneten latenten Variablen ermöglichen. Das einzige Item der gesamten Ratingskala, das eine geringe Trennschärfe ($r_{it} < .30$) aufweist, ist das Item SV13. Damit weicht es deutlich von den Trennschärfen der anderen Skalenitems SV14, SV15 und SV16 ab. Ähnliche Itemtrennschärfen in einer Skala sprechen für eine geeignete Skalierung, was wiederum bedeutet, dass die mit der Verrechnungsvorschrift gebildeten Testwerte für die Interpretation einer Merkmalsausprägung herangezogen werden können (Bühner, 2011, Casale et al. 2015b). Ob dies auch für die Testwerte der Skala SV gilt, ist aufgrund der niedrigen Trennschärfe des Items SV13 fraglich. Auch Hisker (2018) konnte in ihrer Pilotierungsstudie zum DBR-MIS eine geringe Trennschärfe des Items SV13 feststellen. Die ebenfalls im Rahmen der Studie durchgeführten Experteninterviews zeigten jedoch eine hohe Relevanz des Items hinsichtlich der Verwendung der Ratingskala in schulischen Kontexten an, weshalb von einer Entfernung des Items abgeraten wurde (Hisker, 2018). Um die Interpretationseinschränkungen des Testwertes der Skala SV zu lösen, schlug Hisker (2018) eine Umformulierung des Items SV13 vor. Studien, die sich mit einer möglichen Itemumformulierung zur Erhöhung der Trennschärfe befassen, stehen noch aus, sind aber für die Verbesserung der Ratingskala, insbesondere in Hinblick auf das von Casale et al. (2015b) beschriebene Gütekriterium eines gültigen Messmodells, notwendig.

Die konkrete Bestimmung von allgemeinen Schwellenwerten erfolgte über die Mittelwertverteilung in der vorliegenden Gesamtstichprobe pro Skala und die damit einhergehende Zuordnung der auftretenden Mittelwerte zu Prozenträngen. Anhand der Prozenträngе konnten für alle Mittelwertausprägungen die prozentuellen Anteile in der Gesamtstichprobe betrachtet werden. Anschließend wurden in Annäherung an die Zielvorgaben mit Anteilen von 80-10-10% für die Kategorien *unauffällig*, *grenzwertig*

und *auffällig* Schwellenwerte bestimmt. Diese Zielvorgaben wurden von Goodman (1997) für die Bestimmung von SDQ-Grenzwerten vorgeschlagen und wurden angesichts aktueller Prävalenzraten von psychischen Störungen und Verhaltensauffälligkeiten bei Kindern und Jugendlichen in umfassenden Studien und Übersichtsarbeiten, die diesen Vorgaben annähernd entsprechen, ausgewählt.

In den meisten Fällen lagen die empirisch beobachtbaren Anteile hinsichtlich der Prozentrangzuordnungen pro Skala entweder über oder unter den Zielvorgaben, woraufhin die prozentuellen Anteile ausgewählt wurden, die der Vorgabe von jeweils 10% für die Kategorien *auffällig* und *grenzwertig* am ehesten entsprachen. Dies führt dazu, dass sich die Prozentanteile der Gesamtstichprobe in den drei Kategorien pro Skala unterscheiden. In der Skala VPL fallen zum Beispiel 12% der SuS in die Kategorie *auffällig* und 11,5% in die Kategorie *grenzwertig*, während die Prozentanteile in der Skala PS pro Kategorie jeweils nur bei ca. 8% liegen. Insgesamt ergeben sich in den sechs Skalen für die Kategorien *auffällig* und *grenzwertig* kombinierte Anteile zwischen 16,9% und 23,5%. Die Abweichungen von den Zielvorgaben sind damit äußerst gering. Anzumerken ist an dieser Stelle aber, dass die Zielvorgaben, auch wenn sie in Anlehnung an ausgewiesene Prävalenzraten ausgewählt wurden, ein relativ willkürlich gesetztes Kriterium für die Bestimmung von Schwellenwerten darstellen und die Prävalenzen psychischer und verhaltensbezogener Auffälligkeiten bei Kindern und Jugendlichen lediglich annäherungsweise repräsentieren. Außerdem bestehen bei der Bestimmung von Schwellenwerten anhand von Zielvorgaben, verschiedene Möglichkeiten mit diesen umzugehen. Dabei hängen die Möglichkeiten eng mit der Zielsetzung der Schwellenwertberechnung zusammen. Bei der SDQ-Grenzwertbestimmung von Lohbeck et al. (2015) lag der Fokus beispielsweise auf einer hohen Sensitivität bei der Identifizierung von Problemfällen. Um zu verhindern, dass SuS mit Problemen übersehen werden, wurden die Grenzwerte für die drei Kategorien daher so ausgewählt, dass hinsichtlich der Zielvorgabe nicht weniger als 20% der Gesamtstichprobe in die Kategorien *auffällig* und *grenzwertig* fielen (ebd.). Anders als beim SDQ liegt der Einsatzfokus des DBR-MIS auf verhaltensverlaufsdiagnostischen Zwecken (Schurig et al., 2019). Ziel der Untersuchungen im Rahmen dieser Arbeit war es somit, Schwellenwerte zu bestimmen, die den Lehrkräften als Orientierungslinie und Interpretationshilfe bei der Betrachtung von Verhaltensverläufen dienen können. Für

die konkrete Identifikation von Verhaltensauffälligkeiten wird vom LEVUMI-Team eine Vorschaltung des SDQ empfohlen (Schurig et al., 2019). Die ausgewählten Schwellenwerte in Annäherung an die Zielvorgaben ohne besonderen Fokus auf Sensitivität erscheinen für das Forschungsinteresse dieser Arbeit daher sinnvoll.

In Anschluss an die Bestimmung allgemeiner Schwellenwerte für die Skalen des DBR-MIS wurde der Einfluss der Variablen Geschlecht und Alter auf die sechs Teilkonstrukte der Ratingskala analysiert. Mittels U-Test und Varianzanalyse wurde überprüft, ob Mittelwertsunterschiede zwischen Mädchen und Jungen sowie zwischen den drei eingeteilten Altersgruppen bestehen. Für die beiden Geschlechtsgruppen konnten signifikante Unterschiede in den Mittelwerten der Skalen SAV, VPL, SV und PS festgestellt werden. Dabei wurden für die Jungen durchschnittlich höhere Mittelwerte in den Problemdimensionen SAV und VPL und in den Stärkedimensionen SV und PS durchschnittlich niedrigere Mittelwerte als für die Mädchen ermittelt. Dies entspricht der durch Studien belegten Annahme, dass Jungen häufiger externalisierenden Verhaltensauffälligkeiten zeigen als Mädchen (vgl. Kapitel 2.2.1). In den Skalen DAV und PSI konnten dagegen keine signifikanten Unterschiede zwischen den beiden Gruppen festgestellt werden.

