Masterarbeit

Verhaltensverlaufsdiagnostik in der Praxis

Untersuchungen zur Implementierung des Direct Behavior Ratings „PUTSIE“
in einer inklusiven Grundschule

Technische Universität Dortmund

Fakultät der Rehabilitationswissenschaften

Entwicklung und Erforschung inklusiver Bildungsprozesse

Erstgutachter: Prof. Dr. Markus Gebhardt

Zweitgutachter: Dr. Michael Schurig

Vorgelegt von: Silas Krause

E-Mail: silas.krause@tu-dortmund.de

Vorgelegt am: 13.02.19

Inhaltsverzeichnis

1. Einleitung

Die Förderung von Schülerinnen und Schülern mit Verhaltensschwierigkeiten in der Inklusion steht seit Jahren in der Kritik. Dies wird zum Beispiel anhand folgenden Zitats ersichtlich: „Zahlreiche internationale Berichte und Studien weisen auf die Schwierigkeiten einer erfolgreichen Inklusion von Kindern und Jugendlichen mit Verhaltensstörungen hin." (Hennemann et al. 2015, S. 115) Eine Annahme, dieses Problem zu beheben besteht darin, die Passung individueller Bedürfnisse des Schülers besser erfassen zu können und diese im schulischen Angebot zu berücksichtigen. Hierfür sind diagnostische Instrumente notwendig, die das Verhalten über die Zeit erfassen können und den Lehrkräften ein brauchbares Feedback über die Passung ihrer Förderung zum Schüler geben. Die Förderung wird auf diese Weise evidenzbasiert, da die Lehrkräfte dessen Wirksamkeit mit Instrumenten der Wissenschaft überprüfen können. Leider gibt es derzeit jedoch nur wenige Instrumente, die eine solche Art der Diagnostik ermöglichen (vgl. ebd.). Eine Vorgehensweise könnte in der Nutzung sogenannter Direct Behavior Ratings (Kapitel 3) liegen, mit denen das Verhalten der Schüler in kurzen Beobachtungssequenzen regelmäßig erhoben wird, sodass eine Diagnostik des Verhaltens über die Zeit und die Evaluation von Interventionen ermöglicht werden könne. Die psychometrische Güte einzelner Instrumente scheint hierbei schon erwiesen, leider bleibt jedoch die Handhabbarkeit und die Betrachtung möglicher Bedingungen und Hindernisse für eine gelingende Implementierung in bisherigen Untersuchungen weitestgehend unberücksichtigt.

Mit der Arbeit wird daher das Ziel verfolgt, mit Hilfe des selbst entwickelten Direct Behavior Ratings „PUTSIE" die Implementierung eines Direct Behavior Ratings an einem inklusiven Schulsystem genauer zu untersuchen, um mögliche Gelingensbedingungen und Hindernisse bei der Implementierung festzustellen. Die so erlangten Informationen können bei der Weiterentwicklung von Ratinginstrumenten genutzt werden. Hierbei ist zu betonen, dass die Arbeit innerhalb des Projekts „LEVUMI" (https://www.levumi.de) verfasst wird, welches sich genauer mit der Entwicklung von kostenlosen, onlinebasierten Lern- und Verhaltensverlaufsdiagnostischen Instrumenten beschäftigt.

Die Arbeit hat daher folgenden Aufbau: In einem theoretischen Teil A werden zunächst Annahmen der pädagogischen Diagnostik, Verlaufsdiagnostik und hierauf bezogene Inhalte und Annahmen zum Verhalten und inklusivens Schulsystemen genauer dargelegt. Mit dem Direct Behavior Rating wird ein unter Bezug auf theoretische Annahmen möglicherweise geeignetes Instrument zur Verhaltensverlaufsdiagnostik in inklusiven Settings vorgestellt (Kapitel 3). Hieran schließen sich die für die Arbeit relevanten Forschungsfragen zur praktischen Umsetzbarkeit der Ratinginstrumente an. Um die Implementierung eines Ratinginstruments zu untersuchen, wird im Methodenteil B ab Kapitel 4 zunächst die Stichprobe und Stichprobenschule genauer beschrieben. Anschließend erfolgt die Darstellung der Entwicklung des Ratingmaterials, insbesondere des Ratinginstruments „PUTSIE“, sowie die Beschreibung der Durchführung und Auswertung zur Beantwortung der Forschungsfragen (Kapitel 5). Im Ergebnisteil C erfolgt die Ergebnisdarstellung und -diskussion (Kapitel 6). In Kapitel 7 werden die Stärken und Schwächen der Erhebung aus Sicht des Autors dargelegt. Die Arbeit endet mit einem kurzen Fazit in Kapitel 8.

Teil A: Theorie

Im folgenden Teil werden notwendige theoretische Grundlagen, die zu der Forschungsfrage führen dargelegt. Zunächst werden Grundideen und Funktionen pädagogischer Diagnostik in der Praxis dargelegt (Kapitel 2.1). In Kapitel 2.2 erfolgt ein Bezug der Annahmen auf den Bereich des Verhaltens. Anschließend wird mit dem response-to-intervention Ansatz ein Rahmenkonzept inklusiver Schulsysteme dargelegt, um darzustellen, an welche theoretischen Anforderungen ein mögliches diagnostisches Instrument bei der Implementierung anschlussfähig sein muss (Kapitel 2.3). Kapitel 3 dient der umfassenden Darlegung von Direct Behavior Ratings als für die Inklusion möglicherweise geeignetes verlaufsdiagnostisches Instrument zur Erfassung von Schülerverhalten im inklusiven Setting. Abschließend die Darlegung der Forschungslücke sowie die Darstellung der Forschungsfragen in Kapitel 3.3.

2. Begriffliche Bestimmungen und Grundlagen

Das Forschungsvorhaben, welches sich in dieser Arbeit wiederspiegelt, beschäftigt sich mit einem spezifischen Instrument der Verlaufsdiagnostik inklusiven Unterrichts. Um einen Zugang zu diesem Themenbereich zu ermöglichen ist es notwendig, Entwicklungslinien der gewählten Methode, sowie des Inhaltsbereichs, welcher mit der

Methode verknüpft ist darzulegen. Die Methode des Direct Behavior Ratings entstammt aus Annahmen der pädagogischen Verlaufsdiagnostik (Kapitel 2.1) und deren Umsetzung in sogenannten Mehrebenensystemen zur Förderung, wie sie vor Allem durch das Response-to-Intervention-Modell (RTI) repräsentiert werden (Kapitel 2.2). Aus diesem Grund gilt es die Grundlagen und theoretischen Annahmen dieser beiden Bereiche darzulegen. Inhaltlich soll das Instrument das Verhalten von „schwierigen" Schülerinnen und Schüler erfassen. Dies macht zusätzlich eine Klärung des Begriffs Verhalten und Verhaltensstörung notwendig (Kapitel 2.3).

2.1 Pädagogische Diagnostik

„Die Lehrperson muss den Lernstand ihrer Schülerinnen und Schüler regelmäßig erfassen, ihn beurteilen und die Urteile in Noten und Zeugnissen festhalten." (Meyer 2018, S. 78) Dieses Zitat aus der allgemeinen Pädagogik verweist auf die Aufgabe von Lehrkräften, den Lern- und Entwicklungsstand der Schülerschaft zu erheben, wenn auch hier nur ein kleiner Bereich diagnostischer Kompetenz abgebildet wird. Als inhaltlicher Schwerpunkt der Lehrerausbildung wird in den Standards für die Lehrerbildung (2004) daher explizit der Bereich Diagnostik, Beurteilung und Beratung genannt. Hierunter werden die Diagnose und Förderung individueller Lernprozesse, sowie die Leistungsmessung und Leistungsbeurteilung verstanden (vgl. Kultusministerkonferenz 16.12.2004 (i.D.F. vom 12.06.2014)). Dies zeigt, dass diagnostische Prozesse in das Lehrerhandeln einbezogen werden sollen. Dennoch stellt sich die Frage, wie Diagnostik im pädagogischen Bereich genauer beschrieben werden kann und wie mögliche Instrumente aussehen können. Eine anerkannte Definition pädagogischer Diagnostik stammt von Ingenkamp und Lissmann (2008). Diese definieren:

> „Pädagogische Diagnostik umfasst alle diagnostischen Tätigkeiten, durch die bei einzelnen Lernenden und den in einer Gruppe Lernenden Voraussetzungen und Bedingungen planmäßiger Lehr und Lernprozesse ermittelt, Lernprozesse analysiert und Lernergebnisse festgestellt werden, um individuelles Lernen zu optimieren. Zur pädagogischen Diagnostik gehören ferner die diagnostischen Tätigkeiten, die die Zuweisung zu Lerngruppen oder zu individuellen Förderungsprogrammen ermöglichen sowie die mehr gesellschaftlich verankerten Aufgaben der Steuerung des Bildungsnachwuchses oder der Erteilung von Qualifikationen zum Ziel haben." (Ingenkamp und Lissmann 2008 In: Jürgens und Lissmann 2015, S.13)

Durch die Definition werden Kernfunktionen pädagogischer Diagnostik deutlich: *das Erheben von Lernvoraussetzungen und Lernbedingungen; die Analyse der Lernpro-*

zesse, sowie das Feststellen von Lernergebnissen. Zudem wird eine *Zuweisungsfunktion* (zu Förderprogrammen, etc.) und eine *Qualifikationsfunktion* (Zuweisung von Bildungsabschlüssen, etc) benannt. Diese Vielzahl von Funktionen zeigt, dass Lehrerhandeln in eine Vielzahl diagnostischer Prozesse eingebunden ist.

Untersucht man diagnostische Tätigkeiten von Lehrkräften, lassen sich formelle von informellen Diagnosen unterscheiden. Formelle Diagnosen sind „zielgerichtet, theoriegeleitet und systematisch mit wissenschaftlich geprüften Methoden erstellt." (Hesse und Latzko 2017, S. 25) Meist wird durch die Lehrkräfte jedoch informell diagnostiziert, das heißt, dass eher implizite, subjektive Urteile, Einschätzungen und Erwartungen, die während des alltäglichen erzieherischen Handelns gewonnen werden, diagnostisch genutzt werden (vgl. ebd.). Viele Autoren fordern, dass beide Arten der Diagnostik in der Praxis genutzt werden, um eine angemessene diagnostische Kompetenz, „die Fähigkeit eines Urteilers, Personen zutreffend zu beurteilen" (Schrader 2010, S.102) zu erlangen.

Die Anwendungsfelder für pädagogische Diagnostik sind dabei (wie oben angedeutet) äußerst vielseitig. So können Bereiche wie Schulleistung und deren Einflussfaktoren (zum Beispiel Vorwissen, Intelligenz, lernrelevante Emotionen), die Unterrichtsqualität, Lernergebnisse und Lernverläufe, Lernschwierigkeiten aber auch das Sozialverhalten von Schülerinnen und Schülern diagnostisch erhoben werden (vgl. Hesse und Latzko 2017).

In Anlehnung an Schuck ist es wichtig, dass pädagogische Diagnostik in Prozesse eingebunden ist, in denen sie der Hypothesenbildung und -prüfung in aufeinander bezogenen Prozessen von Diagnose und Förderung dient (vgl. Schuck 2014). Er postuliert daher, dass diagnostische Instrumente nur nutzbar sind, wenn sie der Hypothesenprüfung der pädagogischen Praxis dienen und postuliert, dass diese möglichst auch über die Zeit anwendbar bleiben müssen (vgl. ebd.). Ein mögliches Instrument hierfür stellt die sogenannte Lernverlaufsdiagnostik dar, welche im Folgenden erläutert werden soll.

2.1.1 Lernverlaufsdiagnostik

Eine spezielle Methode innerhalb der pädagogischen Diagnostik ist die Lernverlaufsdiagnostik: „Sie soll Fähigkeiten und Fertigkeiten von Schülern in Lernverläufen, also über einen längeren Zeitraum hinweg, erfassen und darstellen und somit den individuellen Lernfortschritt der Schüler an Lehrkräfte rückmelden." (Mühling et al. 2017, S.

557) Ein solches Vorgehen ermöglicht drei Funktionen, die gerade auch für inklusive Schulsysteme relevant sind (siehe Kapitel 2.3): Kinder mit Unterstützungsbedarf können frühzeitig erkannt werden (1). Die Qualität der Förderung wird durch regelmäßige Lern- und Entwicklungsverlaufsdiagnostik evaluiert (2) und unwirksame Förderungen durch kontinuierliche Diagnostik erkannt und optimiert (3) (vgl. Hennemann et al. 2015).

Ansätze zur Lernverlaufsdiagnostik wurden in den 70er Jahren von dem Wissenschaftler Stanley Deno mit dem Ansatz des „curriculumbasierten Messens“ (im Folgenden CBM) entwickelt. Er entwickelte Instrumente, um das Wissen und die Kompetenzen, bezogen auf den aktuell in der Klasse behandelten Lehr-Lernstoff, in regelmäßigen und möglichst kurzen Abständen zu überprüfen (vgl. Hesse und Latzko 2017). In Deutschland findet der Ansatz im pädagogischen Setting erst seit 2006 durch einen Artikel von Klauer Beachtung (vgl. Klauer 2011).

Instrumente zur Verlaufsmessung sind so aufgebaut, dass Schülerinnen und Schüler in relativ kurzen Abständen mit gleichen oder parallelen Aufgaben getestet werden. Dies ermöglicht, Entwicklungsverläufe über die Zeit abzubilden (vgl. Casale et al. 2015b). Hierbei lassen sich zwei Varianten unterscheiden: Zum einen kann eine gleiche Fertigkeit immer wieder erfasst werden, um fortlaufend die Verbesserungen in Form abnehmender Fehler oder zunehmender Lösungsgeschwindigkeit zu dokumentieren. Zum anderen kann die Erweiterung des Wissens überprüft werden, indem von Anfang an geprüft wird, was am Ende erreicht werden soll (vgl. Klauer 2011). Mittlerweile haben sich aus den Ansätzen einige Verfahren entwickelt, wobei diese sich meist auf den Primarbereich und dort auf die Bereiche des Lesens, der Rechtschreibung und Mathematik beziehen. Tests für die Sachfächer fehlen derzeit (vgl. Hesse und Latzko 2017). Beispiele sind der LVD-M 2-4 von Strathmann und Klauer (2012) für den Bereich Mathematik, der LdL (Lernfortschrittsdiagnostik Lesen) von Walter (2009) für die Lesekompetenz, sowie der VSL (Verlaufsdiagnostikum sinnentnehmenden Lesens) von Walter (2013), sowie die verschiedenen Tests der Onlineplattform LEVUMI.

Instrumente, die für die Lernverlaufsdiagnostik geeignet sein sollen, müssen Anforderungen genügen, um Lernverläufe regelmäßig, kleinschrittig und ökonomisch erfassen zu können (vgl. Hesse und Latzko 2017). In Anlehnung an Walter (2008) lassen sich notwendige Merkmale für Messinstrumente wie folgt zusammenfassen: Zunächst

muss ein enger Bezug zu Fach und Klasse ausgewiesen, Einsatz und Auswertung der Tests müssen schnell und unkompliziert, die Tests häufig anwendbar sein und genügend Parallelformen mit gleichem Schwierigkeitsgrad besitzen. Das Gütekriterium der Änderungssensitivität muss erfüllt werden. Abschließend sollen kurzfristige Veränderungen der Leistungspotenziale sensibel erfasst werden können (vgl. Hesse und Latzko 2017, in Anlehnung an Walter (2008)). Klauer formuliert folgende Ansprüche „Die Lernverlaufsdiagnostik erfordert (1) eine klare Definition der wiederholt zu erfassenden Kompetenz, (2) homogen schwierige, (3) änderungssensible Tests und (4) unter Umständen eine Abkehr von der klassischen Testtheorie“ (Klauer 2014, S. 1).

Mit diesen Merkmalen und Zielen kommt die Methode Ansprüchen inklusiver, individualisierter und auf Heterogenität ausgelegter Schulkonzeptionen entgegen. So wird in der KMK-Empfehlung zu inklusiver Bildung von Kindern und Jugendlichen mit Behinderungen in Schulen (2011) formuliert:

> „Individuelle Lernplanungen und Förderpläne sind für eine erfolgreiche inklusive Bildung unverzichtbar. Eine inklusive Unterrichtsgestaltung beruht auf *einer den Lernprozess begleitenden pädagogischen Diagnostik* und einer kontinuierlichen Dokumentation der Lernentwicklung.“ (Kultusministerkonferenz 20.10.2011, S. 10)

Einige Autoren betonen daher die Verlaufsdiagnostik als Gelingensfaktor von inklusiven Schulsystemen um einen Umgang mit heterogenen Lerngruppen zu ermöglichen (vgl. Casale et al. 2015a). Mit der Entwicklung zu einem inklusiven Schulsystem nimmt daher auch in Deutschland die Bedeutung der Lernverlaufsdiagnostik zu, wobei zu betonen ist, dass bis heute Instrumente für den deutschsprachigen Raum nur für wenige Bereiche vorliegen, sodass Ausbau und Forschung im Bereich der Lern- und Entwicklungsverlaufsdiagnostik wichtig sind (vgl. Hennemann et al. 2015). Ein Projekt, welches die Erforschung und Entwicklung der Verlaufsdiagnostik eng mit der pädagogischen Praxis verknüpft, ist die Onlineplattform LEVUMI, eine Kooperation der Universität München/Dortmund, Kiel und Flensburg. Mit der Lernplattform werden dabei drei Ziele verfolgt: Zunächst soll ein frei verfügbares Onlineinstrument zur Lernverlaufsmessung zur Verfügung gestellt werden. Zweitens soll durch die Datenanalyse der Ergebnisse die Forschung zu Lernverlaufsmessung und ihrer Akzeptanz weiterentwickelt werden und drittens diagnostische Maßnahmen verbessert und gezielte Förderungen entwickelt werden (vgl. Mühling et al. 2017; Gebhardt et al. 2015). Da die Arbeit im Rahmen dieses Projektes verfasst wird, wird LEVUMI im Methodenteil (Kapitel 5) genauer vorgestellt.

2.1.2 Verhaltensverlaufsdiagnostik

Casale et al. (2015b) betonen, dass die Entwicklung verlaufsdiagnostischer Tests nicht nur für den Bereich akademischer Kompetenzen notwendig ist, welche mit dem Begriff Lernverlaufsdiagnostik bezeichnet werden, sondern auch der Förderschwerpunkt emotionale und soziale Entwicklung. Auch dieser könne von solchen Testinstrumenten profitieren, sodass diese den Begriff der Verhaltensverlaufsdiagnostik vorschlagen (vgl. ebd.). Diese verfolge analog zur Lernverlaufsdiagnostik drei Ziele. *Erstens*, die konsequente Erfassung der Lern- und Entwicklungsausgangslage, um individuell passende Förderangebote abzuleiten. *Zweitens*, die frühe Identifikation herausfordernden Verhaltens, um der Entstehung von Verhaltensstörungen präventiv entgegenzuwirken und *drittens,* die Evaluation des Erfolgs pädagogischer Handlungsmöglichkeiten im Bereich der Verhaltensförderung, welche dem Anspruch evidenzbasierter Förderung gerecht wird (vgl. Casale et al. 2015a). Ziel zwei bezieht sich explizit auf Verhaltensstörungen, eine Begrifflichkeit, die im folgenden Kapitel genauer dargelegt wird. Hinter dem dritten Ziel steckt der Anspruch, forschungsmethodisch hochwertige Wirksamkeitsnachweise von Handlungsmöglichkeiten in der Sonderpädagogik zu erhalten (vgl. ebd.).