Da die Ergebnisse auf einen teilweise hohen Einfluss der Variablen Geschlecht auf die Teilkonstrukte des DBR-MIS hinweisen, wurden zusätzlich geschlechtsspezifische Schwellenwerte ebenfalls in Annäherung an die Zielvorgaben 80–10–10% bestimmt. Kritisch zu betrachten sind an dieser Stelle die deutlichen Unterschiede in der Stichprobengröße zwischen den Gruppen. So lagen für die Schwellenwertbestimmung mehr als doppelt so viele Daten für Jungen als für Mädchen vor. Dahingehend ist auch die Aussagekraft der allgemeinen Schwellenwerte eingeschränkt, da diese angesichts der Geschlechtsverteilung anhand von Daten einer nicht repräsentativen Stichprobe bestimmt wurden. Dies zeigt sich auch darin, dass die geschlechtsspezifischen Schwellenwerte für Jungen bei den meisten Skalen deutlich näher an den allgemeinen Schwellenwerten liegen als die der Mädchen. Zudem ist die Übertragung der Zielvorgaben von 80–10–10% auf die Stichproben beider Geschlechtsgruppen gleichermaßen für die Bestimmung geschlechtsspezifischer Schwellenwerte kritisch anzusehen, da Prävalenzraten, wie die Ergebnisse der KiGGS-Welle 2 zeigen, in der Regel je nach Geschlecht und Alter variieren (vgl. Kapitel 2.2.2).

Der Einfluss der Variablen Alter auf die Teilkonstrukte des DBR-MIS wurde mittels einer einfaktoriellen Varianzanalyse überprüft. Zwischen den drei eingeteilten Altersgruppen konnten keine signifikanten Unterschiede in den Mittelwerten festgestellt werden. Die Ergebnisse der Varianzanalyse weisen damit darauf hin, dass das Alter von SuS keinen großen Effekt auf die Ausprägung der zu messenden Konstrukte des DBR-MIS hat. Im Rahmen dieser Arbeit wurden daher keine altersspezifischen Schwellenwerte bestimmt. In umfassenderen Prävalenzstudien konnten, wie in Kapitel 2.2.2 beschrieben, zwar auch keine kontinuierlich abnehmenden oder zunehmenden Prävalenzraten über verschiedene Altersgruppen festgestellt werden, trotzdem zeigten sich teilweise Unterschiede. Im Hinblick darauf und angesichts der Altersgruppeneinteilung in dieser Arbeit, sind zusätzliche Analysen bezüglich des Effekts der Variablen Alter notwendig. So bleibt unklar, ob eine andere Gruppeneinteilung zu einem anderen Ergebnis der Varianzanalyse geführt hätte. Zudem besteht in der vorliegenden Stichprobe eine hohe Altersheterogenität. Bezogen auf den Effekt der Variablen Alter könnten dahingehend Studien mit umfassenderen, altershomogenen Stichprobengruppen notwendig sein, um konkrete Unterschiede zwischen verschiedenen Altersgruppen genauer zu analysieren. In Rückbezug auf die dritte Forschungsfrage dieser Arbeit kann aber festgehalten werden, dass anhand der vorliegenden Stichprobendaten und durchgeführten Analysen lediglich ein statistisch bedeutsamer Geschlechtseffekt auf die Skalen des DBR-MIS festgestellt werden konnte und daher insgesamt allgemeine und geschlechtsspezifische Schwellenwerte für eine Altersspanne von 6 bis 18 Jahren bestimmt wurden (siehe Tabelle 18).

Bei der Betrachtung der ermittelten Schwellenwerte fällt auf, dass sowohl die allgemeinen als auch die geschlechtsspezifischen Schwellenwerte der Kategorien *auffällig* und *grenzwertig* pro Skala jeweils sehr nah beieinander liegen. Ein Extrembeispiel sind an dieser Stelle die Schwellenwerte der Skala VPL für Jungen, bei der der Schwellenwert für die Kategorie *auffällig* (A = 6,3) nur eine Ausprägungsstufe hinter dem Schwellenwert für die Kategorie *grenzwertig* (G = 6,0) liegt. Der Grund dafür liegt in der geringen Anzahl möglicher Ausprägungsstufen, die durch die generelle Abstufung der Ratingskala sowie durch die Auswertung über die Berechnung von Skalenmittelwerten bestimmt wird. Zusätzlich spielt auch die Verteilung der Mittelwerte in der Stichprobe eine entscheidende Rolle dabei, wie die nah die ermittelten

Schwellenwerte beieinander liegen. Problematisch bei diesen Schwellenwerten ist, dass teilweise ein Itempunkt mehr oder weniger den Unterschied zwischen einer Einordnung in die Kategorien *auffällig*, *grenzwertig* oder *unauffällig* ausmacht. Dahingehend ist auch die vergleichende Einordnung von individuellen Testmittelwerten, welche für Skalen mit einem fehlenden Itemwert berechnet werden, anhand der ermittelten Schwellenwerte äußerst fraglich. Für die Einschätzung von Skalenmittelwerten in Orientierung an den in dieser Arbeit bestimmten Schwellen könnte die Vollständigkeit der Itemwerte einer Skala somit eine Voraussetzung darstellen.

Kritische Reflexion

Wie bereits in den vorangegangenen Ausführungen deutlich geworden ist, weist die vorliegende Arbeit einige Limitationen bezüglich des Forschungsvorgehens auf, die in der Betrachtung der Ergebnisse berücksichtigt werden müssen. Ein entscheidender methodischer Faktor, der die Aussagekraft der Ergebnisse maßgeblich einschränkt, ist die vorliegende Stichprobe dieser Forschungsarbeit. Im Hinblick auf die ungleiche Geschlechtsverteilung in der Stichprobe wird deutlich, dass diese angesichts der Population von SuS in Deutschland nicht repräsentativ ist (Rudnicka, 2019). Dies führt zu einer Verzerrung der Ergebnisse, die anhand der Gesamtstichprobe gewonnen wurden, da die Variable Geschlecht, wie mittels der Analysen zur dritten Forschungsfrage bestätigt werden konnte, einen Einfluss auf die Verteilung von Skalenmittelwerten des DBR-MIS hat. Ebenso wie für das Geschlecht, wurde aufgrund verschiedener Prävalenzstudien auch für die Variable Alter ein Effekt auf die Skalenmittelwerte des DBR-MIS angenommen. Diese Annahme konnte anhand von Analysen im Rahmen der dritten Forschungsfrage allerdings nicht bestätigt werden, was wiederum auf die hohe Altersheterogenität in der Stichprobe und die Einteilung in drei Altersgruppen zurückzuführen sein könnte. Zudem ist die vorliegende Stichprobengröße für eine Normierung, für die in der Regel umfassende repräsentative Stichproben ($n > 300$) vorausgesetzt werden, zu gering. Insbesondere die Teilstichproben von Jungen (n = 148) und Mädchen (n = 61) liegen deutlich unter dem geforderten Stichprobenumfang für die Bestimmung von Bezugsnormen. Insgesamt erfolgte die Bestimmung von Schwellenwerten in dieser Arbeit damit also anhand nicht repräsentativer (Teil-)Stichproben, in denen hinsichtlich der Skalenmittelwerte keine Normalverteilungen vorlagen, womit einige grundlegende Voraussetzungen für

eine Normierung verletzt wurden. Die ermittelten allgemeinen und geschlechtsspezifischen Schwellenwerte für die Skalen des DBR-MIS sind daher nur sehr eingeschränkt für die Einordnung individueller Testwerte geeignet. Angesichts des ausgeschriebenen Forschungsinteresses dieser Arbeit können die bestimmten Schwellenwerte für die Kategorien *auffällig* und *grenzwertig* allerdings erste grobe Orientierungspunkte für Lehrkräfte in der Betrachtung und Interpretation von individuellen Testwerten und Verhaltensverläufen darstellen. Diese Folgerung bezieht sich allerdings nur auf den primären Einsatzzweck des DBR-MIS in der Verlaufsdiagnostik, in dessen Orientierung mit den Zielvorgaben für die Schwellenwertbestimmung umgegangen wurde. So können die ermittelten Schwellenwerte für das konkrete Ziel einer sensitiven Identifikation von SuS mit auffälligem Verhalten im Sinne eines Screenings nicht empfohlen werden. Eine solche Zielsetzung würde bei der Schwellenwertbestimmung einen anderen Umgang mit den Zielvorgaben voraussetzen und damit höchstwahrscheinlich zu unterschiedlichen Schwellenwerten führen. Da jedoch generell fraglich ist, ob der DBR-MIS auch für Screening-Zwecke geeignet ist, wäre eine Normierung in diesem Sinne aktuell redundant.

Ebenso wie bei den Ergebnissen der Forschungsfragen 2 und 3 kann auch bei den Analyseergebnissen der erstem Forschungsfrage davon ausgegangen werden, dass die Stichprobengröße und die nicht vorliegende Normalverteilung der Indikatorvariablen in der Stichprobe einen Einfluss auf die ermittelten Modellgütewerte haben. Daher kann anhand dieser Werte, die für beide Messmodelle nicht zufriedenstellend waren, weder das Drei-Faktoren-Modell noch das Sechs-Faktoren-Modell des DBR-MIS grundsätzlich abgelehnt werden. Im Rahmen des Forschungsinteresses dieser Arbeit lag der Fokus allerdings vielmehr auf der Modellselektion, welche über den Vergleich der jeweiligen Fit-Indizes eindeutig vorgenommen werden konnte. Trotzdem bleibt zu beachten, dass für den DBR-MIS hinsichtlich des Sechs-Faktoren-Modells zwar eine gute Reliabilität und eine überwiegend geeignete Skalierung nachgewiesen werden konnte, die Frage nach einer hinreichenden Validität des Instruments aber weiterhin offenbleibt und in weiteren Studien überprüft werden muss.

Bezogen auf die limitierte Belastbarkeit der vorgestellten Forschungsergebnisse durch die vorliegende Stichprobe muss an dieser Stelle noch einmal auf die Rahmenbedingungen dieser Arbeit hingewiesen werden. Da aufgrund der Corona-Pandemie

keine eigenständige Datenerhebung mittels des DBR-MIS möglich war, wurde auf den bereits vorliegenden Datensatz einer Erhebung aus dem Jahr 2018 zurückgegriffen. Die vorliegende Stichprobe war in ihrer Erhebung demnach nicht grundlegend auf das Forschungsinteresse dieser Arbeit ausgelegt, ermöglichte aber unter den gegebenen Umständen die Umsetzung desselben im Rahmen dieser Masterarbeit

7. Fazit und Ausblick

Zusammenfassend betrachtet, konnte das übergeordnete Ziel dieser Arbeit, kriteriale Bezugsnormen in Form von Schwellenwerten für den DBR-MIS zu bestimmen, anhand der durchgeführten Untersuchungen und Analysen umgesetzt werden. Mit dieser Zielsetzung wurde sowohl an die Forschungsarbeiten im Rahmen des Projekts LEVUMI als auch an den bisher unzureichend diskutierten Forschungsbereich bezüglich der Normierung von verhaltensverlaufsdiagnostischen Instrumenten angeknüpft. Im Hinblick auf die nicht repräsentative Stichprobe, mittels deren Daten die Analysen zur Beantwortung der drei Forschungsfragen durchgeführt wurden, sind die vorliegenden Ergebnisse und bestimmten Schwellenwerte für die Skalen des DBR-MIS nur eingeschränkt zu betrachten und wissenschaftlich wenig belastbar. Allerdings liegen bisher auch keine vergleichbaren Forschungsarbeiten zum DBR-MIS vor, die beispielsweise für eine weiterführende Interpretation und Diskussion der ermittelten Werte herangezogen werden könnten. In diesem Sinne stellen die Ergebnisse dieser Arbeit trotz Limitationen einen wichtigen Forschungsbeitrag dar, anhand dessen weitere Erkenntnisse über den DBR-MIS und Ansätze zu weiteren Forschungsarbeiten gewonnen werden können.

Die Ergebnisse der konfirmatorischen Faktorenanalysen und Reliabilitätsanalysen im Rahmen der ersten Forschungsfrage weisen z. B. auf vereinzelt fragliche Messzusammenhänge zwischen Skalen und Items des DBR-MIS hin, die die Notwendigkeit weiterer Überprüfungen und Anpassungen zur Verbesserung des Instruments anzeigen. Die Ergebnisse der einfaktoriellen Varianzanalyse bezüglich der Überprüfung des Einflusses der Variablen Alter auf die Skalenmittelwerte des DBR-MIS liefern zudem den Hinweis, dass in nachfolgenden Arbeiten zur Schwellenwertbestimmung altershomogenere Stichproben sinnvoll sind, um Alterseffekte konkreter analysieren zu können. Insgesamt kann angenommen werden, dass die ermittelten allgemeinen und geschlechtsspezifischen Schwellenwerte aufgrund von Verzerrungen durch die nicht repräsentative Stichprobe nur bedingt als kriteriale Bezugsnormen für den DBR-MIS eingesetzt werden können. So können sie für Lehrkräfte bei der Interpretation von individuellen Testwerten des DBR-MIS als grobe Richtwerte dienen, ermöglichen aber keine sensitive Identifikation von SuS mit auffälligem Verhalten.