Die Autoren erarbeiten Testgütekriterien einer Verhaltensverlaufsdiagnostik. Die einzelnen Gütekriterien sollen im Folgenden kurz dargelegt werden, um die Ansprüche denen sich geeignete Instrumente auch im Vergleich zur Lernverlaufsdiagnostik stellen müssen aufzuzeigen.

Objektivität meint die Unabhängigkeit von Testergebnissen vom Testleiter und von Messzeitpunkten. Die Testdurchführung und -auswertung darf zwischen Messzeitpunkten und Testleitern nicht variieren. Da Verhalten jedoch etwas subjektiv wahrgenommenes darstellt, stellt Objektivität eine schwierige jedoch notwendige Herausforderung dar, da sonst nicht unterschieden werden könnte, ob eine Verhaltensveränderung tatsächlich auf die Fördermethode oder auf andere Aspekte zurückgeführt werden kann (vgl. Casale et al. 2015b).

Reliabilität zeigt die Messgenauigkeit an, mit der ein Merkmal erfasst wird. Diese kann über verschiedene Wege beschrieben werden. Die *interne Konsistenz* berechnet sich durch die Korrelation von Items, sodass Zusammenhänge der verschiedenen Items berücksichtigt werden. Die Retest-Reliabilität gibt die Stabilität eines Merkmals über

die Zeit an. Paralleltestreliabilität zeigt die Messergebnisse zwischen zwei gleich konstruierten Tests mit anderen Items. Veränderungsmessungen sind für die Verhaltensverlaufsdiagnostik unabdingbar, da sonst Veränderungen im Verhaltensverlauf auch zufällig zustande kommen könnten. Beispielsweise durch Messfehler. Zu Gunsten der Änderungssensitivität darf diese jedoch bei den Tests nicht zu hoch ausfallen (vgl. Casale et al. 2015b).

Validität beschreibt die Gültigkeit dessen, was ein Test misst. Die Inhaltsvalidität wird zum Beispiel durch theoretische Ableitungen und Expertenbefragungen generiert. Die Konstruktvalidität gibt an, wie präzise das beabsichtigte Merkmal erfasst wird. Diese kann zum einen erlangt werden, wenn zwei Tests, die das gleiche Konstrukt messen, hoch miteinander korrelieren oder wenn zwei Tests, die verschiedenes messen gering miteinander korrelieren. Wie Schülerverhalten gemessen werden kann, wird noch diskutiert. Wichtig ist jedoch für die Verlaufsdiagnostik die Auswahl und Begründung der Messkonstrukte. Als problematisch wird hier angesehen, dass häufig klinisch relevante Verhaltensweisen, welche im Schulalltag nicht immer relevant und beobachtbar sind, erfasst werden. Zudem ist wichtig, dass für die Schule relevante Verhaltensweisen erhoben werden, was mit dem Begriff sozialer Validität erfasst werden kann (vgl. ebd.).

Skalierung zeigt die Verrechnungsvorschrift, mittels der ein Testwert gebildet werden soll. Die Skalierung gibt an, ob die Bildung des Testwertes die Unterschiede zwischen Testpersonen widerspiegelt. Sie ist wichtig, da Über- oder Unterschätzungen von Verhaltensweisen begrenzt werden können, so muss auch die Gewichtung der einzelnen Items zueinander sinnvoll sein (vgl. ebd.).

Die *Ökonomie* beschreibt insbesondere die Anwendbarkeit und Auswertung, Testzeit und finanzielle Kosten, welche in einem angemessenen Kosten-Nutzen-Verhältnis stehen sollen. Sie ist für die Verlaufsdiagnostik besonders wichtig, da ohne ökonomisches Vorgehen keine engmaschige, wiederholte Diagnostik im schulischen Alltag möglich ist. Die Verfahren dürfen also nur wenig personelle Ressourcen erfordern und müssen eng gekoppelt an die Fördermaßnahmen funktionieren (vgl. ebd.).

Anwendung eines gültigen Messmodells. Messmodelle setzen latente (nicht-beobachtbare) Variablen mit manifesten (beobachtbaren) Variablen in Beziehung. In reflexiven Messmodellen versursacht die latente Variable die Merkmalsausprägung der manifes-

ten Indikatoren. Bei formativen Messinstrumenten ist dies umgekehrt. In der klassischen Testtheorie wird die latente Variablenausprägung aus den manifesten Itemwerten und einem Messfehler hergeleitet. In der probabilistischen Messtheorie (auch Item-Response-Theory genannt) wird die Personenfähigkeit und die Itemschwierigkeit in Beziehung gesetzt, um die latente Variablenausprägung zu errechnen. Problematisch zeigt sich die Itemschwierigkeit im Bereich Verhalten. Es muss noch untersucht werden, ob die Verhaltensweisen oder die Situationen, in denen das Verhalten gezeigt wird, die „Items" darstellen. Zudem wird Verhalten immer subjektiv wahrgenommen und kann nicht -wie eine Rechenaufgabe- mit richtig oder falsch bezeichnet werden. Aus diesem Grunde bietet sich die Generalisierbarkeitstheorie als mögliches Instrument bei der Untersuchung von Items an. Durch sie können multiple Fehlerquellen (wie verschiedene Rater, Kinder oder Situationen), die einen Einfluss auf die Situation haben, neben der allgemeinen Testgüte herangezogen werden. Auf diese Weise kann untersucht werden, inwiefern bestimmte Merkmale (sogenannte Facetten) Einfluss auf den Messfehler haben (vgl. Casale et al. 2015b)

Eindimensionalität. Kann das Antwortverhalten auf eine Kompetenz, also die latente Variable zurückgeführt werden, kann man dies Eindimensionalität nennen. Diese ist Grundlage für Tests der Verlaufsdiagnostik. Die durch die Verlaufsdiagnostik erfassten Merkmale dürfen nur auf ein theoretisches Konstrukt zurückgeführt werden. Liegen Items für die Messung eines Konstruktes eigentlich auf mehreren Konstruktdimensionen, kann dies dazu führen, dass Items entgegengesetzt bewertet werden und sich so bei der Auswertung keine Veränderung des Verhaltens zeigt, obwohl eigentlich eine Verbesserung zu erkennen sein müsste (vgl. ebd.).

Änderungssensitivität. Dieses Kriterium, welches auch schon für die Verlaufsdiagnostik allgemein genannt wurde, meint, dass Veränderungen im latenten Merkmal auch über kurze Zeiträume abgebildet werden können. Dies stellt den Anspruch der Verlaufsdiagnostik gegenüber der Statusdiagnostik dar, welche zeit- und situationsübergreifende Diagnostik ermöglichen soll. Wichtig ist es daher, Items so zu konstruieren, dass diese auch über kurze Zeiträume wahrnehmbare Veränderungen aufweisen, sodass eine mittlere Güte der Re-Test-Reliabilität erreicht wird. Unterschiede im Verhalten müssen auch kurzfristig über den Test erfasst werden (vgl. ebd.).

Inferenz. Inferenz beschreibt den Aufwand und die Komplexität der schlussfolgernden Kognitionen zum bewerten eines Items. In Bezug auf die Verhaltensverlaufsdiagnostik

muss hier die Globalität in der Formulierung von Items berücksichtigt werden. Global, also sehr allgemein formulierte Items führen zu einer höheren Inferenz, da die Beantwortung der Items mehr eigenen Definitionsaufwand benötigt. Dies führt zu einer Steigerung der Ökonomie, was den Test jedoch möglicherweise unbrauchbar für die Verlaufsdiagnostik macht (vgl. Casale et al. 2015b).

Direktheit. Verhalten lässt sich umso leichter erfassen, je näher der Zeitpunkt des aufgetretenen Verhaltens liegt. Für die Verhaltensdiagnostik sollten natürlich möglichst direkte Verfahren genutzt werden, um das Verhalten zeitnah zu registrieren, Wahrnehmungsfehler zu vermeiden und schnelle Förderentscheidungen zu treffen (vgl. ebd.).

Orientierung an der individuellen Bezugsnorm. Verlaufsdiagnostik benötigt eine Vorgehensweise im Sinne der formativen Diagnostik. Die Schülerinnen und Schüler sollen mit sich selbst verglichen werden (individuelle Bezugsnorm). Dies steht Ansätzen einer summativen statusdiagnostischen Vorgehensweise gegenüber, bei der Individuen miteinander verglichen werden (soziale Bezugsnorm). Die Orientierung an der individuellen Bezugsnorm führt ebenfalls dazu, dass eine enge Umschreibung von Verhaltensweisen mittels Items erfolgt. Dies lässt sich natürlich kritisieren, da Verlaufsdiagnostik dann womöglich nicht „allumfassend" diagnostiziert. Andererseits muss jedoch auch die Ökonomie der Methode berücksichtigt werden (vgl. ebd.).

An dieser Stelle muss betont werden, dass die Autoren darauf verweisen, dass eine Aufzählung weiterer notwendiger Gütekriterien möglich ist, sie sich jedoch auf die Betrachtung der wichtigsten Merkmale beschränken. Zudem verweisen sie darauf, dass der Ansatz des Direct Behavior Ratings, welches in Kapitel 3 genauer vorgestellt wird, geeignet scheint, um die Kriterien einer Verhaltensverlaufsdiagnostik zu erfüllen (vgl. ebd.).

In diesem Kapitel konnten die Grundlagen pädagogischer Diagnostik, ihre Funktionen und Anwendungsbereiche vorgestellt werden. Zudem wurde genauer auf die Lern- und Entwicklungsverlaufsdiagnostik eingegangen, da in der Arbeit ein Instrument für die Verhaltensverlaufsdiagnostik entwickelt und verwendet werden soll, die Vorteile eines solchen Instruments für ein inklusives Schulsystem wurden dargestellt und mit LEVUMI das Projekt innerhalb dessen ein solches Instrument genutzt werden soll in Kürze skizziert. Abschließend wurden einige Gütekriterien für die mögliche Verlaufsmessung des Verhaltens dargelegt. Im Folgekapitel soll daher eine genauere Erläuterung des Begriffs Verhalten und Verhaltensstörung erfolgen. Zudem wird die bisherige

Erfassung des Verhaltens dargelegt, um mögliche Probleme und Anknüpfungspunkte für die Verhaltensverlaufsmessung aufzuzeigen.

2.2 Verhalten

Begriffe wie Verhaltensstörung oder Erziehung sind auf das Verhalten von Menschen bezogen. Auffällig ist, dass Lehrwerke, welche sich genauer mit dem Phänomen Verhaltensstörung auseinandersetzen häufig auf eine Definition des Begriffs Verhalten verzichten (siehe zum Beispiel Hillenbrand (2008) und Stein (2017)). Mögliche psychologische Definitionen des Begriffs lauten jedoch „Verhalten ist jenes Geschehen, das, an einem Organismus oder von einem Organismus ausgehend, außenseitig wahrnehmbar ist." (Faßnacht 2000) Eine andere Definition versteht unter Verhalten „die Gesamtheit aller beobachtbaren, feststellbaren oder meßbaren Aktivitäten des lebenden Organismus" (Fröhlich 2015, S. 417). Beiden Definitionen ist gemein, dass Verhalten als von lebenden Organismen ausgehend und wahrnehmbar beschrieben wird. Die Definition von Fröhlich beinhaltet sogar den Begriff Messbarkeit, sodass auf eine diagnostische Eignung geschlossen werden kann. Beide Definitionen beschreiben Verhalten zudem als Geschehen, oder als Gesamtheit, worunter eine Vielzahl von Prozessen verstanden werden kann. Ein Diagnostikinstrument muss daher eingrenzen, welches Verhalten erfasst werden soll, sodass Eindimensionalität (Kapitel 2.1.2) erlangt werden kann. Bei abweichendem Verhalten von Kindern im schulischen Kontext wird zumeist von Schülerinnen und Schülern mit Erziehungsschwierigkeiten, Verhaltensstörungen oder Verhaltensauffälligkeiten gesprochen (vgl. Hillenbrand 2008). Durch das Adjektiv abweichend wird klar, dass ein Bezugssystem notwendig ist, in dem das Verhalten gedeutet wird. Nur in Abhängigkeit explizit oder implizit vorhandener Normen können Abweichungen subjektiv wahrgenommen werden (vgl. Hillenbrand 2008). Diese Normen können sich zudem situativ verändern. So wird in der Schule die Einhaltung anderer Normen verlangt, als beim Spielen zu Hause. Dies führt dazu, dass bestimmte Verhaltensweisen erst in der Schule als abweichend wahrgenommen werden. Zudem ist die Wahrnehmung von der Abweichung als solche subjektiv verschieden, da Lehrkräfte als für Wahrung und Setzung der Norm zuständige Verantwortliche diese zum einen unterschiedlich setzen können (zum Beispiel durch Klassenregeln), zudem aber auch unterschiedlich wahrnehmen (siehe Objektivität in Kapitel 2.1.2).

Die Abhängigkeit der Wahrnehmung des abweichenden Verhaltens von dem das Verhalten Bewertenden muss auch bei der Entwicklung von Diagnostikinstrumenten berücksichtigt werden (siehe hierzu das Merkmal der Reliabilität in Kapitel 2.1.2).

Nach Hillenbrand (2008) liefert Myschkers Definition von Verhaltensstörung eine die genannten Einflüsse berücksichtigende Definition zur Beschreibung von abweichendem Verhalten. Diese hat sich zudem in der deutschen Pädagogik durchgesetzt und lautet:

> „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 pädagogisch-therapeutische Hilfe nicht oder nur unzureichend überwunden werden kann.“ (Myschker 2002, S. 44)

Die Definition enthält 4 Elemente. Die Abweichung von zeit- und kulturspezifischen Erwartungen, sowie deren Begründung im Verhalten (1). Die Darlegung möglicher Ursachen (organogen/milieureaktiv) von Verhaltensstörungen (2). Zudem werden mögliche diagnostische Merkmale zur Feststellung dieser genannt (Mehrdimensionalität, Häufigkeit, Schweregrad) und Folgen der Störung (Einschränkung der Entwicklungs-, Lern- und Arbeitsfähigkeit sowie im Interaktionsgeschehen) dargelegt (3). Abschließend wird auf die Notwendigkeit pädagogisch-therapeutischer Hilfe verwiesen (4).

Dennoch weist auch der Begriff Verhaltensstörung das nicht lösbare Dilemma auf, einer Person (Lehrkraft) eine Zuweisungsmacht einzugestehen, das Verhalten aufgrund eigener Normen und Werte als abweichend zu bezeichnen. Da Lehrkräfte jedoch auch einem Erziehungsauftrag folgen ist ein solches Vorgehen zwar insbesondere bei der Diagnostik im Bereich des Verhaltens zu berücksichtigen, jedoch vertretbar.

Abweichendes Verhalten kann sich auf verschiedenste Art und Weise äußern, dies zeigt schon ein Vergleich der Symptome von Aggressivität und Ängstlichkeit, welche beide eine Abweichung von der Norm darstellen können. Eine empirische Klassifikation, welche durch die Zusammenfassung häufig zusammen auftretender Phänomene in Klassen gekennzeichnet ist, wurde von Myschker entwickelt. Dieser unterscheidet externalisierende Störungen von internalisierenden Störungen, sozial unreifem Verhalten und sozialisiert delinquentem Verhalten (vgl. Hillenbrand 2008). Insbesondere

die externalisierenden und internalisierenden Störungen sind dabei mittlerweile empirisch fundiert und werden in der pädagogischen Diagnostik genutzt (vgl. Myschker 2002). Internalisierende Störungen betreffen eher Mädchen und lassen sich durch ängstlich-gehemmtes Verhalten beschreiben. Typische Störungsbilder sind Angst, Minderwertigkeit, Trauer, Interessenlosigkeit, Schlafstörung und somatische Störung (vgl. Hillenbrand 2008). Jungen sind weitaus häufiger von externalisierenden Störungen betroffen. Diese lassen sich auch als ausagierende Störungen beschreiben. Typische Störungsbilder sind Aggression, Hyperaktivität, Aufmerksamkeitsstörung, Impulsivität (vgl. ebd.). Durch die empirische Fundierung der Klassifikation bietet sich diese auch für die Berücksichtigung im diagnostischen Bereich an.

Diagnostik lässt sich in Bezug auf Verhaltensstörungen als „das Gewinnen von Hinweisen und Erkenntnissen sowie eine darauf aufbauende Urteilsbildung zum Zwecke der Beschreibung und Erklärung der Störung“ (Stein 2017, S. 120) beschreiben. Der Autor unterscheidet zwei Anlässe der Diagnostik, zum einen die formelle Diagnostik zur Erstellung von Gutachten, zum anderen die informelle Diagnostik zur Einschätzung und Beurteilung von Problematiken im Kontext von Verhaltensstörungen in Alltagssituationen. Er verweist zudem auf den Unterschied von Statusdiagnostik, bei der eine bestimmte Situation oder ein bestimmter Zustand einmalig untersucht wird, sowie Prozessdiagnostik, bei der Veränderungen über die Zeit erhoben werden können (vgl. ebd.). Im Sinne der Ausführungen zur Verlaufsdiagnostik (Kapitel 2.1) sollen an dieser Stelle klar prozessdiagnostische Methoden fokussiert werden. Zudem sollen die diagnostischen Methoden nicht einer Erstellung von Gutachten dienen, sondern der Auswahl von Fördermethoden und der Überprüfung ihrer Wirksamkeit im Sinne einer evidenzbasierten Förderung (siehe Kapitel 2.2). Auf diese Weise wird eine informelle Zielsetzung verfolgt.

Für den Bereich der sozialen Lernziele in den Lehrplänen einiger Bundesländer gibt es insgesamt nur wenige Diagnoseinstrumente (vgl. Hesse und Latzko 2017). Das gleiche Ergebnis gilt auch für die Verhaltensverlaufsdiagnostik. Casale et al. (2015a) nennen hierfür zwei Gründe: Erstens werden an die Testgüte von Instrumenten zur Verlaufsdiagnostik hohe Anforderungen gestellt, zweitens stellt sich die Frage, wie Verhalten -als wesentlicher Indikator der emotionalen und sozialen Entwicklung von Kindern und Jugendlichen- entsprechend der Gütekriterien erfasst werden kann. Für die

pädagogische Arbeit im Förderschwerpunkt emotionale und soziale Entwicklung werden von der Forschergruppe folgende diagnostische Methoden zusammengefasst: Gesprächsmethoden und Interviewverfahren, Beobachtungsverfahren und Verhaltensbeobachtungen, Beurteilungsverfahren und Verhaltensbeurteilungen, Testverfahren, Dokumentenanalyse und Alltagsbeobachtungen (vgl. Casale et al. 2015b).

Die Autoren stellen heraus, dass insbesondere die systematische Verhaltensbeobachtung und die Verhaltensbeurteilung mittels Ratingskalen Gütekriterien einer Verhaltensverlaufsdiagnostik entsprächen (vgl. Casale et al. 2017). Sie beschreiben die systematische Verhaltensbeobachtung wie folgt.

> „Die systematische Verhaltensbeobachtung liefert Daten über die Häufigkeit bzw. Dauer des Auftretens von konkreten Verhaltensweisen. Der Beobachter muss vorher genau definieren, worauf er beim Kind achten will und wie er dies protokolliert. (…) Systematische Verhaltensbeobachtungen (…) weisen eine hervorragende Objektivität und Reliabilität auf und bilden Verhaltensänderungen sensitiv ab. (…) Allerdings ist ein solches Vorgehen mit großem Aufwand und zusätzlicher personeller Ressource durchführbar und daher für den engmaschigen Einsatz im schulischen Alltag (…) nicht geeignet.“ (Casale et al. 2015a, S. 328–329)

Es zeigt sich also ein Problem in der Ökonomie des Ansatzes für die verlaufsdiagnostische Verwendung, da eine Durchführung mit einer Lehrkraft im schulischen Alltag nicht möglich ist.

Die zweite Methode Verhaltensbeurteilung mittels Ratingskalen zeigt folgende Eigenschaften:

> „Die zweite diagnostische Methode, die Verhaltensbeurteilung mittels Ratingskalen, ist hingegen sehr ökonomisch, da die Beantwortung der Items meist in wenigen Minuten erfolgt. Verhaltensbeurteilungen werden in der Form angewendet, dass ein beobachtetes Verhalten retrospektiv anhand standardisierter Skalen eingeschätzt wird. (…) Gegen den häufigen Einsatz als Instrument zur Verlaufsdiagnostik spricht jedoch, dass die Einschätzungen stark subjektiv erfolgen, eine recht lange Latenz zum auftretenden Verhalten aufweisen und nicht zwingend veränderungssensitiv über diese kurzen Zeitabstände sind.“ (Casale et al. 2015a, S. 329)

Der Vorteil der Methode liegt in dessen Ökonomie, dennoch sind die starke Subjektivität (vgl. Objektivität in Kapitel 2.1.2) und die langen Zeiträume zwischen Auftreten und Bewertung des Verhaltens (vgl. Direktheit in Kapitel 2.1.2) Gründe, die gegen die Eignung als Methode der Verlaufsdiagnostik sprächen.

Es lässt sich also festhalten, dass keine dieser „klassischen“ diagnostischen Instrumente als Instrument der Verlaufsdiagnostik geeignet scheint. Die Autoren verweisen

jedoch auf die Möglichkeit eines Mischinstruments der beiden Methoden der „direkten Verhaltensbeobachtung“ und der „Verhaltensbeurteilung mittels Ratingskalen“. Diese werde durch die „Direkte Verhaltensbeurteilung“ oder „Direct Behavior Ratings“ (DBR) präsentiert (Kapitel 3), welche als Instrument der Verhaltensverlaufsdiagnostik in Frage käme (vgl. Casale et al. 2015b). Bevor das Instrument jedoch vorgestellt wird, müssen die Rahmenbedingungen inklusiver Beschulung, in die es eingebettet werden soll, dargelegt werden.

2.3 Inklusive Schulsysteme und Response-to-Intervention

Soll die Implementierung eines Direct Behavior Ratings an einer Schule untersucht werden, muss zunächst dargelegt werden, für welche Schulsysteme die Verhaltensverlaufsdiagnostik überhaupt geeignet sein soll. In der Literatur zur Verhaltensverlaufsdiagnostik wird dabei immer wieder von „Response to intervention“ gesprochen (vgl. Gebhardt et al. 2016). Der Ansatz wird daher im Folgenden dargelegt und Forschungsergebnisse im Bereich des Verhaltens vorgestellt. Das Modell, welches seit den 1960er Jahren in Amerika in enger Verknüpfung mit der Verlaufsdiagnostik entwickelt wird (vgl. Blumenthal et al. 2014) erhebt den Anspruch, ein Rahmenkonzept für evidenzbasierte Förderung in inklusiven Bildungssystemen darzustellen (vgl. Hillenbrand 2015). Evidenzbasierung drückt „die Anforderung an Vorgehensweisen, Methoden, Verfahren und Programme aus, bestimmten wissenschaftlichen Überprüfungen Stand zu halten und dabei zu positiven Wirkungen zu führen.“ (Hillenbrand 2015, S. 170–171) Das Konzept besteht aus drei Ebenen (Abbildung 1), in der englischen Literatur „tears“ genannt.

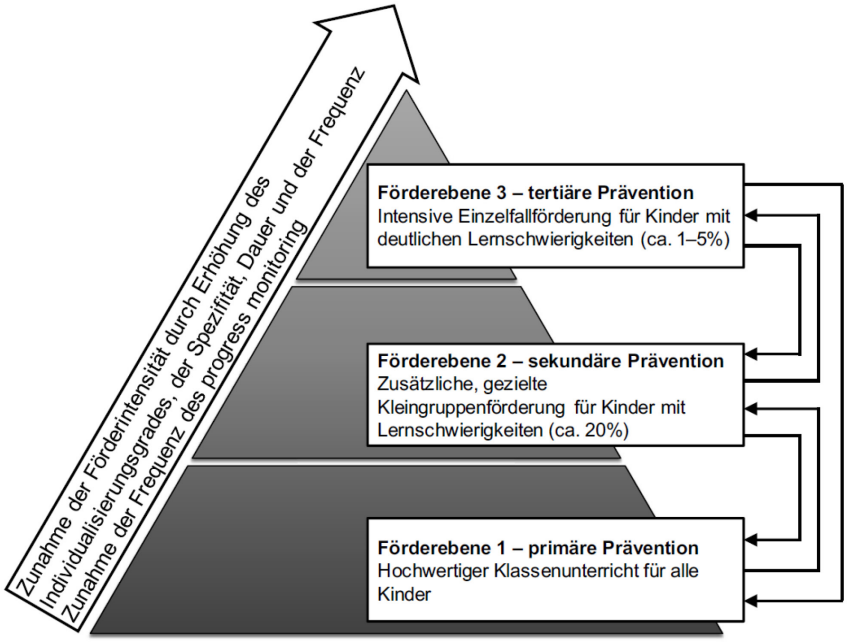


Abbildung 1: Pyramide der Förderebenen des RTI (Voß et al. 2015, S. 19)

Auf Förderebene 1 befindet sich der reguläre, hochwertige Unterricht mit evidenzbasierten Verfahren. Dieser soll durch ein Lernprozess-Monitoring begleitet werden, mit dessen Hilfe frühzeitige Fehlentwicklungen erkannt werden können.

Kinder, die auf dieser Ebene keine oder nur geringe Lernfortschritte verzeichnen und daher im Monitoring auffallen, haben Anspruch auf Ebene 2, die intensivierte Förderung. Die Merkmale auf dieser Ebene sind stark durch die amerikanische Diskussion um den No Child Left Behind Act und den Individuals with Disabilities Education Act mit verschiedenen bildungspolitischen Grundideen geprägt, sodass hier teils widersprüchliche Forderungen (zum Beispiel das vollständige Verschmelzen allgemeiner Pädagogik mit der Sonderpädagogik gegenüber dem Erhalt der Disziplin) zu finden sind (vgl. Blumenthal et al. 2014). Eine gemeinsame Annahme ist jedoch, dass etwa 20% der Schülerschaft eine zusätzliche intensivierte Förderung, welche durch engmaschigere Verlaufsdiagnostik begleitet wird, benötigen. Hierunter werden auch die Instrumente der curriculum-based measurement verstanden, welche in Annahme der Autoren auch in geänderter Form auf den Verhaltensbereich bezogen werden können Wichtig ist, dass Kindern und Jugendlichen auf dieser Ebene noch kein sonderpädagogischer Förderbedarf zugewiesen wird (vgl. ebd.). Die zusätzliche Förderung kann innerhalb der Klasse oder durch eine Kleingruppen-/Einzelförderung gestaltet werden.

Genutzt werden sollen evidenzbasierte Fördermethoden. Zudem können auch multiprofessionelle Teams zur Besprechung der Förderung gebildet werden (vgl. Huber und Grosche 2012).

Förderebene 3 besteht aus einer intensivierten Langzeitförderung durch spezialisierte Fachkräfte für die 3-5% der Schülerschaft, bei denen durch Interventionen auf Förderebene 2 keine oder unzureichende Verbesserungen zu erkennen sind. Die ist gekennzeichnet durch die längerfristige Aufrechterhaltung der in Stufe 2 genannten diagnostischen Maßnahmen, es kann jedoch auch eine zusätzliche Differenzialdiagnostik in den für das Kind relevanten Bereichen genutzt werden. Diese Diagnostik soll Aufschluss über Möglichkeiten zu einer verbesserten, intensivierten Förderung geben. Diese soll von spezialisierten Fachkräften (z.B. Förderpädagogen) durchgeführt werden. Eine Evaluation der Methoden erfolgt wiederum über die Mittel der Verlaufsdiagnostik. Eine Zuweisung sonderpädagogischen Förderbedarfs ist hier möglich, aber nicht zwingend erforderlich (vgl. ebd.).

Abschließend lassen sich die Eigenschaften des RTI in Anlehnung an Blumenthal durch folgende Merkmale beschreiben. Zum einen gibt es nach Intensität gestufte Förderebenen zur Prävention von Lern- und Verhaltensschwierigkeiten. Zweitens werden Förderentscheidungen auf Basis individueller Ergebnisse in Screenings und Lernverlaufsdokumentationen getroffen und drittens ist ein Einsatz evidenzbasierter Lehr- und Fördermethoden sowie -programme zu erkennen (vgl. Blumenthal et al. 2014).

In Deutschland wird das Modell zumeist mit dem wait-to-fail-Problem in Verbindung gebracht. Huber und Grosche (2012) beschreiben darunter das grundsätzliche Problem, dass Sonderpädagogen und an sie gebundene Ressourcen erst abgerufen werden, wenn eine bestimmte Wahrnehmungs- oder Belastungsgrenze an einer allgemeinen Schule überschritten wurde. Probleme eines Schulkindes müssen also zunächst umfassend und massiv werden, bis sie mit den zur Verfügung stehenden Diagnoseinstrumenten erfasst und klassifiziert werden können (vgl. ebd.). Das RTI soll dem Problem als „Rahmenkonzept zur Identifikation, Prävention und Intervention bei Beeinträchtigungen im Lernen und Verhalten“ (Huber und Grosche 2012, S.313) entgegenwirken. Hillenbrand verweist auf die Notwendigkeit solcher Konzepte für eine gelingende Inklusion von Schülerinnen und Schülern mit Verhaltensstörungen (vgl. Hillenbrand 2015).

Der RTI-Ansatz ist zudem das einzige im deutschen Bereich evaluierte Konzept für inklusive Schulsysteme (vgl. Mahlau et al. 2014). Eine breit angelegte Evaluation des Ansatzes findet sich durch das Rügener Inklusionsmodell (RIM). Hier wurde das Rahmenkonzept in einer Treatmentgruppe aus 441 Kindern aus 23 Schulklassen 12 Rügener Grundschulen implementiert und Lern- und Förderergebnisse der Systeme mit einer Kontrollgruppe Stralsunder Grundschulen (N=358) verglichen. Die Forscher betonen folgende Ergebnisse: Der Ansatz erweist sich insgesamt als zielführend bei der Gestaltung eines wohnortnahen, zugänglichen, angemessenen und anpassungsfähigen inklusiven Schulsystems. Die Schülerinnen und Schüler zeigen in der Treatmentgruppe höhere Ergebnisse in den Förderschwerpunkten Lernen, emotionale und soziale Entwicklung und gleichwertige Ergebnisse im Förderschwerpunkt Sprache. RTI führt zu einer deutlichen Abnahme der Häufigkeiten von sonderpädagogischen Förderbedarfen in diesen drei Förderschwerpunkten. Die Einstellung gegenüber dem Konzept ist bei Sonderpädagogen und Schulleitern positiv, bei Grundschullehrern tendenziell positiv, Eltern erleben die Struktur mindestens genauso positiv wie bisher, fühlen sich jedoch besser über den Entwicklungsstand ihrer Kinder informiert (vgl. Voß et al. 2015). Blumenthal und Voß (2016) betrachteten die Rügener Schülergruppe (N=430) in Hinblick auf ihre sozial-emotionale Situation, indem Sie Lehrereinschätzungen von Verhaltensbereichen über den SDQ und als Einschätzungen der Schülerinnen und Schüler mit dem Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen (FEESS 3-4) und einem soziometrischen Verfahren zum Ende der vierten Klasse im Querschnitt erhoben und mit der Stralsunder Kontrollgruppe verglichen (vgl. Blumenthal und Voß 2016). Es zeigten sich folgende Ergebnisse: Das RIM weist bessere Ergebnisse in Bezug auf das Problem- aber auch prosoziale Verhalten für Kinder ohne sowie mit leichten Minderleistungen, bei prosozialem Verhalten auch mit deutlichen Minderleistungen auf. Zudem zeigen sich positive auf das Klassenklima, das Angenommensein, sowie des Selbstkonzepts und der sozialen Integration, zumindest bei leichten Minderleistungen, beim Klassenklima auch bei deutlichen Minderleistunden (vgl. ebd.).

Hillenbrand (2015) formuliert in seinem Beitrag zu evidenzbasierter Praxis im Förderschwerpunkt emotionale und soziale Entwicklung einige Merkmale für die Arbeit mit Kindern des Förderschwerpunkts in inklusiven Settings, welche stark an das Respon-

seto Intervention-Modell (im Folgenden RTI) angelehnt sind. Er schlägt drei Förderstufen vor, die zunehmende Intensität und Individualisierung von Maßnahmen aufweisen:

Stufe 1, universelle Stufe. Der reguläre Unterricht erfolgt evidenzbasiert, Screenings ermöglichen das frühzeitige Erkennen von Problementwicklungen. Wissenschaftlich fundierte Maßnahmen kommen auf universeller Ebene für die ganze Klasse zum Einsatz. Insbesondere Maßnahmen des Classroom Managements lassen sich hier ansiedeln.

Stufe 2, gezielte Unterstützung. Die Kooperation der Fachkräfte erfolgt hier im Team zur Entwicklung individualisierter Förderpläne. Die Durchführung der Interventionen erfolgt meist in der allgemeinen Klasse. Auf dieser Ebene sind beispielsweise Maßnahmen wie das Good Behavior Game zu nennen.

Stufe 3, spezialisierte Unterstützung. Intensivere Maßnahmen auf Basis eines individuellen Förderplans werden verwendet. Es finden eine engmaschige Diagnostik und indizierte, sehr individuell angepasste Maßnahmen mit hoher Intensität Verwendung. Die Förderung kann in speziellen Kleingruppen stattfinden, am besten ist jedoch die Förderung im normalen Klassensetting. Ziel ist die Rückführung in die Klasse oder die erfolgreiche Stabilisierung des Schülers (vgl. Hillenbrand 2015).

Zur Umsetzung dieser Ansprüche ist insbesondere eine regelmäßige Evaluation der Maßnahmen mit validen Instrumenten der Begleitdiagnostik ab Stufe zwei notwendig, welche ein kontinuierliches Monitoring des Verhaltens erlaubt. Hinweise auf eine mangelnde Wirksamkeit (Response) von Maßnahmen (Intervention) sollen auf diese Weise schneller wahrgenommen und Fördermaßnahmen dementsprechend schneller angepasst werden können (vgl. Casale et al. 2017).

Für die Förderung sozialer-emotionaler Kompetenzen wurden im Rahmen des Rügener Inklusionsmodells folgende diagnostische Möglichkeiten und Interventionen für die drei verschiedenen Ebenen des RTI-Ansatzes festgelegt.

Auf Förderebene 1 erfolgt ein Grobscreening. Es werden verschiedene Maßnahmen des Classroom-Managements, sowie allgemeine Programme zur Förderung emotionaler und sozialer Kompetenz als präventive Angebote verwendet (vgl. Mahlau et al. 2014).

Förderebene 2 fokussiert zusätzlich individuelle Screeningverfahren, wie den SEVE-Test (Hartke und Vrban 2010). Als Maßnahmen werden verhaltenssteuernde Maßnahmen, z.B. "49 Handlungsmöglichkeiten" (ebd.) vorgeschlagen.

Förderebene 3 bleibt dem Sonderpädagogen vorenthalten. Hier geht es darum, die Verhaltensproblematik genau zu beschreiben und so neue Ansatzpunkte für die Förderung zu erlangen. So können spezifische auf den Schüler zugeschnittene Förderangebote gemacht werden (vgl. Mahlau et al. 2014).

Interessant ist, dass aus beiden Ansätzen, obwohl sich diese auf den RTI-Ansatz beziehen, keine verlaufsdiagnostischen Instrumente ersichtlich werden. Es wird lediglich von Screeningverfahren gesprochen. Es stellt sich daher die Frage, wie für das Rahmenkonzepte geeignete verlaufsdiagnostische Instrumente im Bereich des Verhaltens aussehen können.

Durch die positiven Effekte des RTI auch auf die Förderung von Kindern mit Verhaltensauffälligkeiten, sollen die Ebenen im Methodenteil für die Beschreibung des Schulsystems dienen, an der die Erhebung durchgeführt wird. Auf diese Weise wird die bisherige Arbeitsweise mit theoretischen Annahmen verglichen.

In diesem Kapitel wurde das Rahmenmodell des Response-to-Intervention dargelegt, in welches sich ein verlaufsdiagnostisches Instrument einbetten lassen muss. Der Einbezug der Verlaufsdiagnostik, sowie die Eignung für die Förderung von Schülerinnen und Schülern mit Verhaltensschwierigkeiten wurde anhand des Rügener Inklusionsmodells verdeutlicht. Nun gilt es, ein mögliches Instrument aufzuzeigen, mit dem die Verhaltensverlaufsdiagnostik ermöglicht wird. Hierzu wird im Folgenden die Konzeption von und Arbeit mit Direct Behavior Ratings vorgestellt.

3. Direct Behavior Rating (DBR)

Das Kapitel beschreibt die Methode der Direkten Verhaltensbeurteilung (in der englischsprachigen Literatur „Direct Behavior Rating“) als mögliches Instrument der Verlaufsdiagnostik des Schülerverhaltens. Zunächst werden grundsätzliche Eigenschaften und Annahmen dargelegt, in einem zweiten Kapitel der vorhandene Forschungsstand zu der Methode vorgestellt und abschließend eine Forschungslücke erarbeitet, aus der schließlich die Forschungsfragen der Arbeit hergeleitet werden.