Mit der Bestimmung von Schwellenwerten für den DBR-MIS ist allerdings noch nicht die Frage geklärt, ob eine solche kriteriale Bezugsnorm den diagnostischen Wert des Instruments für den Einsatz in der schulischen Praxis steigert. Vielmehr ist die Bereitstellung dieser Bezugsnorm im Hinblick auf das von Casale et al. (2015b) genannte Gütekriterium der individuellen Bezugsnorm für Instrumente der Verhaltensverlaufsdiagnostik (vgl. Kapitel 2.3.2) besonders kritisch zu sehen. Dahingehend sind weiterführende Forschungsarbeiten notwendig, die die Anwendung des DBR-MIS, mit Fokus auf die Interpretation der individuellen Testergebnisse unter Berücksichtigung der allgemeinen und geschlechtsspezifischen Schwellenwerte, untersuchen. Hinsichtlich dieses Forschungsinteresses wäre z. B. eine weitere Erprobung des DBR-MIS durch Lehrkräfte in schulischen Kontexten und eine daran anschließende qualitative Erhebung der Erfahrungen und Meinungen dieser Lehrkräfte zu der Arbeit mit den Schwellenwerten denkbar. Weitere Forschungsarbeiten in diese Richtung könnten damit erste Ergebnisse bezogen auf die Frage, ob Vergleichsnormen für verhaltensverlaufsdiagnostische Instrumente überhaupt sinnvoll und notwendig sind, liefern. Auch im Hinblick auf die Frage, ob kriteriale Bezugsnormen für verhaltensverlaufsdiagnostische Instrumente in Form von Schwellenwerten oder in Form von Entwicklungsverlaufslinien aussägekräftiger und für die Interpretation von Testergebnissen hilfreicher sind, können neue Erkenntnisse nur über weitere umfassende Studien gewonnen werden.

Abschließend kann festgehalten werden, dass die Ergebnisse dieser Arbeit neue Fragenstellungen aufwerfen und damit den hohen Forschungsbedarf bezüglich der Bereitstellung von Bezugsnormen für verhaltensverlaufsdiagnostische Instrumente verdeutlichen. Auch wenn sich die vorliegenden Ergebnisse ausschließlich auf den DBR-MIS beziehen und damit keine generalisierbaren Aussagen zu DBR oder anderen verhaltensverlaufsdiagnostischen Instrumenten ermöglichen, stellt die dahinterliegende Forschung im Rahmen des Projekts LEVUMI einen wichtigen Beitrag für die Weiterentwicklung und Etablierung der Verhaltensverlaufsdiagnostik im deutschsprachigen Raum dar. Insbesondere die Befunde zur Skalenstruktur und Reliabilität des DBR-MIS weisen darauf, dass dieser ein vielversprechendes Instrument für die Erhebung von Verhaltensverläufen von SuS in schulischen Kontexten darstellt, jedoch weiterhin Verbesserungspotenzial besteht.

Literaturverzeichnis

Altendorfer-Kling, U., Ardelt-Gattinger, E. & Thun-Hohenstein, L. (2007). Der Selbstbeurteilungsbogen des SDQ anhand einer österreichischen Feldstichprobe. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie,* 35 (4), 265–271.

Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland 2018. Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung.* Verfügbar unter https://www.bildungsbericht.de/de/bildungsberichte-seit-2006/bildungsbericht-2018/pdf-bildungsbericht-2018/bildungsbericht-2018.pdf [14.05.2020]

Barkmann, C. & Schulte-Markwort, M. (2004). Prävalenz psychischer Auffälligkeiten bei Kindern und Jugendlichen in Deutschland - ein systematischer Literaturüberblick. *Psychiatrische Praxis,* 31 (6), 278–287.

Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderung. (2017). *Die UN-Behindertenrechtskonvention. Übereinkommen über die Rechte von Menschen mit Behinderung*. Verfügbar unter https://www.behindertenbeauftragte.de/SharedDocs/Publikationen/UN_Konvention_deutsch.pdf?__blob=publicationFile&v=2 [14.05.2020]

Bierhoff, H. (2002). Theorien hilfreichen Verhaltens. In D. Frey & M. Irle (Hrsg.), *Theorien der Sozialpsychologie*, Band 2 (2. Aufl.) (S. 178-197). Bern: Huber.

Blumenthal, Y. & Hartke, B. (2015). Der Response to Intervention-Ansatz. Eine Grundlage für ein präventives und inklusives Beschulungskonzept. In R. Krüger & C. Mähler (Hrsg.), *Gemeinsames Lernen in inklusiven Klassenzimmern. Prozesse der Schulentwicklung gestalten* (S. 49-61). Köln: Link.

Brezinka, V. (2003). Zur Evaluation zur Präventivintervention für Kinder mit Verhaltensstörungen. *Kindheit und Entwicklung,* 12 (2), 71–83.

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

Briesch, A. M., Chafouleas, S. & Riley-Tillman, T. Chris (Eds.). (2016). *Direct Behavior Rating. Linking Assessment, Communication, and Intervention*. New York, London: The Guilford Press.

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

Bundschuh, K. & Winkler, C. (2019). *Einführung in die sonderpädagogische Diagnostik* (9. überarb. Aufl.). München: Ernst Reinhardt Verlag.

Casale, G. (2017). „Nützt es was oder nützt es nichts?“ Direct Behavior Rating (DBR) als diagnostische Methode zur zeitnahen Überprüfung des Fördererfolgs bei unterrichtlichem Schülerinnen- und Schülerverhalten. *Potsdamer Zentrum für empirische* Inklusionsforschung, Nr. 1. Verfügbar unter https://www.uni-potsdam.de/fileadmin/projects/inklusion/PDFs/ZEIF-Blog/Casale_2017_Direct_Behavior_Rating.pdf [20.05.2020]

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

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

Casale, G., Hennemann, T., Huber, C. & Grosche, M. (2015b). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung*, 41 (1), 37–53.

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

Chafouleas, S., Jaffery, R., Riley-Tillman, T. C., Christ, T. J. & Sen, R. (2013): The Impact of Target, Wording, and Duration on Rating Accuracy for Direct Behavior Rating. *Assesment for Effective Intervention*, 39 (1), 39–53.

Christ, T. J., Riley-Tillman, T. C. & Chafouleas, S. (2009). Foundation for the developement and use of Direct Behavior Rating (DBR) to asses and evaluate student behavior. *Assesment for Effective Intervention*, 34, 201–213.

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

DeVries, J. M., Rathmann, K. & Gebhardt, M. (2018). How Does Social Behavior Relate to Both Grades and Achievement Scores? *Frontiers in Psycholoy*, 9. Verfügbar unter https://doi.org/10.3389/fpsyg.2018.00857 [20.05.2020]

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

Dorsch, F., Wirtz, M. A. & Strohmer, J. (Hrsg.). (2014). *Dorsch - Lexikon der Psychologie* (17. überarb. Aufl.). Bern: Huber.

Edlund, C. V. (1969). Rewards at home to promote desirable behavior. *Teaching Exceptional Children,* 1, 121-127.

Flott-Tönjes, U., Albers, S., Ludwig, M., Schumacher, H., Storcks-Kemming, B., Thamm, J. & Witt, H. (2017). *Fördern planen. Ein sonderpädagogisches Planungs- und Beratungskonzept für Förderschulen und Schulen des gemeinsamen Lernens*. Oberhausen: Athena.

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

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

Gebhardt, M., Diehl, K. & Mühling, A. (2015a). Online-Lernverlaufsmessung für alle Schülerinnen und Schüler in inklusiven Klassen. *Zeitschrift für Heilpädagogik*, 66, 444–453.

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

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

Gehrke, M. (2019). *Angewandte empirische Methoden in Finance & Accounting. Umsetzung mit R*. Berlin, Boston: De Gruyter Oldenbourg.

Gold, A., Gawrilow, C. & Hasselhorn, M. (2016). Grundlagen schulpsychologischer Diagnostik. In K. Seifried, S. Drewes & M. Hasselhorn (Hrsg.), *Handbuch Schulpsychologie. Psychologie für die Schule* (2. überarb. Aufl.) (S. 117-127). Stuttgart: Kohlhammer.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire. A research note. *Journal of Child Psychology and Psychiatry*, 38, 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 Strengts and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of abnormal child psychology*, 38 (8), pp. 1179–1191. Verfügbar unter https://doi.org/10.1007/s10802-010-9434-x

Götz, J. & Hauenschild, K. (2015). Didaktische Brennpunkte inklusiven Unterrichts. In R. Krüger & C. Mähler (Hrsg.), *Gemeinsames Lernen in inklusiven Klassenzimmern. Prozesse der Schulentwicklung gestalten* (S. 39-48). Köln: Link.

Haller, A., Klasen, F., Petermann, F., Barkmann, C., Otto, C., Schlack, R. & Ravens-Sieberer, U. (2016). Langzeitfolgen externalisierender Verhaltensauffälligkeiten. Ergebnisse der BELLA-Kohortenstudie. *Kindheit und Entwicklung*, 25 (2), 31–40.

Hartke, B. (2017). Gelingende Inklusion. Das Rügener Inklusionsmodell (RIM). In B. Hartke (Hrsg.), *Handlungsmöglichkeiten Schulische Inklusion. Das Rügener Modell kompakt* (S. 11-19). Stuttgart: Kohlhammer.

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

Hascher, T. (2005). Diagnostizieren in der Schule. In A. Bartz, C. Kloft, J. Fabian, S. Huber, H. Rosenbusch & H. Sassenscheidt (Hrsg.), *PraxisWissen SchulLeitung. Basiswissen und Arbeitshilfen zu den zentralen Handlungsfeldern der Schulleitung* (S. 1-8). Kronach: CarlLink.

Hesse, I. & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3. überarb. & erweiterte Aufl.). Opladen, Toronto: Verlag Barbara Budrich.

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

Himme, A. (2007): Gütekriterien der Messung: Reliabilität, Validität und Generalisierbarkeit. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (2. überarb. & erweiterte Aufl.) (S. 375-390). Wiesbaden: Gabler.

Hisker, S. (2018): *Veränderungen im Direct Behavior Rating (DBR) über die Zeit. Eine Pilotierung mit fünf Messzeitpunkten in Grund- und Gesamtschulen*. Masterarbeit, Technische Universität Dortmund.

Hölling, H., Schlack, R., Petermann, F., Ravens-Sieberer, U. & Mauz, E. (2014). Psychische Auffälligkeiten und psychosoziale Beeinträchtigungen bei Kindern und Jugendlichen im Alter von 3 bis 17 Jahren in Deutschland – Prävalenz und zeitliche Trends zu 2 Erhebungszeitpunkten (2003–2006 und 2009–2012). Ergebnisse der KiGGS-Studie – Erste Folgebefragung (KiGGS Welle 1). *Bungesgesundheitsblatt - Gesundheitsforschung – Gesundheitsschutz*, 57 (7), 807–819.

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

Ihle, W. & Esser, G. (2002). Epidemiologie psychischer Störungen im Kindes- und Jugendalter: Prävalenz, Verlauf, Komorbidität und Geschlechtsunterschiede. *Psychologische Rundschau*, 53 (4), 159–169.

Jantzer, V., Haffner, J., Parzer, P., Roos, J., Steen, R. & Resch, F. (2012): Der Zusammenhang von ADHS, Verhaltensproblemen und Schulerfolg am Beispiel der Grundschulempfehlung. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 61 (9), 662–676. Verfügbar unter https://psydok.psycharchives.de/jspui/bitstream/20.500.11780/3687/1/Jantzer_PdKK_2012_9.pdf

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

Klipker, K., Baumgarten, F., Göbel, K., Lampert, T. & Hölling, H. (2018). Psychische Auffälligkeiten bei Kindern und Jugendlichen in Deutschland - Querschnittergebnisse aus KiGGS Welle 2 und Trends. *Journal of Health Monitoring*, 3 (3), 37–45.

Kowalewski, C. (2009). *Psychische Störungen bei Kindern und Jugendlichen in ambulanter kinder- und jugendpsychiatrischer Versorgung. Vergleichende Analyse zur Diagnosenverteilung im Klientel kinder- und jugendpsychiatrischer Praxen in Deutschland*. Inaugural-Dissertation. Philipps-Universität Marburg. Verfügbar unter http://archiv.ub.uni-marburg.de/diss/z2009/0605/pdf/dck.pdf [30.05.2020]

Krause, S. (2019): *Verhaltensverlaufsdiagnostik in der Praxis. Untersuchungen zur Implementierung des Direct Behavior Ratings „PUTSIE“ in einer inklusiven Grundschule*. Masterarbeit, Technische Universität Dortmund.

Kultusministerkonferen. (2000). *Empfehlungen zum Förderschwerpunkt emotionale und soziale Entwicklung*. Verfügbar unter https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2000/2000_03_10-FS-Emotionale-soziale-Entw.pdf [25.05.2020]

Kutscher, N. (2008). Heterogenität. In T. Coelen & H. Otto (Hrsg.), *Grundbegriffe Ganztagsbildung. Das Handbuch* (S. 61-70). Wiesbaden: VS Verlag für Sozialwissenschaften/ GWV Fachverlage GmbH Wiesbaden.