3.1 Theoretische Grundlagen

In Kapitel 2.2 wurde bereits dargelegt, dass die klassischen Instrumente der Statusdiagnostik des Verhaltens als solche nicht für die Verlaufsdiagnostik übernommen werden können. Zudem wurden Direct Behavior Ratings als mögliche, für die Verlaufsdiagnostik geeignete, Mischform von direkter Verhaltensbeobachtung und Verhaltensbeurteilung mittels Ratingskalen beschrieben. Um dies zu begründen soll im Folgenden die Entwicklung des Ansatzes sowie dessen theoretischen Überlegungen skizziert werden. In dem Kapitel wird insbesondere auf das Herausgeberwerk „Direct behavior rating. Linking assessment, communication, and intervention" von Briesch (2016) verwiesen, welches einen umfangreichen Einblick in Theorie und Anwendung der Ratinginstrumente liefert.

Die Ursprünge des Direct Behavior Ratings lassen sich auf den amerikanischen Schulpsychologen Calvin Edlund zurückführen. Er beschrieb eine Intervention, in welcher der Klassenlehrer zunächst akzeptables Sozialverhalten beschrieb und das Verhalten der Kinder anschließend nach jeder Instruktionsphase nach dieser Definition bewertete. Am Ende eines Schultages wurde dem Kind das Verhalten zurückgemeldet und die Ergebnisse mit nach Hause gegeben. In den folgenden Jahren entstanden aus dem Ansatz eine Vielzahl an Methoden, die die Belohnung positiven Verhaltens im Elternhaus in den Fokus rückten (vgl. Briesch 2016). In dem Vorgehen lassen sich durch die regelmäßige Rückmeldung über das Verhalten und dessen fortlaufende Bewertung Merkmale einer Verlaufsdiagnostik des Verhaltens erkennen. Daher gingen weitere diagnostische Instrumente aus der Grundüberlegung hervorgingen, welche als Direct Behavior Ratings bezeichnet wurden. Eine mögliche Beschreibung der Verfahren erfolgt durch Casale et al. (2017):

> „Im englischsprachigen Raum hat sich das sogenannte *direct behavior rating* als eine neuartige Methode zur Verlaufsdiagnostik von Schülerverhalten entwickelt (…). Es vereint die Vorteile der systematischen und direkten Verhaltensbeobachtung und der Verhaltensbeurteilung mittels Ratingskalen. In einem festgelegten kurzen Zeitraum wird ein bestimmtes konkret operationalisierbares Zielverhalten beobachtet und direkt im Anschluss an diesen Zeitraum auf einer Ratingskala eingeschätzt. Aufgrund dieser ökonomischen Vorgehensweise kann das Rating sehr häufig (…) wiederholt werden. Die Ergebnisse lassen sich über die Zeit in einem Liniendiagramm darstellen, sodass Verläufe und Entwicklungen von Schülerverhalten sichtbar werden." (Casale et al. 2017, S.259)

Aus der Definition lassen sich die Hauptmerkmale der Methode, sowie ihr Bezug auf „klassische" diagnostische Verfahren der Verhaltensdiagnostik erkennen.

Zum einen wird die *Direktheit der Beobachtung* herausgestellt. Das Rating soll direkt im Anschluss an das Beobachtungsintervall erfolgen. Dies verspricht weniger Fehler durch lange Zeiträume zwischen der Beobachtung und Bewertung des Verhaltens. Dennoch können sich hier Einschränkungen über die Wahl des Beobachtungsintervalls zeigen, liegt ein zu kurzer Zeitraum vor, lässt sich das Verhalten gegebenenfalls nicht beobachten, während ein zu langes Beobachtungsintervall gegebenenfalls zu Vergessen von aufgetretenem Verhalten führen kann. Die möglichst geringe zeitliche Distanz zeigt Verknüpfungen zur Methode der psychologisch objektivsten Beobachtungsmethode des Verhaltens, der direkten Verhaltensbeobachtung (vgl. Briesch 2016).

Zudem wird nicht das Verhalten insgesamt, sondern ein *spezifisches Verhalten* beobachtet. Insgesamt sollen durch die Ratings Verhaltensweisen beurteilt werden, welche beobachtbar und messbar sind. Allgemeine Merkmale wie Intelligenz können nicht erfasst werden. Dies zeigt den Anspruch der Methode, ein genau definiertes Verhalten zu erfassen, welches möglichst wenig intersubjektive Deutungen zulässt (vgl. ebd.).

Abschließend dient die Verwendung von Ratingskalen dazu, das wahrgenommene Verhalten einer Person *zu quantifizieren*, ob Häufigkeit, Länge, und/oder Intensität wahrgenommen wird, ist hierbei abhängig vom gewählten Verhalten (vgl. ebd.). Die Verwendung von Skalen zeigt die Ähnlichkeit zu der diagnostischen Methode der Verhaltensbeobachtung mittels Ratingskalen.

Die bisherigen Darstellungen der Methode und ihren Hauptmerkmalen, weisen auf die Funktion des Instruments, der visuellen Darstellung von Verhaltensverläufen zur diagnostischen Einschätzung von Schülerverhalten, sowie zur Überprüfung von Fördermaßnahmen über die Zeit hinweg hin. Briesch (2016) benennt daher *assessment* (Beurteilung) und *intervention* (Förderung) als zwei Hauptfunktionen der direkten Verhaltensbeurteilung. Die Methode soll also einerseits das Erfassen des Verhaltens mittels Verlaufsgraphen und eine Evaluation der verwendeten Fördermaßnahmen ermöglichen. Diese Möglichkeiten decken sich auch mit den Zielen einer Verlaufsdiagnostik, wie sie in Kapitel 2.1 und den Zielen des RTI, wie sie in Kapitel 2.2 dargestellt wurden. Andererseits kann die Methode auch als Intervention genutzt werden, indem die Schüler oder Eltern Rückmeldungen über ihr Verhalten erhalten, wie es ursprünglich von Edlund verfolgt wurde. Eine weitere Funktion wird mit *communication* (Kommunikation) (vgl. ebd.) benannt und im Folgenden dargelegt.

Briesch (2016) betont mit der Funktion *Kommunikation*, dass durch das Instrument des Direct Behavior Ratings ein Verhalten nicht nur erfasst und dargestellt, sondern kommunizierbar wird. So wird in Anlehnung an Bronfenbrenners Systemtheorie argumentiert, dass Verhalten in verschiedene Kontexte eingebettet ist und durch diese beeinflusst wird. Wichtig sei daher, dass über diese Kontexte hinaus über Verhalten und Ziele miteinander kommuniziert werden könne. Direct Behavior Ratings böten eine Möglichkeit zur einfachen, effektiven Kommunikation zwischen verschiedenen an der Förderung beteiligten Akteuren (inklusive der Schülerinnen und Schüler selbst). So wird moniert, dass Lehrkräfte häufig nur Rückmeldungen zu negativem Störverhalten an Schülerinnen und Schüler sowie Erziehungsberechtigte weitergeben. Durch die regelmäßige Rückmeldung von Verhalten können jedoch auch positive Entwicklungen einfach erkannt und kommuniziert werden. Eine zweite Argumentationslinie bezieht sich auf die Kommunikation innerhalb von Response-to-intervention Modellen. Es müssen regelmäßige Absprachen erfolgen, um Schülerinnen und Schülern geeignete Ebenen und Fördermaßnahmen zuzuweisen. Dadurch, dass die Beurteilung des Verhaltens durch die Einschätzung nur wenig Zeit in Anspruch nähme, wird davon ausgegangen, dass die Methode auch für die längerfristige Diagnostik genutzt werden kann. Ein dritter Grund wird darin genannt, dass durch die Rückmeldung über bestimmte Verhaltensweisen und Fördermethoden zu einer Generalisierung führen könnten, indem sich die Systeme in ihrer Art der Förderung, ihren Erwartungen an den Schüler angleichen (vgl. ebd.).

Merkmale, welche das Instrument insbesondere für die Eignung in inklusiven Schulsystemen nutzbar machen, sind nach Ansicht der Autoren Effizienz, Flexibilität, Wiederholbarkeit und Vertretbarkeit (vgl. ebd.).

Unter dem Merkmal *Effizienz* verstehen die Autoren die auch für Lernverlausdiagnostik notwendige Eigenschaft als Diagnostik- und Interventionsinstrument ökonomisch zu sein. Bei Bekanntheit des Ratings könne es in ein paar Sekunden ausgefüllt werden. Als Interventionskonzept (zum Beispiel durch die Verknüpfung mit daily report cards) ist es zudem ökonomisch, da es keine zusätzlichen Ressourcen von außerhalb benötige (vgl. ebd.).

Flexibilität meint, dass Direct Behavior Ratings für verschiedene Verhaltensauffälligkeiten, Rater, Situationen, Zeiträume, Ratingsysteme und Häufigkeiten von Ratings geeignet und anpassbar seien (vgl. ebd.).

Die *Wiederholbarkeit* spricht das für verlaufsdiagnostische Instrumente notwendige Kriterium an, dass das Diagnostikinstrument über die Zeit hinweg auf die gleiche Weise genutzt werden kann (vgl. Briesch 2016).

Abschließend muss das Instrument *vertretbar* sein. Dies meint, dass diagnostisch genaue, reliable und valide Daten produziert werden. Eine Einschätzung für Direct Behavior Ratings hierzu erfolgt in Kapitel 3.2 in welchem die wichtigsten Forschungsergebnisse zur Methode dargelegt werden (vgl. ebd.).

Trotz der eindeutigen Merkmale, welche das Diagnostikinstrument betreffen, zeigen sich Umsetzungen von Direct Behavior Ratings äußerst vielfältig. Grob lassen sich zwei Gruppen von Ratinginstrumenten anhand ihrer Skalen unterscheiden. So gibt es zum einen die Gruppe von DBRs mit Single-Item-Skalen, zum anderen Messinstrumente, bei denen Multi-Item-Skalen verwendet werden.

Single-Item-Skalen zeichnen sich dadurch aus, dass mit einem einzelnen Item eine übergeordnete Verhaltensdimension erfasst wird, das Verhalten wird also über die Bewertung eines Verhaltens eingeschätzt, diese Skalierung eignet sich insbesondere, wenn ein breit gefächertes Verhaltensproblem beurteilt werden soll (vgl. Casale et al. 2017). Multi-Item-Skalen, sind durch mehrere Items (in der Regel drei bis fünf) gekennzeichnet, die beantwortet werden, um eine übergeordnete Verhaltensdimension zu erfassen. Je nach Bedarf können Ergebnisse der Items individuell analysiert oder als gemeinsamer Summenscore dargestellt oder analysiert werden. Multi-Item-Skalen sind besonders sinnvoll für die Erfassung des konkreten Fördererfolgs bei Schülerinnen und Schülern. Je nach Bedarf können Ergebnisse der einzelnen Items individuell analysiert oder zu einem Summenscore aufaddiert werden (vgl. ebd.).

Trotz seiner postulierten Eignung für die Verlaufsdiagnostik wird auch Kritik an der Verwendung von Instrumenten der Direkten Verhaltensbeurteilung geübt. So äußert Willmann (2018) in Bezug auf die Ökonomie des Instruments einige Bedenken, indem er die engmaschige Diagnostik für maladaptive Verhaltensroutinen in Frage stellt, da diese häufig großer Persistenz unterlägen und sich nur langsam veränderten Das Argument ist insofern in Frage zu stellen, da Direct Behavior Ratings gut beobachtbare Verhaltensmerkmale erfassen sollen und sich hierfür keine Persönlichkeitsmerkmale eignen (siehe oben). Es kommt also immer auf die Benennung der einzelnen Items zur Erfassung des Schülerverhaltens an. Des Weiteren wird argumentiert, dass das Ver-

fahren das Oberflächenverhalten von Schülern im Unterricht messe und dabei lediglich soziale Aspekte fokussiert würden, die inneren Motivlagen, der subjektive Sinn und die individuelle Bedeutung sowie die emotionalen Hintergründe des Verhaltens schlichtweg ausgeblendet würden (vgl. Willmann 2018). Bei diesem Argument muss eventuell die Funktion des Instrumentes noch einmal deutlich gemacht werden. Vorrangig dienen Direct Behavior Ratings der Diagnostik und der Evaluation von Fördermaßnahmen, nicht aber als isolierte Fördermethode. Es konnte durch die Ausführungen zur Kommunikation mit Hilfe des Instrumentes gezeigt werden, dass die Methode durchaus für ein ganzheitliches Verständnis im Sinne des Einbezugs verschiedener Bezugssysteme sowie des Schülers selbst genutzt werden kann. Dennoch kann zu ergründen sein, ob die Lehrkräfte in der Praxis in diesem Sinne mit dem Instrument arbeiten.

An dieser Stelle konnten Merkmale, Funktionen und Eigenschaften von Direct Behavior Ratings vorgestellt werden. Insbesondere dessen Eignung als Instrument zur Verlaufsdiagnostik des Verhaltens konnte dargelegt werden. Im Folgenden gilt es zu untersuchen, inwieweit und unter welchen Umständen sich das Instrument in empirischer Forschung als effektiv erwiesen hat, um dessen Vertretbarkeit aus psychometrischer Perspektive aufzuzeigen.

3.2 DBR in empirischer Forschung

Direct Behavior Ratings erheben den Anspruch als Instrument der Lernverlaufsdiagnostik in Frage zu kommen. Aus diesem Grund fassen Huber und Rietz (2015) bisherige Methodenstudien (Stand 2015) in einem systematischen Review zusammen. Insgesamt fließen 17 Studien in das Review mit ein. 16 Studien stammen allein aus der Forschergruppe um Chafouleas, Christ, Riley-Tillman und Briesch.

Die Autoren weisen zunächst darauf hin, dass die Untersuchung der Instrumente nach klassischen Testgütekriterien durch unterschiedliche Beobachtungsgegenstände, -situationen und Beobachter gegebenenfalls nicht auf den Bereich von Verhaltensbeobachtungen übertragbar sind. Hier ist noch ein Konflikt verschiedener Autoren erkennbar. Ein Wert, welcher bei der Erforschung des Instruments eine besondere Rolle spielt, ist jedoch die Übereinstimmung bei der Bewertung des Verhaltens zwischen verschiedenen Beobachtern (Intrarater und Interrater-Reliabilität). Es stellt sich jedoch die Frage, ob im Sinne eines vorrangig informellen Verfahrens verschiedene Personen Verhalten

gleich bewerten müssen, wenn es vielmehr um die Erfassung von Entwicklungsverläufen geht (vgl. Huber und Rietz 2015).

Huber und Rietz (2015) beschränken sich in ihren Ausführungen darauf, die Gütekriterien der Validität und Reliabilität anhand vorhandener Studien zu beschreiben. Hierzu untersuchen sie in einem ersten Teil grundlegende Studien zur Beobachtungsgüte von Direct Behavior Ratings, indem sie Studien zur Übereinstimmung mit einem „wahren" Verhaltenswert, der Varianz in den Messdaten (Interrater Reliabilität) sowie die Stabilität von DBR-Beurteilungen desselben Experimentalvideos zu zwei Zeitpunkten (Intrarater-Reliabilität) betrachten. In einem zweiten Teil werden anschließend spezifische Aspekte der Methode DBR fokussiert. Genauer werden Ergebnisse zu Skalendesign, Itemanzahl, Anzahl der Messzeitpunkte, Wahl des Beobachtungsziels, Formulierung des Beobachtungsziels, Valenz der Zielformulierung, Länge der Verhaltensstichproben und die Auswirkungen von Beobachtertrainings zusammengetragen (vgl. ebd.). Die Ergebnisse werden im Folgenden kurz dargelegt.

Für die *Übereinstimmung mit einem „wahren" Verhaltenswert* beziehen sich die Autoren auf Studien, in denen mögliche Urteilsfehler genauer betrachtet wurden. Steege (2001) ließ das Verhalten von fünf auffälligen Schülern durch ihre Lehrkräfte sowie geschulte Rater (für direkte systematische Verhaltensbeobachtungen) auf einer Ratingskala einschätzen. Es ergaben sich Übereinstimmungen bei 94% bis 95% der Situationen. Sodass die Autoren eine hohe Kriteriumsvalidität attestieren (vgl. Huber und Rietz 2015). Eine Forschungsgruppe um Riley-Tillman et al. (2008) kommt jedoch zu dem Ergebnis, dass die Beobachtungsgüte stark vom beobachteten Verhaltensaspekt abhängig ist (siehe hierzu Teil 2 des Reviews). Die Befunde schwankten in 67% bis 93% der Fälle jedoch nur um +/- 1 Punkt von dem Wert, der durch eine direkte systematische Verhaltensbeurteilung zustande kam (vgl. Huber und Rietz 2015).

In Bezug auf die *Interrater-Reliabilität* ließ sich feststellen, dass zwischen 47% und 48% der Varianz in den Beurteilungen durch das Schülerverhalten erklärbar sind (vgl. Briesch, Chafouleas, Riley-Tillman 2010). Zusätzlich lässt sich erkennen, dass etwa 20 % der Varianz in Beurteilungen durch Interaktionen zwischen Lehrkraft und Schüler erklärbar sind im Gegensatz zu Direkten systematischen Verhaltensbeobachtungen, bei denen dieser Aspekt lediglich 1% der Varianz ausmache (vgl. ebd.). Dies lässt insgesamt auf eine starke Prägung durch beobachtende Personen schließen. Christ, Riley-Tillman, Chafouleas und Jaffery (2011) konnten jedoch herausstellen, dass die

Varianz zudem stark vom zu beobachtenden Verhalten abhängig ist. (vgl. Huber und Rietz 2015).

Auf die *Intrarater-Reliabilität* wird lediglich in einer Studie Bezug genommen. So zeigten Riley-Tillman et al. (2011) wiederholt Videoausschnitte und ließen diese bewerten. Es zeigte sich, dass die Test-Retest-Reliabilität mit $r>0.7$ in einer hohen Weise korrelierte, die Bewertungen also große Übereinstimmungen zeigten (vgl. Huber und Rietz 2015).

Der zweite Teil des systematischen Reviews fokussiert verschiedene Aspekte von Direct Behavior Ratings. Die wichtigsten Ergebnisse der Autoren sollen im Folgenden kurz dargelegt werden.

Bezogen auf Studien zu *Skalendesigns* zeigen sich keine Einflüsse der Skalenbreite und Skalenlänge. Die Studien hierzu beziehen sich auf die Generalisierbarkeitstheorie, welche eine genauere Analyse des Messfehlers ermöglicht, sodass untersucht werden kann, wie viel Varianz durch Skalenlänge oder -breite erklärt werden kann. Eine Sekundäranalyse von Christ, Riley-Tillman und Chafouleas (2011) empfiehlt die Verwendung von sechsstufigen Skalen, wobei die Grundlage der Empfehlung ungeklärt bleibt. Riley-Tillman et al. (2011) untersuchten die Skalenbeschriftung in Zeit (in Sekunden) gegenüber prozentualer Zeitanteil je Zeiteinheit) auf die Beobachtungsgüte. Auch hier konnte kein signifikanter Einfluss gefunden werden (vgl. Huber und Rietz 2015).