LeBel, T. J., Kilgus, S. P., Briesch, A. M. & 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.

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

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

Mähler, C. & Krüger, R. (2015). Einführung. In R. Krüger und C. Mähler (Hrsg.): *Gemeinsames Lernen in inklusiven Klassenzimmern* (S. XI-XVI). *Prozesse der Schulentwicklung gestalten*. Köln: Link.

Mühling, A., Gebhardt, M. & Diehl, K. (2017). Formative Diagnostik durch die Onlineplattform LEVUMI. *Informatik-Spektrum*, 40 (6), 556–561.

Myschker, N. & Stein, R. (2018). *Verhaltensstörungen bei Kindern und Jugendlichen. Erscheinungsformen - Ursachen - hilfreiche Maßnahmen* (8. erweiterte & aktualisierte Aufl.). Stuttgart: Kohlhammer.

Oerter, R. (2002). *Entwicklungspsychologie* (5. überarb. Aufl.). Weinheim: Beltz.

Petermann, F. & Lehmkuhl, G. (2010). Prävention von Aggression und Gewalt. *Kindheit und Entwicklung*, 19 (4), 239–244.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2006a): *Quantitative Methoden. Einführung in die Statistik* (2. erweiterte Aufl.). Berlin, Heidelberg: Springer Medizin Verlag.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2006b). *Quantitative Methoden 2. Einführung in die Statistik* (2. erweiterte Aufl.). Berlin, Heidelberg: Springer Medizin Verlag.

Rendtorff, B. (2014). Heterogenität und Differenz. Über Banalisierung von Begriffen und den Verlust ihrer Produktivität. In H. Koller, R. Casale & N Ricken (Hrsg.), *Heterogenität. Zur Konjunktur eines pädagogischen Konzepts* (S. 115-130). Paderborn: Schöningh.

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

Riley-Tillman, T. C., Christ, T. J., Chafouleas, S., Boice-Mallach, C. H. & Briesch, A. M. (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.

Rudnicka, J. (2019). *Anzahl der Schüler an allgemeinbildenden Schulen in Deutschland im Schuljahr 2018/2019 nach Schulart und Geschlecht*. Verfügbar unter https://de.statista.com/statistik/daten/studie/150544/umfrage/anzahl-der-schueler-nach-schularten-im-schuljahr-2008-2009/ [21.07.2020]

Saalfrank, W. & Zierer, K. (2017). *Inklusion.* Paderborn: Ferdinand Schöningh.

Sauerland, A. (2018): *Konstruktion eines Direct Behavior Rating*. Masterarbeit Technische Universität Dortmund.

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M. & Chafouleas, S. M. (2009). The impact of training on the accuracy of Direct Behavior Ratings

(DBR). *School Psychology Quarterly*, 24 (2), 73– 83. Verfügbar unter http://dx.doi.org/10. 1037/a0016255

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

Schuchardt, K. (2015). Heterogenität und interindividuelle Unterschiede beim Lernen. In R. Krüger & C. Mähler (Hrsg.), *Gemeinsames Lernen in inklusiven Klassenzimmern. Prozesse der Schulentwicklung gestalten* (S. 3-12). Köln: Link.

Schurig, M., Jungjohann, J. Gebhardt, M. (2019). *Handbuch für Lehrkräfte im Anwendungsbereich Verhalten und Empfinden. Lern-Verlaufs-Monitoring LEVUMI*. Technische Universität Dortmund. Verfügbar unter http://dx.doi.org/10.17877/DE290R-20376.

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

Souvignier, E., Förster, N. & Zeuch, N. (2016). Lernverlaufsdiagnostik. In K. Seifried, S. Drewes & M. Hasselhorn (Hrsg.), *Handbuch Schulpsychologie. Psychologie für die Schule* (2. überarb. Aufl.) (S. 140-149). Stuttgart: Kohlhammer.

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

Stein, R. & Müller, T. (2015a). Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung. Zur Einleitung. In R. Stein & T. Müller (Hrsg.), *Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung* (S. 11-18). Stuttgart: Kohlhammer.

Stein, R. & Müller, T. (2015b). Verhaltensstörungen und emotional-soziale Entwicklung: zúm Gegenstand. In R. Stein & T. Müller (Hrsg.), I*nklusion im Förderschwerpunkt emotionale und soziale Entwicklung* (S. 19-43). Stuttgart: Kohlhammer.

Strathmann, A.M. & Klauer, K. J. (2012). *LVD-M 2-4. Lernverlaufsdiagnostik-Mathematik für zweite bis vierte Klassen*. Göttingen: Hogrefe.

Tenorth, H. & Tippelt, R. (Hrsg.). (2012). *Beltz Lexikon Pädagogik.* Weinheim: Beltz.

Tschira, A. (2005). *Wie Kinder lernen - und warum sie es manchmal nicht tun. Über die Spielregeln zwischen Mensch und Umwelt im Lernprozess* (2. überarb. Aufl.). Heidelberg: Carl-Auer-Systeme-Verlag.

Ophuysen, S. van & Lintorf, K. (2013). Pädagogische Diagnostik im Schulalltag. In S. Beutel, W. Bos & R. Porsch (Hrsg.), *Lernen in Vielfalt. Chance und Herausforderung für Schul- und Unterrichtsentwicklung* (S. 55-76). Münster: Waxmann.

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

Voß, S. (2017). Datenbasierte Förderentscheidungen. In B. Hartke (Hrsg.): *Handlungsmöglichkeiten Schulische Inklusion. Das Rügener Modell kompakt* (S. 33-56). Stuttgart: Kohlhammer.

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

Voß, S. & Gebhardt, M. (2017b). Verlaufsdiagnostik in der Schule. *Empirische Sonderpädagogik*, (2), S. 95–97. Verfügbar unter https://www.researchgate.net/publication/320517829_Verlaufsdiagnostik_in_der_Schule

Walgenbach, K (2014). *Heterogenität – Intersektionalität – Diversity in der Erziehungswissenschaft*. Opladen: Barbara Budrich Verlag.

Walter,J. (2013). *VSL. Verlaufsdiagnostik sinnerfassenden Lesens*. Göttingen: Hogrefe.

Weiber, R. & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS* (2. erweiterte Aufl.). Berlin, Heidelberg: Springer.

Wember, Franz B.; Stein, Roland; Heimlich, Ulrich (Hg.) (2014): Handlexikon Lernschwierigkeiten und Verhaltensstörungen. 1. Aufl. Stuttgart: Kohlhammer.