Die *Anzahl der Items* hängt natürlich von der Verwendung von Multi-Item-Skalen (MIS) oder Single-Item-Skalen (SIS) ab. Volpe und Briesch (2012) untersuchten die Formate bezogen auf ihre Kriteriums-Validität mit Hilfe der Generalisierbarkeitstheorie vergleichend miteinander. Die Autoren kamen zu dem Ergebnis, dass bei einer SIS weitaus mehr Beurteilungsvarianz unerklärt blieb (ca. 33%) als bei einer MIS (5% und 26%). Im Feld stieg der Anteil unerklärter Varianz einer SIS zudem stärker an, als bei einer MIS. Die Autoren untersuchten des Weiteren, wie viele Wiederholungsmessungen notwendig sind, um eine akzeptable Messgenauigkeit im Vergleich zu einer direkten systematischen Verhaltensbeobachtung durch geschulte Rater zu erlangen. Die Autoren stellten fest, dass bei der Arbeit mit MIS schneller Übereinstimmungen erreicht werden konnten. Hier zeigt sich gegebenenfalls der Vorteil, dass bei MIS

durch die verschiedenen Items und ihre Zusammenfassung zu Mittelwerten (mathematisch) schneller Annäherungen erreicht werden können (vgl. Huber und Rietz 2015).

Für die *Anzahl der Messzeitpunkte* wurde die Frage gestellt, wie viele Messzeitpunkte notwendig sind, um reliable Ergebnisse zu erzielen. Starker Bezug wird auf Christ, Riley-Tillman und Chafouleas (2009) genommen. Diese kommen zu dem Ergebnis, dass die Anzahl der benötigten Messzeitpunkte in starker Weise von den gewählten Verhaltenszielen abhängt. Für SIS in den Bereichen Teilnahme am Unterricht und störendes Verhalten waren demnach fünf Beurteilungssituationen notwendig. Des Weiteren kann aufgezeigt werden, dass längere Verhaltensstichproben zu einer höheren diagnostischen Güte führten, als kurze Sequenzen (vgl. Huber und Rietz 2015).

Bezogen auf die *Wahl des Beobachtungsziels* wurden ja schon zu Beginn des Kapitels Zusammenhänge mit der Art des beobachteten Verhaltens zur Beobachtungsqualität des Tests vermutet. Es ließ sich feststellen, dass bei bestimmten Beobachtungszielen („Teilnahme am Unterricht", „störendes Verhalten") eine größere Validität und Interraterübereinstimmung festgestellt werden konnte. Dahingegen treten bei „angemessenem Verhalten" oder „respektvollem Verhalten" größere Diskrepanzen zwischen Ratern eines DBRs und systematischen direkten Verhaltensbeurteilungen auf. Es stellt sich also heraus, dass die einzelnen Beobachtungsziele eine sehr unterschiedliche Güte besitzen können (vgl. Chafouleas, Jaffery, Riley-Tillman, Christ & Sen 2013 In: Huber und Rietz 2015).

Betrachtet man Studien zu *Formulierungen des Beobachtungsziels* kommt eine Studie (Christ, Riley-Tillman, Chafouleas, Jaffery, 2011) in Betracht, in der die Formulierung des Beobachtungsziels bei SIS in Bezug auf Reliabilität und Validität (Kriterium war die Übereinstimmung der Ratings mit SDO-Ratern) untersucht wurde. Festgestellt werden konnte, dass die Validität globaler Zielformulierungen über fast alle Verhaltensbereiche hinweg einer spezifischen Zielformulierung überlegen waren. Zudem zeigt sich auch in Bezug auf die Interraterreliabilität bei globalen Zielformulierungen höhere Korrelationen (vgl. Huber und Rietz 2015).

Ein weiterer Beobachtungsaspekt bei Direct Behavior Ratings ist die *Valenz von Zielformulierungen*, also die Wertung, die eine Zielformulierung enthält. In zwei Studien (Riley-Tillman et al. 2009 und Chafouleas et al. 2013) wurde der Einfluss positiver und negativer Zielformulierungen bei SIS in Bezug auf die Interraterreliabilität von

DBR- und SDO-Ratern untersucht. Es ließ sich herausarbeiten, dass positive globale Zielformulierungen bei der Beobachtung zu einer höheren Beobachtungsgenauigkeit führten und eine negative Zielformulierung für den Bereich „störendes Verhalten“ (vgl. Huber und Rietz, 2015). Als Ergebnis ließe sich eventuell herleiten, dass die Anwesenheit eines zu beobachtbaren Verhaltens einfacher zu beobachten ist als ein Fehlen.

Die *Länge der Verhaltensstichproben* wurde von Riley-Tillman et al. (2011) genauer betrachtet. Es konnte festgestellt werden, dass die Test-Retest-Reliabilität für zehnminütige Beobachtungssequenzen etwas geringer als für 20-minütige Sequenzen ausfiel. Zudem wird festgehalten, dass ein Mittelwert mehrerer Ratings zu einer höheren Reliabilität führte, sodass der Mittelwert mehrerer kurzer Verhaltensstichproben ein robusterer Kennwert für Verhaltensbeurteilungen sei, als eine Messung über einen längeren Zeitraum (vgl. Huber und Rietz 2015).

Auswirkungen von Beobachtertrainings auf die Messgenauigkeit wurden von Schlientz et al. (2009) sowie LeBel et al. (2009) untersucht. Die Autoren kommen zu dem Ergebnis, dass die Beobachtungen der trainierten Beobachter signifikant höher waren als die der Kontrollgruppe. Die Genauigkeit wurde durch die Interrater-Übereinstimmung und die Abweichung von Ergebnissen im Vergleich zu einer direkten systematischen Verhaltensbeobachtung ermittelt. LeBel kommt zu dem Ergebnis, dass ein kurzes einführendes Training im Vergleich zu einem intensiven Beobachtertraining keine signifikante Verbesserung der Beobachtungsgüte nach sich zieht. Insgesamt kann man also zusammenfassen, dass eine kurze Einführung in den Umgang mit DBR-Skalen ausreicht um eine hohe Genauigkeit der Messungen zu erhalten (vgl. Huber und Rietz 2015). Dies scheint eine Implementierung und Nutzung des Instruments ohne große Vorkenntnisse möglich zu machen.

Insgesamt folgern die Autoren aus den Befunden, dass sich Direct Behavior Ratings als „Fundament für evidenzbasierte Entscheidungsprozesse in multiprofessionellen Teams, wie sie beispielsweise im Response-to-intervention-Konzept empfohlen werden“ (ebd. S.93) eignen. Als Anwendergruppe für die Ratings werden die Lehrkräfte herausgestellt, da die Methode auf hochfrequenten Messungen beruht, sodass andere Personengruppen als Anwender nicht in Frage kämen. Die Autoren betonen zudem, dass die Befundlage im Feld noch zu gering sei. Dies wird auch daran deutlich, dass lediglich 17 Studien herangezogen werden, welche häufig von denselben Autoren

stammen und lediglich kleine Stichprobengrößen aufweisen. Zudem wird genannt, dass insbesondere Instrumente zur Verhaltensverlaufsdiagnostik erst noch entwickelt werden müssten. Dies wird auch von Casale et al. (2015a) betont. Die Autoren weisen auf eine gute Testgüte hin und konstatieren zwei Funktionen, die die Direkte Verhaltensbeurteilung (DBR) ermöglicht: Zum einen könne diese für eine Überprüfung der kindbezogenen, individuellen Passung einer Fördermaßnahme und daraus resultierend ggf. eine Modifikation dieser genutzt werden. Zum anderen könne sie zur Überprüfung der Wirksamkeit von wissenschaftlich nicht überprüften Fördermaßnahmen im Sinne einer Evidenzbasierung im Einzelfall genutzt werden (vgl. ebd.).

Eine Forschergruppe, welche die Eignung von Direct Behavior Ratings insbesondere unter empirischer Perspektive in Deutschland betrachtet, besteht aus den Autoren Casale, Grosche, Volpe und Hennemann. Interessant ist ihre Forschung, da sie insbesondere die Verwendung von MIS gegenüber SIS vergleicht. Von Erkenntnisinteresse für die Autoren ist insbesondere die Reliabilität von Direct Behavior Ratings, da diese die Auseinandersetzung mit möglichen Messfehlern des Instrumentes ermöglicht. Durch die Methode und Anwendung der Generalisierbarkeitstheorie lässt sich der Faktor des nicht erklärbaren Messfehlers in Facetten, die diesen erklären können unterteilen. Im Kontext von Direct Behavior Ratings sind dies insbesondere die beurteilenden Personen (Inter-Rater-Reliabilität), die genutzten Items (interne Konsistenz) und die Messzeitpunkte (Test-Retest-Reliabilität) (vgl. Casale et al. 2017).

Zu Beginn werden bisherige Ergebnisse zur Fragestellung der Forschergruppe wie folgt zusammengefasst:

Einfluss der Rater: Bisherige Studien zur Inter-Rater-Reliabilität konnten zeigen, dass die Facette Rater je nach Personengruppe zwischen 0 und 41% der Gesamtvarianz schwankt. Insgesamt konnten Ratervarianzen von 3 bis 5% bei Multi-Item-Skalen festgestellt werden.

Einfluss der Messzeitpunkte: Zwischen 0 und 20 % der Gesamtvarianz konnte durch Messzeitpunkte erklärt werden. Hier ergaben sich keine Unterschiede zwischen Multi-Item-Skalen und Single-Item-Skalen.

Anzahl an notwendigen Messungen für reliable Befunde: Die Ergebnisse weisen darauf hin, dass ein ausreichender statistischer Generalisierbarkeitskoeffizient bereits nach einem bis sechs Messzeitpunkten erreicht werden kann, das gleiche gilt nach zwei

bis zwölf Messungen für den Zuverlässigkeitskoeffizienten. Ein wichtiger Aspekt in einer Studie nach Chafouleas scheint zu sein, dass Lehrkräfte die regelmäßig in der Klasse arbeiten, schneller reliable Ergebnisse für die Bewertung ihrer Schülerschaft erzielen (vgl. Casale et al. 2017).

In der Studie, in der mit fünf Schülern mit externalisierenden Verhaltensauffälligkeiten gearbeitet wurde, welche von einer Grundschul- und einer sonderpädagogischen Lehrkraft regelmäßig mit Hilfe eines DBRs mit Single-Item-Skalen und Multi-Item-Skalen bewertet wurden, zeigte sich, dass Multi-Item-Skalen wesentlich geringere Anteile der Varianz bei den Facetten Rater (2,5%) und Messzeitpunkte (5,2%) im Vergleich zu Single-Item-Skalen aufweisen (Rater (18,1%), Messzeitpunkte (17,9%)). Die Autoren leiten hieraus her, dass das Verhalten mit Multi-Item-Skalen von beiden Lehrkräften ähnlich und weniger sprunghaft eingeschätzt werden konnte.

Zudem zeigt sich, dass die Messzeitpunkte einen starken Einfluss auf die Gesamtvarianz haben (SIS (19,8%); MIS (22,1%)). Die gleichen Schüler wurden zu unterschiedlichen Zeitpunkten unterschiedlich beurteilt. Dies kann durch die mögliche Sprunghaftigkeit des Verhaltens über die Zeit erklärt werden und zeigt die Änderungssensibilität der Instrumente.

Die Autoren folgern aus den Ergebnissen, dass Direct Behavior Ratings zeitnah eine Datengrundlage für pädagogisch-praktische Entscheidungen über den Erfolg von Fördermaßnahmen schaffen können und für mehrstufige Fördersysteme im Sinne des RTI (Kapitel 2.3) geeignet scheinen. Zudem verweisen die Ergebnisse darauf, dass die Nutzung der konkreten und spezifischen MIS den SIS vorzuziehen ist (vgl. Casale et al. 2017, Gebhardt et al., 2019).

Insgesamt lässt sich feststellen, dass es vielerlei Forschung zur psychometrischen Eignung von DBRs sowie ihrer formalen Gestaltung gibt. Diese weisen auf die generelle Eignung als diagnostisches Instrument der Verhaltensverlaufsdiagnostik sowie auf Hinweise für einzelne Aspekte bei der Testgestaltung hin. Diese können, wie in Kapitel 5.1 dargestellt, für die Erstellung des Ratinginstrumentes „PUTSIE“ verwendet werden. Weitgehend unberücksichtigt bleiben jedoch Fragestellungen zur tatsächlichen Anwendbarkeit des Instrumentes durch die Anwender (Lehrkräfte). Diese wird von den Autoren postuliert, jedoch nicht nachgewiesen. Im Folgekapitel soll diese Forschungslücke genauer betrachtet werden.

3.3 Forschungslücke „Implementierung des Direct Behavior Ratings“

Die Anwendbarkeit von Direct Behavior Ratings und dessen Eignung für den inklusiven Unterricht wird von den Autoren angenommen, jedoch nicht genauer untersucht. Es stellt sich die Frage, ob die Anforderungen, die mit der Durchführung und Auswertung der Methode verbunden sind auch durch Lehrkräfte im schulischen Alltag wahrgenommen werden können und von diesen akzeptiert wird. Hierauf verweist unter anderem auch die Forschungsgruppe des Rügener Inklusionsmodells:

> „Da sich in der Forschung über Schulreformen und die Umsetzung von innovativen Unterrichtskonzepten häufig zeigt, dass die involvierten praktisch tätigen Pädagogen nicht vollständig für umfassende Veränderungen gewonnen werden können, ihre Akzeptanz jedoch eine wichtige Komponente für eine erfolgreiche Implementation der Reform darstellt, ist die Erfassung ihrer Einstellung von großer Bedeutung.“ (Voß et al. 2015, S. 134–135)

Gebhardt et al. (2015) verweisen zudem darauf, dass Testungen der Verlaufsdiagnostik auch auf ihre Anwendbarkeit überprüft werden müssten. Hierzu gibt es bisher jedoch lediglich wenige bis keine Forschungsergebnisse im Bereich der Verhaltensverlaufsmessung. Dies verwundert insofern, da im Bereich evidenzbasierter sonderpädagogischer Förderung als Säulen einer Förderung auch auf die Praxis verwiesen wird und Erreichbarkeit, Praktikabilität und individuelle Passung als wichtige Aspekte für die Umsetzbarkeit von evidenzbasierten Fördermaßnahmen genannt werden (vgl. Casale et al. 2015a). Wenn im Sinne einer evidenzbasierten Förderung durch die Forschung evaluierte Methoden Anwendung finden sollen, muss auch die Anwendbarkeit dieser in der schulischen Praxis berücksichtigt werden. Dies postuliert auch Prenzel (2010) in einem Diskussionsbeitrag, indem er darauf verweist, dass häufig theoretisch evidente, gut erarbeitete Konzepte nicht gut und leicht in die Praxis transferierbar seien. Er postuliert, dass es stärker das Ziel von Wissenschaft sein müsse, durch Wissens- und Lerntransfers aus der Forschung entwickelte Methoden in die Bildungspraxis zu übertragen. Dies könne seiner Meinung nach durch eine „nutzeninspirierten Grundlagenforschung“, welche Nutzenüberlegungen in das Forschungsinteresse miteinbezieht und diese prüft gewährleistet werden (vgl. Prenzel 2010).

In Bezug auf Direct Behavior Ratings meint ein solches Vorgehen, zu untersuchen, inwieweit aus der Theorie hergeleitete Annahmen zum Umgang mit dem Instrument tatsächlich in der Praxis an- beziehungsweise übernommen werden können, um Rückschlüsse zu ziehen, an welchen Stellen eine Implementierung der Methode erschwert

ist. Dies gilt insbesondere auch für Tests im Projekt LEVUMI, da die Tests den Lehrkräften online nur mittels eines Handbuchs und Videos erklärt werden und sich die Lehrkräfte die Methode so größtenteils selbstständig aneignen müssen. Auch aus diesem Grund müssen die Tests auf Anwendbarkeit und Interpretierbarkeit geprüft werden (vgl. Gebhardt et al. 2015).

Die Arbeit widmet sich daher der Frage, inwieweit die Methode von Direct Behavior Ratings aus Sicht der Nutzer (Lehrkräfte) anwendbar scheint und von diesen akzeptiert wird. So konnte im Abschnitt gezeigt werden, dass eine gelingende Implementierung insbesondere mit der Handhabbarkeit und den Einstellungen von Lehrkräften zu einem Instrument zusammenhängt. Da Direct Behavior Ratings durch verschiedene Items, Skalen, etc. sehr heterogen erscheinen und nur ein festgelegtes Diagnoseinstrument, welches von allen Lehrkräften auf ähnliche Weise genutzt wird als überprüfbar erscheint, muss dieses jedoch zuvor dargelegt werden. Für die Erhebung und Weiterentwicklung im Bereich von LEVUMI wurde daher das Direct Behavior Rating „PUTSIE“ erstellt, welches im Methodenteil genauer erläutert wird (Kapitel 5.1). Die Forschungsfragen, welche diese Arbeit begleiten lauten daher:

> Ist das Direct Behavior Rating „PUTSIE“ für Lehrkräfte handhabbar?
>
> Wird das Ratinginstrument „PUTSIE“ als Instrument zur Verhaltensverlaufsdiagnostik von Lehrkräften akzeptiert?

Teil B: Methode

Um die Forschungsfragen zu beantworten soll das Direct Behavior Rating „PUTSIE“ an einem inklusiven Schulsystem implementiert und die Implementierung im Sinne der Forschungsfragen evaluiert werden. Kapitel 4 des Methodenteils dient daher der Beschreibung der Stichprobe, insbesondere der Darstellung des Schulsystems an der die Erhebung durchgeführt werden soll. In Kapitel 5 wird anschließend das Forschungsdesign genauer dargestellt. Zunächst wird das Material (insbesondere das Direct Behavior Rating „PUTSIE“) mit dem die Erhebung durchgeführt wurde, sowie dessen Erstellung dargelegt. Anschließend erfolgt die Beschreibung des Ablaufs sowie deren theoretische Begründung. Kapitel 5 endet mit der Darstellung und Begründung der Erhebungs- und Auswertungsmethode.

4. Stichprobenbeschreibung

Die Beschreibung der Stichprobe soll auf drei Weisen geschehen. Zum ersten soll das Schulsystem an dem die Erhebung durchgeführt wurde vorgestellt werden um die Umgebungsfaktoren zu erläutern, in die das Instrument eingebettet wird. Zweitens wird das Schulsystem zudem mit den in Kapitel zwei erläuterten Eigenschaften des Response-to-Intervention Modells abgeglichen. Drittens folgt eine Beschreibung der an der Studie teilnehmenden Schüler und Lehrkräfte, um die Stichprobe genauer darzulegen.