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

Anhang A

Ratingskala zur Verhaltensdiagnostik in der Schule (DBR-MIS)

Nr.	Items	Nie						Immer
Externalisierendes Verhalten (EXT)								
Störendes und auflehnendes Verhalten (SAV)								
1	Verhält sich wütend und aufbrausend	1	2	3	4	5	6	7
2	Missachtet Regeln und hört nicht auf die Lehrkraft	1	2	3	4	5	6	7
3	Streitet sich mit Mitschüler_innen/provoziert durch eigenes Verhalten seine Mitschüler_innen	1	2	3	4	5	6	7
Verhaltensprobleme beim Lernen (VPL)								
4	Zappelt, ist (motorisch) unruhig/ überaktiv	1	2	3	4	5	6	7
5	Bricht Aufgaben häufig früh ab	1	2	3	4	5	6	7
6	Lässt sich schnell und leicht ablenken	1	2	3	4	5	6	7
Internalisierendes Verhalten (INT)								
Depressives und ängstliches Verhalten (DAV)								
7	Wirkt besorgt, betrübt oder bedrückt	1	2	3	4	5	6	7
8	Wirkt ängstlich/ fürchtet sich	1	2	3	4	5	6	7
9	Wirkt nervös (sucht Nähe zu Erwachsenen)	1	2	3	4	5	6	7
Probleme in sozialen Interaktionen (PSI)								
10	Arbeitet/spielt meist alleine	1	2	3	4	5	6	7
11	Wird von Mitschüler_innen gehänselt oder geär-gert, lässt sich provozieren	1	2	3	4	5	6	7
12	Arbeitet/spielt häufiger mit Erwachsenen als mit Mitschüler_innen	1	2	3	4	5	6	7
Positives Schulverhalten (PSV)								
Schulbezogenes Verhalten (SV)								
13	Meldet sich im Unterricht	1	2	3	4	5	6	7
14	Hält sich an Gesprächsregeln	1	2	3	4	5	6	7
15	Richtet Aufmerksamkeit/Konzentration auf die Be-arbeitung der Aufgabe	1	2	3	4	5	6	7
16	Arbeitet ruhig am Platz und verweigert nicht die Mitarbeit	1	2	3	4	5	6	7
Prosoziales Verhalten (PS)								
17	Verhält sich anderen gegenüber rücksichtsvoll	1	2	3	4	5	6	7
18	Verhält sich anderen gegenüber hilfsbereit	1	2	3	4	5	6	7
19	Verhält sich in Partner- und Gruppensituationen kooperativ	1	2	3	4	5	6	7

(Quelle: Schurig et al., 2019, S. 5)

Anhang B

Prozentrangzuordnungen der Mittelwerte von Jungen pro Skala

Tabelle 19: Zuordnung der Mittelwerte von Jungen der Skala SAV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	22	15,7%	15,7%
1,3	11	7,9%	23,6%
1,6	10	7,1%	30,7%
2	13	9,3%	40,0%
2,3	7	5,0%	45,0%
2,6	8	5,7%	50,7%
3	9	6,4%	57,1%
3,3	4	2,9%	60,0%
3,6	6	4,3%	64,3%
4	9	6,4%	70,7%
4,3	7	5,0%	75,7%
4,6	8	5,7%	81,4%
5	**2**	**1,4%**	**82,9%**
5,3	4	2,9%	85,7%
5,6	5	3,6%	89,3%
6	**10**	**7,1%**	**96,4%**
6,3	4	2,9%	99,3%
7	1	0,7%	100,0%

Anmerkungen. *n = 140,* Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 20: Zuordnung der Mittelwerte von Jungen der Skala VPL zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	8	5,6%	5,6%
1,3	1	0,7%	6,3%
1,6	5	3,5%	9,8%
2	5	3,5%	13,3%
2,3	6	4,2%	17,5%
2,6	8	5,6%	23,1%
3	11	7,7%	30,8%
3,3	7	4,9%	35,7%
3,6	7	4,9%	40,6%
4	7	4,9%	45,5%
4,3	13	9,1%	54,5%
4,6	11	7,7%	62,2%
5	12	8,4%	70,6%
5,3	10	7,0%	77,6%
5,6	12	8,4%	86,0%
6	**10**	**7,0%**	**93,0%**
6,3	**7**	**4,9%**	**97,9%**
6,6	2	1,4%	99,3%
7	1	0,7%	100,0%

Anmerkungen. *n* = 143, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 21: Zuordnung der Mittelwerte von Jungen der Skala DAV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	29	21,3%	21,3%
1,3	8	5,9%	27,2%
1,6	17	12,5%	39,7%
2	15	11,0%	50,7%
2,3	9	6,6%	57,4%
2,6	6	4,4%	61,8%
3	8	5,9%	67,6%
3,3	8	5,9%	73,5%
3,6	6	4,4%	77,9%
4	**7**	**5,1%**	**83,1%**
4,3	4	2,9%	86,0%
4,6	6	4,4%	90,4%
5	**3**	**2,2%**	**92,6%**
5,3	1	0,7%	93,4%
5,6	5	3,7%	97,1%
6	1	0,7%	97,8%
6,3	1	0,7%	98,5%
6,6	1	0,7%	99,3%
7	1	0,7%	100,0%

Anmerkungen. *n* = 136, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 22: Zuordnung der Mittelwerte von Jungen der Skala PSI zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
0,6	2	1,6%	1,6%
1	18	14,2%	15,7%
1,3	7	5,5%	21,3%
1,6	8	6,3%	27,6%
2	10	7,9%	35,4%
2,3	7	5,5%	40,9%
2,6	10	7,9%	48,8%
3	10	7,9%	56,7%
3,3	31	16,5%	73,2%
3,6	12	9,4%	82,7%
4	**5**	**3,9%**	**86,6%**
4,3	3	2,4%	89,0%
4,6	**3**	**2,4%**	**91,3%**
5	3	2,4%	93,7%
5,3	2	1,6%	95,3%
5,6	4	3,1%	98,4%
6	2	1,6%	100,0%

Anmerkungen. n = 127, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 23: Zuordnung der Mittelwerte von Jungen der Skala SV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1,5	1	0,8%	0,8%
1,75	3	2,5%	3,3%
2	**7**	**5,7%**	**9,0%**
2,25	4	3,3%	12,3%
2,5	**9**	**7,4%**	**19,7%**
2,75	6	4,9%	24,6%
3	12	9,8%	34,4%
3,25	7	5,7%	40,2%
3,5	14	11,5%	51,6%
3,75	10	8,2%	59,8%
4	5	4,1%	63,9%
4,25	4	3,3%	67,2%
4,5	5	4,1%	71,3%
4,75	9	7,4%	78,7%
5	4	3,3%	82,0%
5,25	4	3,3%	85,2%
5,5	4	3,3%	88,5%
5,75	4	3,3%	91,8%
6	1	0,8%	92,6%
6,25	5	4,1%	96,7%
6,5	1	0,8%	97,5%
6,75	1	0,8%	100,0%