Die staatliche Grundschule, an der die Erhebung durchgeführt wurde liegt im Zentrum einer Mittelstadt in NRW. In ihrem Einzugsgebiet ist ein hoher Anteil von Menschen mit zugeordnetem Migrationshintergrund erkennbar, welcher sich auch in der Schülerschaft wiederspiegelt. Insgesamt besuchen 188 Schülerinnen und Schüler die Schule. Hiervon wird etwa 85% ein Migrationshintergrund zugewiesen. Die Klassen sind jeweils zweizügig von Klasse eins bis vier. Jede Klasse wird von einem Klassenlehrer oder einer Klassenlehrerin betreut, sodass es acht Lehrkräfte mit einer Klassenleitung gibt. In einer Klasse wird die Klassenleitung aufgeteilt. Das Kollegium besteht aus insgesamt 24 Lehrkräften. Hiervon sind zwei sonderpädagogische Lehrkräfte mit jeweils einer 3/4-Stelle an der Schule angestellt. Durch den hohen Anteil von Kindern mit Migrationshintergrund gibt es Sprachlehrkräfte für Türkisch und Arabisch sowie eine Lehrkraft für Islamkundeunterricht, zusätzlich eine Integrationskraft für Schülerinnen und Schüler mit Migrationshintergrund in der ersten Klasse. Eine Schulsozialarbeiterin und eine Sozialpädagogin verstärken das Team. Der Unterricht an der Versuchsschule findet im klassischen 45-Minuten-Rhythmus statt. Direkt in der Schule befindet sich der offene Ganztag unter privater Trägerschaft. Dieser wird von 70 Kindern besucht. Hier finden sich Freizeitangebote, eine Hausaufgabenbetreuung (wird von einer Grundschullehrkraft betreut) sowie das Mittagessensangebot.

Das „Gemeinsame Lernen“, worunter die Schule den gemeinsamen Unterricht von Kindern mit sonderpädagogischem und ohne sonderpädagogischem Förderbedarf versteht, wird im Schulprogramm durch Leitziele definiert.

Ziel 1 formuliert, dass die Gesamtverantwortung aller Kinder bei der allgemein bildenden Schule liegt und Grundschul- als auch Förderschullehrkraft den Unterricht gemeinsam verantworten. Unterrichtseinheiten sollen gemeinsam geplant und differen-

ziertes Material von beiden Kollegen für alle bereitgestellt werden. Zudem wird betont, dass viel Unterricht gemeinsam durchgeführt und Regel- als auch Förderschullehrkraft für alle Kinder als Ansprechpartner zur Verfügung stehen.

Ziel 2 bezieht sich explizit auf die Diagnostik des Entwicklungs- und Lernstands der Schüler, welcher dokumentiert werden soll, um weitere Fördermaßnahmen zu entwickeln. Hierzu nutzt die Schule das Computerprogramm „Förderplaner“ zur Erstellung von individuellen Förderplänen. Des Weiteren wird auf verschiedene informelle Testverfahren verwiesen, beispielsweise auf Heuers Material „Beurteilen. Beraten. Fördern“ (Heuer, 2008) zur Diagnose, Therapie und Bericht-/Gutachtenerstellung bei Lern-, Sprach- und Verhaltensauffälligkeiten in Vor-, Grund- und Sonderschule und den Mann-Zeichen-Test nach Hermann Ziler. Als individuelle zusätzliche Förderprogramme der Schule werden Schreibtanz, das Marburger Konzentrationstraining, ein Antiaggressionstraining sowie ein Motorikkurs genannt. Es wird betont, dass die Programme auch Regelschülerinnen und -schülern offenstehen.

Die Befähigung, verschiedene Lernwege zur Kompetenzerweiterung zu nutzen ist die dritte Zielsetzung der Schule für das gemeinsame Lernen. Anforderungen sollen an individuelles Arbeitstempo und Fähigkeiten angepasst sein, sodass individuelles Material erarbeitet und genutzt wird. Durch die Arbeit mit Wochen- beziehungsweise Tagesplänen, welche Arbeitsaufträge in dreifach differenzierter Form enthalten, wird zudem selbstständiges und eigenverantwortliches Lernen erprobt. Die individuelle Leistung im Bereich des Verhaltens wird täglich durch Verstärkerpläne, die in jeder Klasse für jedes Kind installiert sind zurückgemeldet. Zusätzlich wird Wert auf Partner- und Gruppenarbeit, sowie Helfersysteme zur Förderung von Sozialkompetenzen gelegt. An dieser Stelle wird auch auf das Programm „Faustlos“ mit einer wöchentlichen Gesprächsrunde zur Rückmeldung des eigenen Verhaltens verwiesen. Da dieses ein Rahmenprogramm für die Prävention und Intervention bei Verhaltensstörungen ist und somit Bereiche des Direct Behavior Ratings betrifft, soll es im Folgenden kurz dargelegt werden.

„Faustlos“ nennt sich das Programm zur Gewaltprävention der Schule. Ziel des Programms ist es, mit allen Kindern konstruktive Formen der Problem- und Konfliktbewältigung zu trainieren und damit die soziale Kompetenz zu steigern. Das Programm basiert auf der deutschsprachigen Version des „Second Step Curriculums“. Dieses wurde vom Commitee for Children in Seattle entwickelt. Das Curriculum wendet sich

an Kinder der Jahrgangsstufe eins bis drei und ist für Gruppen konzipiert. Das Curriculum wird von den jeweiligen Klassenlehrern verantwortet und in den Unterricht integriert. Eine Unterrichtsstunde pro Woche wird für die Vermittlung der Inhalte verwendet. Mit den Kindern, Eltern und Lehrkräften wird ein Vertrag abgeschlossen, der es verbietet Kindern (im Herzen) weh zu tun und Sachen zu beschädigen. Kommt es zu Vertragsverstößen folgt ein Interventionsgespräch, indem die Beteiligten den Tathergang schildern, die Sicht von Opfer und Täter zum Geschehenen dargestellt wird und anschließend die Festlegung von Strafe, beziehungsweise Wiedergutmachung erfolgt. Zusätzlich gibt es jeden Freitag eine Reflexionsstunde, in der jedes Kind sein Sozialverhalten der letzten Woche einschätzen soll und dieses gegebenenfalls mit einem Verstärker belohnt wird.

Als viertes Ziel wird die regelmäßige Rückmeldung über den individuellen Entwicklungs- und Lernstand als Ziel für das gemeinsame Lernen festgelegt. Hierzu werden seitens der Schule drei Strategien genannt. Regelmäßige Klassenpflegschaftssitzungen und Infoabende für Eltern von Kindern mit sonderpädagogischem Förderbedarf sollen dem Austausch dienen. Des Weiteren gibt es halbjährliche Elternsprechtage mit Regel- und Förderschullehrkraft zur individuellen Beratung. Bei Bedarf finden zusätzliche Elterngespräche statt. Als dritte Strategie werden interdisziplinäre Gespräche mit Therapeuten, Jugendamt, Familienhelfern oder Ärzten genannt.

Es wurde schon erwähnt, dass die Schule einen hohen Anteil an Kindern mit Migrationshintergrund aufweist. Die Schule weist daher von sich aus darauf hin, dass die alltägliche Kommunikation zwischen den Kindern und Lehrkräften meistens gelingt, jedoch häufig grundlegende Deutschkenntnisse erst in der Schule erworben werden. Aus diesem Grund wird insbesondere auf die sprachliche Arbeit mit den Kindern Wert gelegt. Sprachhandeln wird durch tägliche Erzählrunden geübt. Gerade in der Schuleingangsphase wird auf Doppelbesetzung und Förderangebote im sprachlichen Bereich Wert gelegt. Insbesondere Seiteneinsteiger (ohne Deutschkenntnisse) erhalten durch die Integrationskraft sprachliche Einzel- oder Kleingruppenförderung.

Vergleicht man die obigen Darstellungen mit Anforderungen an ein inklusives Schulsystem für den Bereich von Kindern mit Verhaltensschwierigkeiten welches mit dem Ansatz des Response-to-Intervention arbeitet (siehe insbesondere die Ausführungen zu Hillenbrand (2015) in Kapitel 3.1) lässt sich folgendes feststellen:

Förderebene 1: Die Schule verweist auf einige Testinstrumente zur regelmäßigen Erfassung des Lern- und Entwicklungsstandes der Kinder. Es sind an dieser Stelle jedoch keine Screeningprogramme zur Evaluation ganzer Klassen genannt. Durch universelle Programme wie „Faustlos“ und Verstärkerpläne, welche als evidenzbasiert gelten können, wird das Verhalten aller Schüler regelmäßig eingeschätzt und gefördert. Zusätzlich werden durch klare Strukturen, welche unter anderem durch die sprachlichen Einschränkungen der Kinder notwendig werden, Merkmale des Classroom-Managements genutzt (Verstärkerpläne, Klassenregeln mit Vertrag, sichtbare Klassenregeln in jeder Klasse).

Förderebene 2: Schülerinnen und Schüler die trotz der auf Förderebene 1 genannten Maßnahmen Auffälligkeiten zeigen erhalten individualisierte Förderangebote, welche von der Klassenlehrkraft durchgeführt, jedoch mit der sonderpädagogischen Lehrkraft besprochen werden. Die Schülerinnen und Schüler werden zu Elterngesprächen eingeladen, in denen die Lehrkräfte ihnen Rasterbögen austeilen, welche in den Klassen 1 bis 3 auch in Rasterzeugnissen enthalten sind. Die Bögen und Zeugnisse beschreiben das Arbeits- und Sozialverhalten des Schülers aus Sicht der Lehrkraft und dienen zusätzlich der Bestimmung individueller Fördermaßnahmen. Die Lehrkräfte nutzen die Bögen als Förderpläne. Eine weitere Erfassung des Verhaltens mittels verlaufsdiagnostischer Methoden erfolgt nicht. In Klasse 1 und 2 erfolgt auf dieser Ebene eine Zuweisung zu Antiaggressions- oder Konzentrationstraining, in einer Klasse wird zusätzlich mit dem Klasse-Kinder-Spiel gearbeitet. Die Angebote werden durch die sonderpädagogischen Lehrkräfte gestaltet.

Förderebene 3: Auf dieser Ebene werden die Schülerinnen und Schüler beschult, die zusätzlich weitere individuelle Begleitung und Förderung benötigen. Diese findet sich durch die sonderpädagogischen Lehrkräfte, aber auch Angebote der Schulsozialarbeit. Es finden Gespräche mit schulexternen Fachkräften (meist Jugendamt und Kinder- und Jugendpsychiatrie) statt. Durch die sonderpädagogische Lehrkraft werden weitere Maßnahmen besprochen. In Klasse 1 und 2 findet diese Arbeit präventiv ohne die Vergabe eines sonderpädagogischen Förderschwerpunkts statt. Ab Klasse 3 wird die Arbeit meist mit der Vergabe eines Förderschwerpunkts im Bereich der sozialen und emotionalen Entwicklung verknüpft. In Bezug auf die Förderpläne werden auf dieser Stufe den Schülerinnen und Schülern neben den Rasterzeugnissen und -berichten noch

zwei Förderziele zugewiesen, welche durch die sonderpädagogische Lehrkraft dargelegt und deren Entwicklung beschrieben wird. Die Förderung findet hier meist im Klassenverbund statt, durch die Zuweisung zu den oben genannten Fördergruppen jedoch auch in Kleingruppen. Eine Einzelförderung ist im Bereich Verhalten am Schulsystem nur im Rahmen der Konfliktgespräche nach „Faustlos“ üblich.

Zusammenfassend lässt sich festhalten, dass die Schule im Sinne des RTI-Modells viele Interventionsmaßnahmen nutzt und umsetzt. Eine explizite Arbeit auf den drei Ebene mit dem Rahmenmodell wird jedoch nicht offen ersichtlich oder kommuniziert. Zudem zeigt sich eine Lücke bei der regelmäßigen Verwendung von diagnostischen Instrumenten, wie sie laut RTI ab Ebene 2 verwirklicht werden soll. Auf Ebene 1 findet ein Monitoring fast ausschließlich auf Grund der gemachten Beobachtungen der Lehrkräfte ohne diagnostische Instrumente statt. Es zeigt sich also, dass die Lehrkräfte der Versuchsschule bisher keine Erfahrungen mit Instrumenten der Lernverlaufsdiagnostik gemacht haben und diese zunächst kennenlernen müssen, das Instrument „PUTSIE“ jedoch eine sinnvolle Ergänzung der bisherigen Arbeitsweise der Lehrkräfte darstellen kann. Aus diesem Grund scheint die Schule für die Durchführung einer Implementationsstudie geeignet (Kuhl et al., 2017).

An der Erhebung, welche in Kapitel 5 genauer dargestellt wird, nahmen insgesamt 9 Lehrkräfte teil. Zwei davon schieden aus Krankheitsgründen oder anderweitigen Verpflichtungen vorzeitig aus der Erhebung aus oder konnten nicht regelmäßig teilnehmen, sodass sie in der Auswertung unberücksichtigt bleiben. Von den restlichen Lehrkräften sind 6 Lehrkräfte mit einer Klassenleitung beauftragt. Zwei Lehrkräfte betreuen eine erste Klasse, ebenso eine Lehrkraft eine zweite, eine Lehrkraft eine dritte und zwei Lehrkräfte je eine vierte Klasse. Eine sonderpädagogische Lehrkraft nahm zusätzlich an den Erhebungen der ersten und zweiten Klassen teil. Je Klasse hatten die Lehrkräfte die Möglichkeit bis zu zwei Kinder zu bewerten. Insgesamt wurden daher elf Schülerinnen und Schüler mit „PUTSIE“ bewertet. Eine übersichtliche Darstellung der Stichprobe ist Anhang A zu entnehmen.

5. Forschungsdesign

Die Darstellung und Begründung des geplanten Ablaufs sowie des Materials und der Auswertungsmethode, welche zur Beantwortung der Forschungsfragen genutzt werden soll, stellt den Inhalt dieses Kapitels dar.

Zuvor soll jedoch das Projekt LEVUMI erläutert werden, welches in Kapitel 2 knapp skizziert wurde und an welches diese Arbeit angebunden ist. Ziele der Lernplattform sind, ein frei verfügbares Onlineinstrument zur Lernverlaufsmessung zur Verfügung zu stellen, die Forschung zu Lernverlaufsmessung und ihrer Akzeptanz voranzubringen sowie diagnostische Maßnahmen zu verbessern und gezielte Förderung zu entwickeln (vgl. Mühling et al. 2017). Dabei wird Wert daraufgelegt, dass die verfügbaren Tests insbesondere für Schüler mit sonderpädagogischem Förderbedarf konzipiert werden. Durch die Anbindung an eine Onlineplattform erhalten Lehrkräfte einen kostenlosen, einfachen Zugang zu Instrumenten der Lernverlaufsdiagnostik, die Forschergruppe im Gegenzug einen ständigen Feldzugang, welcher genutzt werden kann um die Qualität (insbesondere die psychometrischen Gütekriterien) der Tests zu überprüfen und diese gegebenenfalls zu verbessern. Ein Vorteil für die Schulen ist das Onlineformat auch daher, weil wenig Anforderungen an die IT-Infrastruktur der Einzelschule gestellt wird. Lerner können die Testungen gegebenenfalls sogar mit ihren eigenen Endgeräten nutzen (vgl. ebd.).

Derzeit finden sich für die Lernplattform Tests in den Bereichen Deutsch (Leseflüssigkeit, Rechtschreibung, sinnentnehmendes Lesen, Wortschatz) und Mathematik (Zahlen lesen, Zahlenreihen, Zahlenstrahl) (vgl. www.levumi.de, letzter Zugriff am 21.01.19). Bei den Tests wird darauf geachtet mit sogenannten robusten Indikatoren zu arbeiten. Es werden Aufgabentypen gesucht und in die Tests aufgenommen, die die geforderte Kompetenz möglichst gut repräsentieren und hoch mit der relevanten Leistung korrelieren. Dieses Vorgehen steht Ansätzen des curriculum-based-measurements (CBM) gegenüber, bei dem Tests anhand von am Ende eines Schuljahres zu erlangenden Kompetenzen, beispielsweise Lehrplänen konstruiert werden (vgl. Mühling et al. 2017). Ein Vorgehen mit robusten Indikatoren ermöglicht eine geeignetere Diagnostik bei Kindern mit sonderpädagogischem Förderbedarf, da durch den Fokus auf die individuelle Bezugsnorm anhand robuster Indikatoren auch kleine Fortschritte erkannt werden können, während CBM durch die Orientierung an Curricula meist die Bezugsgruppe als Norm fokussieren und „schwache“ Schülerinnen und Schüler immer stärker den Anschluss zur Bezugsnorm (Curriculum) verlieren (vgl. Mühling et al. 2017). Ziel von LEVUMI ist es, die Tests in den vorhandenen Bereichen (Lesen und Mathematik) zu erweitern und durch neue Fächer und Bereiche zu ergänzen (vgl. ebd.). Das Projekt ist daher anschlussfähig für Verlaufsdiagnostik im Bereich des Verhaltens.

Direct Behavior Ratings stellen die Art der Verlaufsmessung dar, welche untersucht werden soll. Um dies zu leisten, muss zunächst ein geeignetes Ratinginstrument vorhanden sein. Da zur Zeit der Durchführung noch kein Ratinginstrument für den Bereich Verhalten auf der Plattform LEVUMI veröffentlicht wurde, wurde das Rating „PUTSIE“ entwickelt, dessen Konstruktion in Kapitel 5.1 genauer dargestellt werden soll. Bei der Erstellung muss berücksichtigt werden, dass das Instrument für die Arbeit als Onlineinstrument geeignet sein muss, auch wenn die Erhebung zunächst als Pen-and-Paper Variante durchgeführt wird. Zudem ist genau zu kennzeichnen, wie die Implementierung und Erhebung gestaltet werden soll, um die Ergebnisse vergleichbar zu machen (Kapitel 5.2). In Kapitel 5.3 wird die notwendige Auswertung der Ratingbögen und ihre Darstellung als Verlaufsgraphen erläutert. In Kapitel 5.4 wird wiederum die Auswertung und Datenerhebung für die Beantwortung der Forschungsfrage dargelegt.

5.1 Der Ratingbogen „PUTSIE“

Um eine Untersuchung von Direct Behavior Ratings in der Praxis durchführen zu können ist ein Ratinginstrument notwendig, dessen Konstruktion an dieser Stelle erläutert werden soll.

LEVUMI fokussiert Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf. Der Förderbedarf im Bereich Verhalten wird durch den Förderschwerpunkt emotionale und soziale Entwicklung abgedeckt. In diesem Bereich gibt es empirische Klassifikationssysteme, wie sie in Kapitel 2.2 mit der Einteilung in externalisierende und internalisierende Verhaltensstörungen dargelegt wurden. Das System eignet sich nur für eine grobe Einteilung von Verhalten, die nicht fein genug für die Erfassung mittels eines Ratings ist. Hierfür sind feiner gegliederte Tests notwendig. Ein solcher ist beispielsweise der Strengths and Difficulties Questionnaire (SDQ), welcher 1997 von Goodman entwickelt wurde. Dieser erfasst folgende Verhaltensdimensionen: Emotionale Probleme, Verhaltensprobleme, Hyperaktivität, Verhaltensprobleme mit Gleichaltrigen und prosoziales Verhalten. Eine Anlehnung an die Verhaltenskategorien für die Erstellung eines DBRs bietet sich für das Rating schon aus Gründen der Vergleichbarkeit an. Die Dimensionen des SDQs wurden in Anlehnung an das psychologische Diagnosemanual „Diagnostic and Statistical Manual of Mental Disorders“ der American Psychiatric Association (2000) hergeleitet. Die stark an klinische Symptome an-

gelehnten Verhaltenskategorien lassen sich gut in externalisierende und internalisierende Verhaltensauffälligkeiten einteilen. Da die Items nicht aus dem SDQ übernommen werden sollen, bietet sich die aktuelle Version des psychologischen Manuals (DSM-V (Falkai et al. 2018)) für die Generierung von Items für das Rating an. Aus diesem Grund wurden die zu erfassenden Dimensionen stärker an die DSM-V angelehnt. Dies führt zu folgenden Änderungen der Verhaltensdimensionen:

Der Dimension Hyperaktivität wird in Anlehnung an die Aufmerksamkeitsdefizit Hyperaktivitätsstörung der DSM-V in die zwei Bereiche „Impulsivität" und „Unaufmerksamkeit" aufgeteilt.

Die Dimension „Verhaltensprobleme" wurde in Anlehnung an die Störung mit oppositionellem Trotzverhalten der DSM-V in den Verhaltensbereich „Trotzverhalten" umgewandelt.

Die Dimension „emotionale Probleme" wird in Anlehnung an Symptome einer Major Depression in den Verhaltensbereich „Emotionalität" abgewandelt.

Die Kategorie Verhaltensprobleme mit Gleichaltrigen wurde in Anlehnung an Sauerland (2018) und deren Items für ein Rating in Anlehnung an den SDQ übernommen. Gleiches gilt für den Bereich schulbezogenes Verhalten, welcher von Gebhardt (2017) durch den engen Zusammenhang von Verhaltensauffälligkeiten und Lernschwierigkeiten als notwendig für die Verlaufsmessung des Verhaltens im schulischen Bereich genannt und daher in dem Ratingbogen eingefügt wurde (vgl. Sauerland 2018).

Anschließend wurden die einzelnen Verhaltensbereiche den Oberbegriffen externalisierende/internalisierende Verhaltensstörungen zugewiesen. So wurden die Bereiche „Unaufmerksamkeit", „Impulsivität" und „Trotzverhalten" aufgrund ihrer symptomatischen Beschreibung den externalisierenden Störungen zugewiesen und die Bereiche „Emotionalität" und „Probleme mit Gleichaltrigen" den internalisierenden Störungen. Hierbei wurden die Symptombeschreibungen der DSM-V (Falkai et al. 2018) herangezogen.

Die Anfangsbuchstaben der einzelnen Verhaltensbereiche „**P**robleme mit Gleichaltrigen, **U**naufmerksamkeit, **T**rotzverhalten, **s**chulbezogenes Verhalten, **I**mpulsivität und **E**motionalität lassen sich zu dem Kunstwort „PUTSIE" zusammenfügen und sind namensgebend für das Instrument.

Da die Verhaltenskategorien feststehen, muss nun entschieden werden, welches Skalendesign Verwendung finden soll. Die Wahl fällt auf Multi-Item-Skalen, da die klinischen Verhaltensbereiche nur schwer mit einzelnen Items beschrieben werden können, beziehungsweise ein Item „der Schüler zeigt impulsives Verhalten“ viel zu global und unökonomisch zu beantworten wäre. Zudem konnten Huber und Rietz (2015) aufzeigen, dass die Verwendung von Multi-Item-Skalen zu einer höheren Aufklärung des Messfehlers führen kann. So bleibt das Rating für die Überprüfung der Testgütekriterien geeigneter.

Eine weitere Frage stellt sich in Bezug auf die Items. Wie viele Items können die Lehrkräfte bewerten, ohne, dass die Güte des Tests eingeschränkt wird? Es wurde entschieden, dass aus Gründen der Ökonomie je Verhaltenskategorie maximal sechs Items verwendet werden sollten.

Die Items für die verschiedenen Kategorien mussten nun ausgewählt werden. Hier zeigt sich die Schwierigkeit, dass die DSM-V der Statusdiagnostik von psychischen Erkrankungen dient. So werden möglicherweise globale Items, welche nicht veränderungssensitiv genug für die Verlaufsdiagnostik sind, gewählt. Da jedoch lediglich die Symptome einzelner Störungsbilder als Items verwendet werden, scheint zu große Globalität zumindest nicht zwangsweise notwendig. Die Items müssen zudem klar umschrieben und operationalisiert sein, sodass sie ohne Erläuterung verstanden und zum Rating genutzt werden können (Voß und Gebhardt 2017). Dies ist insbesondere auch deshalb notwendig, da die Lehrkräfte bei einer möglichen Implementierung einer Onlineversion nur den Ratingbogen und ein Handbuch aber wenig weitere Hilfsmittel zur Verfügung haben.

Die zugeordneten Items für die verschiedenen Kategorien wurden daher aus den möglichen Symptomen, die in der DSM-V für die Verhaltensbereiche genannt werden abgeleitet. Insgesamt wurden die Items ausgewählt, welche sich nach Ansicht des Autors im schulischen Bereich beobachten lassen und nicht zu global formuliert wurden. Zudem wurde darauf geachtet, dass die Items die Anwesenheit eines Verhaltens überprüfen, da dies leichter zu erfassen ist. Die Items und dessen Zuordnung zu verschiedenen Verhaltensbereichen sollen im Folgenden tabellarisch dargelegt werden.

Verhaltenskategorie	**Items**	**Quelle**	**Erläuterung**
Schulbezogenes Verhalten	*(1) „Meldet sich im Unterricht“;* *(2) „Hält sich an Gesprächsregeln“;* *(3) „Richtet die Konzentration auf die Bearbeitung der Aufgabe“;* *(4) „Arbeitet ruhig am Platz“;* *(5) „Arbeitet mit“*	(Sauerland 2018, S.80, f.)	Die Items wurden aus einer im Projekt LEVUMI angegliederten Masterarbeit zur Konstruktion eine DBRs übernommen.
Externalisierende Verhaltensprobleme			
Trotzverhalten	(6) „Lässt sich leicht ärgern“; (7) „Weigert sich, Regeln zu befolgen“; (8) „Widersetzt sich den Anweisungen von Autoritätspersonen“; (9) „Ärgert andere absichtlich“	DSM-V (Falkai et al. 2018, S.635)	Die Items wurden der Störung mit oppositionellem Trotzverhalten entnommen.
Unaufmerksamkeit	(10) „Macht Flüchtigkeitsfehler bei den Schularbeiten“; (11) „Hat Schwierigkeiten, längere Zeit die Aufmerksamkeit bei Aufgaben aufrechtzuerhalten“; (12) „Scheint nicht zuzuhören, wenn andere ihn bzw. sie ansprechen“; (13) „Bringt Schularbeiten nicht zu Ende“; (14) „Hat Schwierigkeiten sich zu organisieren“; (15) „Lässt sich durch äußere Reize ablenken“	DSM-V (Falkai et al. 2018, S.78)	Die Items wurden aus den Symptomen des diagnostischen Kriteriums Unaufmerksamkeit der Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung übernommen.
Impulsivität	(16) „Zappelt mit Händen und Füßen“; (17) „Steht in Situationen auf, in denen Sitzenbleiben erwartet wird“; (18) Handelt, als wäre er bzw. sie „getrieben“; (19) „Redet viel“; (20) „Kann in Unterhaltungen nur schwer warten, bis er bzw. sie an der Reihe ist“; (21) „Bringt Schularbeiten nicht zu Ende“	DSM-V (Falkai et al. 2018, S. 77)	Die Items wurden aus den Symptomen des diagnostischen Kriteriums Hyperaktivität und Impulsivität der Aufmerksamkeits- Hyperaktivitätsstörung entnommen.

Internalisierende Verhaltensprobleme			
Emotionalität	(22) „Wirkt oft traurig“; (23) „Zeigt vermindertes Interesse an Aktivitäten“; (24) „Kann sich nur schwer entscheiden“; (25) „Hat Angst vor sozialen Situationen“; (26) „Klagt über körperliche Beschwerden“	DSM-V (Falkai et al. 2018, S. 217)	Die Items wurden aus den Symptomen der diagnostischen Kriterien einer Major Depression der DSM-V entnommen.
Probleme in der Gruppe	(27) „Arbeitet lieber alleine“; (28) „Spielt lieber alleine“; (29) „Wird von Mitschüler_innen gehänselt oder geärgert“	(Sauerland 2018, S. 54)	Die Items wurden aus einer im Projekt LEVUMI angegliederten Masterarbeit zur Konstruktion eine DBRs übernommen.

Tabelle 1: Darstellung der Items und ihrer Herleitung

Die Items wurden in einen Ratingbogen überführt. Wie der Tabelle ersichtlich ist, lassen sich die Verhaltensbereiche Trotzverhalten, Unaufmerksamkeit und Impulsivität den externalisierenden Verhaltensstörungen zuweisen. Dies liegt daran, dass die hier genannten Verhaltensweisen eher durch ausagierendes Verhalten gekennzeichnet werden. Eine Zuordnung von Unaufmerksamkeit kann an dieser Stelle jedoch kritisch betrachtet werden, da die dort genannten Items wie „bringt Schularbeiten nicht zu Ende“ oder „scheint nicht zuzuhören, wenn andere ihn oder sie ansprechen“ nicht unbedingt mit ausagierenden Verhaltensweisen verknüpft sind. Die Verhaltensbereiche Emotionalität und Probleme in der Gruppe sind dagegen internalisierende Verhaltensweisen. Die Items erfragen häufig die soziale Isolation und soziales Rückzugsverhalten, sodass die Symptome keine ausagierenden Verhaltensweisen erfragen.

Abschließend muss die Skalierung des Ratingbogens begründet werden. Wie in Kapitel 3.2 dargelegt wurde, wird für die Verhaltensbewertung eine mindestens sechsstufige Skala empfohlen. Für „PUTSIE“ wird daher eine siebenstufige Likert-Skalierung verwendet. Die Likert-Skala ermöglicht die Skalierung persönlicher Urteile anhand einer mehrstufigen Skala. Einer Aussage wird eine Einschätzung auf einer Skala zugeordnet. Die Beschriftung der Skala in Bezug auf DBRs (zeitlich/prozentual) scheint keinen großen Einfluss auf die Reliabilität zu haben (siehe Kapitel 3.2). Daher werden für die Skalierung als Extremwerte für das Verhalten „Nie“ (1) und „Immer“ (7) angegeben und somit eine zeitliche Skalierung verwendet. Die Skalierung wurde auf

diese Weise dem entwickelten Rating von Sauerland (2018) nachempfunden. Falls das Ratinginstrument von Sauerland und das hier vorgestellte Instrument auf LEVUMI veröffentlicht werden, steigert dies die Vergleichbarkeit und Handhabbarkeit beider Instrumente, da Lehrkräfte beim Wechsel des Ratingbogens keine großen Umstellungen beim Bewerten berücksichtigen müssen. Da die Darstellung des Ratinginstrumentes erfolgt ist, kann nun dargelegt werden, auf welche Art und Weise die Implementierung der Methode an der Schule erfolgen soll.

5.2 Vorgehen und Gesprächsleitfaden

Um Einstellungen und den Umgang mit dem Ratinginstrument „PUTSIE“ aus Sicht von Lehrkräften zu ergründen soll das Instrument in einer Schule implementiert und in regelmäßigen Abständen durch die Lehrkräfte evaluiert werden. Im Sinne empirischer Forschung kommen für die Erhebung von Einstellungen und Erfahrungen nur verbale Daten in Frage, welche in der qualitativen Forschung mittels Erzählung oder Leitfadeninterview gewonnen werden (vgl. Mayer 2013). Sind konkrete Aussagen über einen Gegenstand, wie in diesem Fall die Einstellung und Handhabbarkeit des Ratinginstruments „PUTSIE“ aus Lehrerperspektive das Ziel der Datenerhebung, werden Leitfadeninterviews empfohlen. Diese zeichnen sich durch Gesprächssituationen mit einem Leitfaden aus. Der Leitfaden beinhaltet offen formulierte Fragen, auf die frei geantwortet wird. Auf diese Weise wird die Vergleichbarkeit der Daten erhöht, zudem gewinnt das Vorgehen an Struktur (vgl. ebd.). Da die Lehrkräfte mit dem Direct Behavior Rating „PUTSIE“ arbeiten sollen, das Instrument und die Methode jedoch zuvor nicht angewendet haben, bietet sich zudem eine Implementierung über die Zeit (im Längsschnitt) an. Auf diese Weise können Fehler und Veränderungen in der Handhabung, sowie Veränderungen in der Einstellung zum Instrument erkannt und einbezogen werden. Es wird angenommen, dass zudem Veränderungen im Verhalten einen längerfristigen Prozess darstellen und kurzfristige Erhebungen mit dem Rating womöglich keine Verhaltensänderungen erfassen können. Es wird daher ein achtwöchiger Zeitraum festgehalten.

Durch die Analyse des Schulsystems wird davon ausgegangen, dass die Lehrkräfte zu Beginn der Erhebung noch keine Erfahrung mit Instrumenten der Lernverlaufsdiagnostik gesammelt haben. Aus diesem Grund erschien es sinnvoll, zunächst das Projekt LEVUMI, sowie die Grundlagen von DBRs vorzustellen. Der Autor besuchte die

Schule daher einen Monat vor geplantem Erhebungsbeginn, um das Konzept im Rahmen eines zwanzigminütigen Vortrags innerhalb einer Lehrerkonferenz vorzustellen. Der Vortrag beinhaltete grundsätzliche Überlegungen zu DBR, Informationen zu LEVUMI, sowie eine Darstellung des Ablaufs der Erhebung. Auf diese Weise wurden den Lehrkräften von vornherein die Anforderungen bei Teilnahme an der Erhebung verdeutlicht. Die Informationen sind gesammelt in einem Handout (Anhang B (nur auf der beiliegenden CD)) zu finden, welche den Lehrkräften mit einer ersten Kopie des Ratingbogens „PUTSIE“ ausgeteilt wurden. An der Erhebung interessierte Lehrkräfte konnten sich im Anschluss in eine Liste eintragen.

Für die Erhebung wurde mit der Schulleitung ein etwa achtwöchiger Zeitraum vereinbart, in dem die Lehrkräfte mit dem Instrument arbeiten sollten. Die ersten Gespräche fanden kurz nach den Herbstferien statt, sodass ein Erhebungszeitraum von Ende Oktober bis Mitte Dezember ohne Unterbrechung durch Schulferien, etc. ermöglicht wurde. Innerhalb dieser Zeit sollten je Rating vier leitfadengestützte Interviews als „Planungsgespräche“ geführt werden. Diese bestehen aus Leitfadeninterviews. Durch die Beschreibung als „Gespräche“ soll ersichtlich werden, dass in den Gesprächen keine reine Befragung, sondern auch Erläuterungen zum weiteren Vorgehen sowie Hilfestellungen, welche für die erfolgreiche Implementierung des Instrumentes notwendig sind erfolgen konnten. Jedes der Gespräche sollte, in Abhängigkeit ob ein oder zwei Kinder bewertet wurden 15 Minuten (ein Kind) beziehungsweise 30 Minuten (zwei Kinder) dauern. Begründen lässt sich dies zum einen durch die postulierte Ökonomie, beziehungsweise Effizienz der Methode im Schulalltag (Kapitel 3.1): Lehrkräfte müssen mit der Methode schnell arbeiten können. Zum anderen muss auch an die Ökonomie der Ergebnisauswertung gedacht werden, da die Gespräche im Anschluss durch den Autor transkribiert und inhaltsanalytisch ausgewertet werden müssen.

Vor dem ersten Gespräch erhielten die Lehrkräfte das Material für die gesamte Erhebung in einer Mappe. Diese beinhaltete die Ratingbögen, eine Kurzanleitung sowie Dokumentationsbögen (siehe Anhang C). Die Auswertung und Erstellung der Verlaufsgraphen erfolgt durch den Autor (siehe Kapitel 5.3). Hierzu wurden die Lehrkräfte angehalten, die Ratingbögen am Tag des nächsten Planungsgesprächs im Lehrerzimmer zu hinterlegen. Die Auswertung erfolgte mittels einer Auswertungstabelle in Excel, in der die Ratingwerte der einzelnen Items eingetragen und für die Erstellung

des Gesamtwertes der Verhaltenskategorie summiert wurden (Summenscore). Die Lehrkräfte wurden angehalten, alle zugehörigen Subitems zu einem Verhaltensbereich vollständig anzukreuzen. Sie durften frei wählen, wie viele und welche Verhaltensbereiche sie bei den jeweiligen Schülern bewerten wollten. Eingeschränkt wurde die Wahl der Beobachtungszeitpunkte. Es wurde darauf geachtet, dass die Lehrkraft das Verhalten lediglich einmal pro Tag einschätzte. Dies erleichtert im Nachhinein die Vergleichbarkeit der erhobenen Daten. Zudem wurde der Beobachtungszeitraum auf maximal einen Schultag festgelegt, damit regelmäßige Ratingsituationen pro Tag zustande kommen.

Grob wurde folgender Ablauf der Erhebung verfolgt:

1. Erstes Gespräch, gefolgt von einer einwöchigen „Baselineerhebung“

2. Zweites Gespräch, gefolgt von einer zweiwöchigen Interventionsphase

3. drittes Gespräch, gefolgt von einer zweiwöchigen Interventionsphase

4. Abschließendes viertes Gespräch

Da Verhaltensveränderungen häufig länger Zeit benötigen wurde die zweiwöchige Interventionsphase gewählt. Die Inhalte der einzelnen Gespräche sollen im Folgenden durch die Darlegung des Gesprächsleitfadens genauer vorgestellt werden. Der Leitfaden, wie er für die Erhebung genutzt wurde, ist im Anhang (Anhang D) zu finden.

Gespräch 1: Ziel des ersten Gesprächs und erste Aufgabe bei der Verwendung der Methode ist es, die Kinder zu bestimmen, welche durch das Rating erfasst werden sollen, sowie ihre Problemlagen genauer zu beschreiben, um darauf aufbauend Verhaltensbereiche des Ratingbogens zuzuordnen, sowie die Zeiträume beziehungsweise Situationen zu bestimmen in denen das Verhalten bewertet werden soll. Hierzu werden folgende Fragen des Leitfadens in Anlehnung an Brieschs (2016) Schrittfolge mit dem „Individualized Behavior Rating Scale Tool“ gestellt:

Leitfaden Gespräch 1
1. Welchen Schüler möchten Sie mit dem Rating bewerten? (Name, Geschlecht, Alter, Migrationshintergrund, familiärer Hintergrund)
2. Warum möchten Sie gerade diese_n Schüler_In beobachten? (Verhaltensprobleme, wann und wo?)

3. Sie sehen innerhalb des Fragebogens verschiedene Beobachtungsbereiche (grau hinterlegt), welche dieser Bereiche könnte Ihrer Meinung nach für Sie am relevantesten/interessantesten sein?

4. Wann tritt das Verhalten auf, in welchen Situationen und über welche Zeiträume möchten Sie das Verhalten einschätzen/bewerten (maximales Beobachtungsintervall: ein Schultag, minimales Beobachtungsintervall: eine Schulstunde).

5. Wer von Ihnen bewertet das Verhalten/bewerten Sie das Verhalten alleine?

6. Haben Sie noch Fragen?

7. Was versprechen Sie sich von der Arbeit mit dem Direct Behavior Rating?

Abb. 2 „Leitfaden Gespräch 1“

Durch die Fragen erfolgen zum einen für Direct Behavior Ratings notwendige grundsätzliche Festlegungen von Beobachtungszeiträumen, Beobachtungsbereichen sowie des Raters. Zusätzlich wird insbesondere durch Frage 7 auch die Einstellung der Lehrkräfte zu Beginn der Methode erfasst wird.