Anmerkungen. *n* = 122, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 24: Zuordnung der Mittelwerte von Jungen der Skala zu PS zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	6	4,3%	4,3%
1,3	3	2,2%	6,5%
1,6	**7**	**5,1%**	**11,6%**
2	7	5,1%	16,7%
2,3	**8**	**5,8%**	**22,5%**
2,6	7	5,1%	27,5%
3	8	5,8%	33,3%
3,3	9	6,5%	39,9%
3,6	9	6,5%	46,4%
4	11	8,0%	54,3%
4,3	5	3,6%	58,0%
4,6	7	5,1%	63,0%
5	7	5,1%	68,1%
5,3	8	5,8%	73,9%
5,6	7	5,1%	79,0%
6	12	8,7%	87,7%
6,3	2	1,4%	89,1%
6,6	6	4,3%	93,5%
7	9	6,5%	100,0%

Anmerkungen. *n* = 138, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Anhang C

Prozentrangzuordnung der Mittelwerte von Mädchen pro Skala

Tabelle 25: Zuordnung der Mittelwerte von Mädchen der Skala SAV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
0,6	1	1,7%	1,7%
1	22	36,7%	38,3%
1,3	6	10,0%	48,3%
1,6	8	13,3%	61,7%
2	3	5,0%	66,7%
2,3	3	5,0%	71,7%
3	5	8,3%	80,0%
3,3	**2**	**3,3%**	**83,3%**
3,6	3	5,0%	88,3%
4	1	1,7%	90,0%
4,6	**2**	**3,3%**	**93,3%**
5	3	5,0%	98,3%
5,3	1	1,7%	100,0%

Anmerkungen. n = 60, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 26: Zuordnung der Mittelwerte von Mädchen der Skala VPL zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1	10	17,2%	17,2%
1,3	5	8,6%	25,9%
1,6	5	8,6%	34,5%
2	5	8,6%	43,1%
2,3	4	6,9%	50,0%
2,6	7	12,1%	62,1%
3	6	10,3%	72,4%
3,3	2	3,4%	75,9%
4	3	5,2%	81,0%
4,3	**2**	**3,4%**	**84,5%**
4,6	2	3,4%	87,9%
5	2	3,4%	91,4%
5,6	**2**	**3,4%**	**94,8%**
6	1	1,7%	96,6%
6,3	2	3,4%	100,0%

Anmerkungen. *n* = 58, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 27: Zuordnung der Mittelwerte von Mädchen der Skala DAV zu Prozenträngen

Mittel-wert	Anzahl	Anzahl in %	Prozent-rang
1	14	24,1%	24.1%
1,3	3	5,2%	29.3%
1,6	6	10,3%	39.7%
2	9	15,5%	55.2%
2,3	4	6,9%	62.1%
2,6	4	6,9%	69.0%
3	2	3,4%	72.4%
3,3	2	2,4%	75.9%
3,6	1	1,7%	77.6%
4	1	1,7%	79.3%
4,3	**3**	**5,2%**	**84.5%**
4,6	2	3,4%	87.9%
5	1	1,7%	89.7%
5,3	**2**	**3,4%**	**93.1%**
6	2	3,4%	96.6%
6,3	1	1,7%	98.3%
6,6	1	1,7%	100.0%

Anmerkungen. n = 58, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 28: Zuordnung der Mittelwerte von Mädchen der Skala PSI zu Prozenträngen

Mittel-wert	Anzahl	Anzahl in %	Prozent-rang
0,6	2	3,8%	3,8%
1	9	17,3%	21,2%
1,3	1	1,9%	23,1%
1,6	4	7,7%	30,8%
2	3	5,8%	36,5%
2,3	7	13,5%	50,0%
2,6	2	3,8%	53,8%
3	4	7,7%	61,5%
3,3	10	19,2%	80,8%
3,6	**2**	**3,8%**	**84,6%**
4	2	3,8%	88,5%
4,3	**3**	**5,8%**	**94,2%**
5,3	1	1,9%	96,2%
5,6	1	1,9%	98,1%
7	1	1,9%	100,0%

Anmerkungen. *n* = 52, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 29: Zuordnung der Mittelwerte von Mädchen der Skala SV zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
2,25	1	1,9%	1,9%
2,5	1	1,9%	3,8%
2,75	**4**	**7,5%**	**11,3%**
3,25	2	3,8%	15,1%
3,5	**4**	**7,5%**	**22,6%**
3,75	4	7,5%	30,2%
4	4	7,5%	37,7%
4,25	3	5,2%	43,4%
4,5	4	7,5%	50,9%
4,75	2	3,8%	54,7%
5	6	11,3%	66,0%
5,25	3	5,7%	71,7%
5,5	2	3,8%	75,5%
5,75	3	7,5%	83,0%
6	3	5,7%	88,7%
6,25	1	1,9%	90,6%
6,75	3	5,7%	96,2%
7	2	3,8%	100,0%

Anmerkungen. *n* = 53, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Tabelle 30: Zuordnung der Mittelwerte von Mädchen der Skala PS zu Prozenträngen

Mittelwert	Anzahl	Anzahl in %	Prozentrang
1,3	1	1,8%	1,8%
2	1	1,8%	3,6%
2,6	2	3,6%	7,1%
3	**1**	**1,8%**	**8,9%**
3,3	4	7,1%	16,1%
3,6	**2**	**3,6%**	**19,6%**
4	4	8,9%	28,6%
4,6	3	5,4%	33,9%
5	5	8,9%	42,9%
5,3	3	5,4%	48,3%
5,6	6	10,7%	58,9%
6	7	12,5%	71,4%
6,3	6	10,7%	82,1%
6,6	1	1,8%	83,9%
7	9	16,1%	100,0%

Anmerkungen. *n* = 56, Farbliche Markierungen nach Kategorien: grün = unauffälliger Bereich, gelb = grenzwertiger Bereich, rot = auffälliger Bereich.

Eidesstattliche Versicherung
(Affidavit)

Weniger, Franziska
Name, Vorname
(Last name, first name)

176280
Matrikelnr.
(Enrollment number)

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

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

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

Ist das schon auffällig? Bestimmung kriterialer Bezugsnormen für den Einsatz des verhaltensverlaufsdiagnostischen Instruments DBR-UIS in der schulischen Praxis

*Nichtzutreffendes bitte streichen
(Please choose the appropriate)

Witten, 29.07.2020
Ort, Datum
(Place, date)

F. Weniger
Unterschrift
(Signature)

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

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

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

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

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

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

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

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

Witten, 29.07.2020
Ort, Datum
(Place, date)

F. Weniger
Unterschrift
(Signature)

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