In Bezug auf die inhaltliche Ausrichtung des ersten Gesprächs ist wichtig, dass an dieser Stelle noch keine Festlegung einer Intervention erfolgt, sondern zunächst eine sogenannte Baseline erhoben wird. Dies wird ermöglicht, indem das gewählte Verhalten mittels des Ratingbogens erhoben, jedoch noch keine weitere Intervention angeschlossen wird. Die so erhobenen Daten werden verwendet, um zu untersuchen, ob das Verhalten tatsächlich beobachtbar und eine Intervention vertretbar ist und um Vergleichswerte für spätere Interventionsphasen zur Verfügung zu haben, auf die man sich beziehen kann (vgl. Briesch 2016). Die Autorin empfiehlt die Baselineerhebung über eine Woche, wobei dies einen Richtwert darstellt, welcher empirisch nicht nachgewiesen werden konnte. In Anlehnung an die Ergebnisse von Huber und Rietz (2015), welche herausarbeiteten, dass für reliable Ergebnisse drei bis fünf Erhebungszeitpunkte notwendig sind, wurde für die Baselineerhebung der Zeitraum einer Woche festgelegt. Die Festlegung von mehr Erhebungszeitpunkten als unbedingt notwendig ermöglicht dabei die Freiheit, trotz kurzem Fehlen von Lehrkräften oder Schülerinnen und Schülern die Erhebung dennoch im geplanten Zeitraum mit reliablen Ergebnissen weiterzuführen.

Gespräch 2: Die Evaluation der gewählten „Verhaltensbereiche“, der Ratingzeitpunkte sowie einer „Förderhypothese“, welche das Förderziel und die Intervention

zum Erreichen des Ziels und die Festlegung der Verantwortlichkeiten zwischen Klassenlehrkraft und gegebenenfalls sonderpädagogischer Lehrkraft enthält, waren Ziele des zweiten Gesprächs.

Leitfaden Gespräch 2:

1. Kamen Sie mit dem Erhebungsinstrument (Ratingbögen) zurecht, gab es Besonderheiten während der „Baselineerhebung". Ergaben sich Fragen?

2. Sie sehen hier (grafisch) die Ergebnisse ihres_r Schülers_In, was können Sie anhand ihrer Ergebnisse erkennen (das Verhalten ist gleichgeblieben, schwankt stark)

3. Es ist Ziel, nun ein Förderziel und eine Fördermaßnahme festzulegen, welche innerhalb der nächsten zwei Wochen von Ihnen durchgeführt wird. Wie könnte Ihrer Meinung nach ein solches Ziel mit einer verknüpften Maßnahme aussehen. (Hier kann ggf. Unterstützung durch das Werk „Schwierige Schüler- 49 Handlungsmöglichkeiten bei Verhaltensauffälligkeiten (Hartke, Vrban (2012))" erfolgen).

4. Wer führt die Fördermaßnahme durch, wer übernimmt das Rating?

5. Haben Sie noch Fragen?

Abb. 3 „Leitfaden Gespräch 2"

Durch Frage 1 lassen sich Probleme der Lehrkräfte in den Bewertungssituationen oder mit dem Ratingbogen erfassen. Frage zwei fokussiert die Analysekompetenz der Lehrkräfte bei der Auswertung von Graphen, welche in der Theorie häufig als nicht ausreichend vorhanden gilt. Den Lehrkräften wird hierzu der Graph auf dem Computer präsentiert, welcher zuvor vom Autor mittels der ausgefüllten Bögen erstellt wurde. Die Fragen 1 und 2 werden den Gesprächen 2 bis 4 vorangestellt, um gegebenenfalls Veränderungen über die Zeit festzustellen. Bei Frage 2 können Hinweise und Hilfestellungen durch den Interviewer erfolgen, um Fehlinterpretationen zu korrigieren. Frage 3 soll genauer erläutert werden. Es ist Anliegen der Arbeit, Bedingungen zu schaffen, welche dem RTI-Modell gleichen. Interventionen sollen demnach evidenzbasiert sein. Ein Buch, welches übersichtlich evidenzbasierte Verfahren für den Bereich Verhalten vorstellt ist das Buch „Schwierige Schüler – 49 Handlungsmöglichkeiten bei Verhaltensauffälligkeiten" (Hartke und Vrban 2010). Das Werk wird so eingebunden, dass entweder die Interventionen der Lehrkräfte mit den Methoden abgeglichen und dort einsortiert werden, oder auf Wunsch der Lehrkräfte eine Vorstellung geeigneter Verfahren für das Problemverhalten in Gespräch 2 durch den Interviewer erfolgte, sodass

die Lehrkräfte aus einer kleinen Anzahl an Interventionen wählen konnten. Zu betonen ist, dass die Durchführung der Förderung alleinig bei den Lehrkräften lag, und diese als Experten für die Förderung der Schülerinnen und Schüler gelten. Es fand keine Durchführung von Fördermaßnahmen durch den Interviewer statt.

Gespräch 3: Die Ziele dieses Gespräches sind insbesondere die Evaluation der ersten Interventions- oder Förderphase anhand der gewonnenen Graphen, um Rückschlüsse über die Intervention und das sinnvolle weitere Vorgehen zu erhalten. Dieses Vorgehen lässt sich auch mit dem Begriff des Data-based-decision-makings anhand von Direct Behavior Ratings umschreiben. Zunächst wird der Graph ausgewertet und beschrieben. Anschließend wird die Frage gestellt, inwiefern gewünschter Erfolg eingetreten ist. Falls nicht soll untersucht werden, ob die Förderung korrekt implementiert wurde, falls Ja, kann die derzeitige Förderung geändert oder eine neue Intervention gewählt werden, falls nicht können Strategien entwickelt werden, die die richtige Umsetzung der Intervention gewährleisten. Ist ein Erfolg eingetreten, kann die Intervention fortgesetzt werden, ohne sie zu ändern, Modifizierungen geleistet werden, um die Durchführbarkeit zu erhöhen oder Strategien zur Generalisierung des Zielverhaltens besprochen werden (vgl. Briesch 2016, S. 230). Die Umsetzung dieses „decision-making-trees“ (ebd.) soll sich auch in den Fragen des Leitfadens wiederspiegeln.

Leitfragen Gespräch 3:

1. Kamen Sie mit dem Rating zurecht, gab es Probleme, Fragen oder andere Besonderheiten während der ersten Interventionsphase?

2. Sie sehen hier (grafisch) die Ergebnisse der ersten Interventionsphase. Wie interpretieren Sie diese Ergebnisse?

3. Wie möchten Sie in Hinblick auf die weitere Förderung fortfahren?

Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf die Fördermethode?

- *Bei keiner Veränderung:* Gab es Probleme bei der Implementierung der Methode? Muss die Intensität der Intervention minimiert/maximiert werden? Soll die Methode geändert werden?
- *Bei positiver Veränderung*: Soll die vorhandene Methode weiterverfolgt/intensiviert werden, oder im Sinne eines fading-outs ausgeblendet werden?

(Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf den Beobachtungszeitraum?)

(Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf die gewählten Beobachtungsbereiche?)

4. Haben Sie noch Fragen?

Abb. 4 „Leitfaden Gespräch 3“

Frage 3 greift den decision-making tree nach Briesch (2016) genauer auf, indem Nachfragen zum weiteren Vorgehen in Abhängigkeit der Graphenergebnisse gestellt werden. Das Vorgehen ist zudem in der Kurzanleitung im Ratingmaterial der Lehrkräfte erläutert (Anhang C). Im Anschluss an das Gespräch findet die zweite zehntägige Interventionsphase statt.

Gespräch 4: Das letzte Gespräch der Erhebung dient der Evaluation der zweiten Förderphase (Grapheninterpretation und mögliche Änderungen in der Intervention), zum anderen jedoch auch der abschließenden Einschätzung der Methode durch die Lehrkräfte.

Leitfragen Gespräch 4:

1. Kamen Sie mit dem Rating zurecht, gab es Besonderheiten während der zweiten Interventionsphase?

2. Sie sehen hier (grafisch) die Ergebnisse der Interventionsphase. Wie interpretieren Sie den Graphen – auch im Vergleich zu den vorherigen Ratings?

3. Wie möchten Sie in Hinblick auf die weitere Förderung fortfahren?

Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf die Fördermethode?

- Bei keiner Veränderung: Gab es Probleme bei der Implementierung der Methode? Muss die Intensität der Intervention minimiert/maximiert werden? Soll die Methode geändert werden?
- Bei positiver Veränderung: Soll die vorhandene Methode weiterverfolgt/intensiviert werden, oder im Sinne eines „fading-outs“ ausgeblendet werden?

Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf den Beobachtungszeitraum?

Welche Folgerungen ergeben sich aus Ihrer Sicht in Hinblick auf die gewählten Beobachtungsbereiche?

4. Welche Möglichkeiten und Grenzen birgt das Ratinginstrument Ihrer Meinung nach in der Arbeit im inklusiven Setting und wie würden Sie es in Ihrer weiteren Arbeit einbinden?

5. Können Sie sich vorstellen, die Methode und die Ergebnisse für die Kommunikation mit außerschulischen Partnern oder auch Eltern zu verwenden?
6. Ergaben sich Probleme/Fragestellungen im Umgang mit dem Instrument, die für Sie von vorneherein besser oder anders hätten vermittelt werden müssen?

Abb. 5 „Leitfaden Gespräch 4“

Der Leitfaden für das vierte Gespräch wiederholt die Fragen aus Gespräch 3. Es schließen sich jedoch ab Frage 4 abschließende, allgemeine Fragen zur Arbeit mit dem Instrument an. So wird nach wahrgenommenen Möglichkeiten und Grenzen gefragt, sowie ob und wie die Methode weiterhin verwendet werden könne. Zudem wird explizit der Aspekt Kommunikation mit Hilfe der Methode hervorgehoben (Frage 5), welcher in Kapitel 3 als ein Vorteil dargestellt wurde, jedoch von den Lehrkräften nicht während der Erhebungszeit angewendet wurde. Abschließend werden Probleme oder Fragestellungen im Umgang mit dem Instrument erhoben. Die empirische Erhebung der Gesprächsdaten endet mit dem vierten Interview.

Insgesamt soll die Vorgehensweise mit Leitfadengestützten Planungsgesprächen eine Implementierung der Methode ermöglichen, die Unterstützung der Lehrkräfte im Umgang mit dem Instrument nicht ausschließt und durch den gleichen Aufbau der Gespräche dennoch zu Vergleichbarkeit der Daten führt. Bisher unbeachtet ist die Auswertung der Verlaufsgraphen für die einzelnen Planungsgespräche. Diese soll im Folgenden dargelegt werden.

5.3 Die Graphenerstellung

Wie in Kapitel 5.2 dargelegt, erfolgte die graphische Darstellung der Ratingergebnisse durch den Verfasser. Es muss daher dargelegt werden, wie die Graphen für die Lehrkräfte aufbereitet und diesen präsentiert werden.

Die Datenaufbereitung wurde an Brieschs „Steps in Preparing DBR Data for Interpretation and Decision Making“ (ebd. 2016, S. 215) angelehnt. Briesch formuliert sechs Schritte, die für die Datenaufbereitung von Direct Behavior Ratings berücksichtigt werden sollen: Die Auswahl eines Daten-Management-Systems, die Dateneingabe, die Generierung von zusammenfassenden Statistiken, das Erstellen von Balkendiagrammen, das Erstellen von Liniendiagrammen. (vgl. ebd.)

1. Die Auswahl des Daten-Management-Systems: Die Auswahl eines Instruments zum Datenmanagement fällt auf Microsoft-Excel. Dies hat vor allem den Grund, dass dem Autor das Programm zur Verfügung steht und die Lehrkräfte auf Wunsch schon vor der Implementierung des Ratings auf der Onlineplattform LEVUMI mit dem Ratingbogen auf ihren Schulrechnern weiterarbeiten könnten, da sie mit dem Programm vertraut sind und dieses zum Großteil besitzen. Auswertungsvorlagen können zudem leicht weitergereicht werden und sind so ohne große statistische Kenntnisse und Einarbeitungszeiten verwendbar. Auch Briesch (2016) verweist auf das Programm und dessen Eignung zur Datenaufbereitung. Wichtig ist schon an dieser Stelle die Berücksichtigung des Datenschutzes. Daher wurden keine Schülernamen verwendet, sondern lediglich Buchstaben zur Beschreibung der Schüler.

2. Dateneingabe: An dieser Stelle soll die Datenaufbereitung für die Erstellung kurz dargelegt werden. Wie schon im vorherigen Kapitel dargelegt, werden die ausgefüllten Ratingbögen am Tag des Planungsgesprächs im Lehrerzimmer hinterlegt. Nun gilt es die Ratingergebnisse so in die Datenmaske zu überführen, dass anschließend ein Verlaufsgraph daraus gewonnen werden kann. Je beobachtetem Kind wurde in Excel ein Datenblatt erstellt. Jede Spalte des Datenblattes repräsentiert zunächst ein Item. Jede Zeile einen Beobachtungszeitpunkt. Den einzelnen Items werden der zugehörige Verhaltensbereich sowie die Zugehörigkeit zu externalisierender/internalisierender Verhaltensstörung zugewiesen. Beginnt ein neuer Verhaltensbereich wird zudem eine Spalte „Gesamt“ angefügt, in der der Summenscore der dem Verhaltensbereich zugehörigen Items errechnet wird. Gleiches geschieht für die Bereiche des externalisierenden/internalisierenden Verhaltens, in dem die Summenscores der ihnen untergeordneten Verhaltensbereiche in jeweils einer Spalte errechnet werden. Bei vollständiger Ausfüllung der untergeordneten Verhaltensbereiche können so auch Summenscores zu diesen Verhaltenskategorien gewonnen werden. Die ausgefüllten Datenblätter befinden sich auf der beiliegenden CD (Anhang E).

Die Daten der Ratingbögen werden überführt, indem die angekreuzten Werte „1“ bis „7“ auf den Likert-Skalen der einzelnen Items in der passenden Spalte des Datenblattes eingetragen werden. Handelt es sich um den ersten Wert, wird die erste Zeile der dem Item zugeordneten Spalte genutzt. Handelt es sich um den zweiten Wert, die zweite und so weiter.

Für die Erhebung wurde entschieden, dass fehlende Werte gedoppelt werden können und diese bei der graphischen Darstellung in gepunktete Graphen umgewandelt werden, um das Fehlen der Werte zu kennzeichnen.

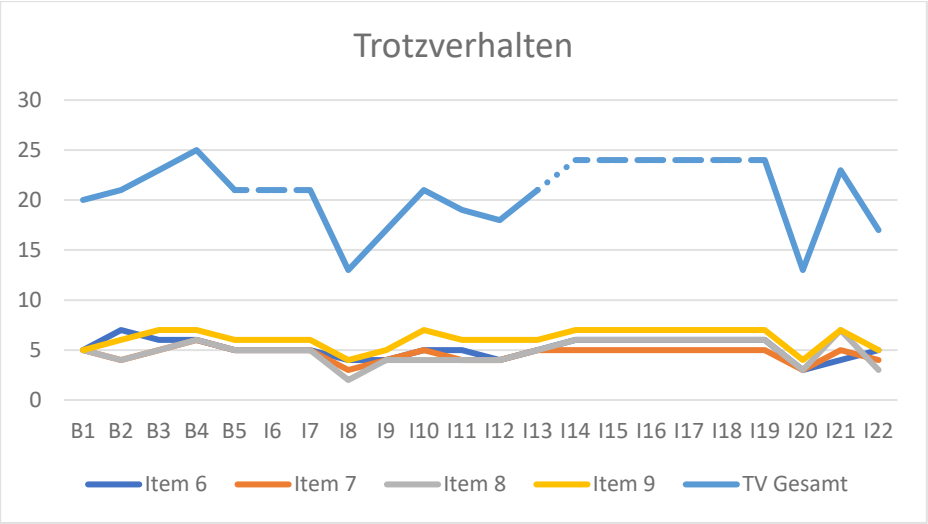
3. Das Generieren zusammenfassender Statistiken: Wie schon im obigen Abschnitt dargestellt, werden die Beobachtungswerte einzelner Verhaltensbereiche je Beobachtungszeitpunkt durch Addieren der einzelnen Itemwerte zusammengefasst. Auf diese Weise lassen sich Summenwerte für die einzelnen Verhaltensbereiche generieren. Es besteht die Möglichkeit die Werte als Summenscore oder als Median zusammenzufassen. Für die vorliegende Arbeit werden Summenscores für die Auswertung der Ratings verwendet. Hierbei zeigt sich der Nachteil, dass je nach Anzahl der Items, welche einem Verhaltensbereich zugeordnet werden, unterschiedlich hohe Werte zugeordnet werden können. So kann bei einer Multi-Item-Skala mit 5 Items und einer siebenstufigen Likert-Skala eine maximale Punktzahl von 35 Punkten erlangt werden, bei einer Multi-Item-Skala mit drei Items nur 21 Punkte. Andererseits ist die Verwendung von Summenscores bei der Verwendung von DBRs (Casale et al. 2015b, S. 48) üblich, weshalb die Verwendung von Summenscores für die Auswertung am einfachsten erscheint.

4. Erstellung von Balkendiagrammen. Die Erstellung von Balkendiagrammen wird von Briesch (2016) für die Darstellung von Veränderungen ganzer Gruppen über die Zeit genutzt. In dieser Arbeit kann diese Art der Auswertung jedoch unberücksichtigt bleiben, da die Erhebung sich auf die Einzelförderung mittels Direct Behavior Ratings beschränkt. Es kann jedoch erwähnt werden, dass Balkendiagramme die Darstellungen von Gruppenentwicklungen über die Zeit und die Gegenüberstellung verschiedener Gruppen innerhalb der Klasse (zum Beispiel Mädchen und Jungen) ermöglichen (vgl. ebd.).

5. Erstellung von Liniendiagrammen. Abbildung 2 zeigt beispielhaft die Graphendarstellung zum Verhaltensbereich „Trotzverhalten“, wie er aus den Daten gewonnen wurde und den Lehrkräften zur Auswertung vorgelegt wurde. Auf der X-Achse wird, analog zu nahezu jeder Verlaufsdiagnostik die Zeit repräsentiert. Da die Lehrkräfte angewiesen wurden, möglichst täglich das Verhalten zu bewerten entspricht eine Einheit auf der X-Achse einem Tag. Die einzelnen Zeitpunkte werden zudem so benannt, dass erkennbar ist, in welcher Phase (Baselineerhebung/Interventionsphase) der Datenpunkt erhoben wurde. Auf diese Weise wird die Verwendung von Phase Change

Lines, senkrechten Strichen, die markieren, ab wann eine Intervention verändert, hinzugefügt oder entfernt wird, umgangen, wobei diese bei einigen Graphen durch gepunktete Linien dennoch dargelegt wurden (vgl. Briesch 2016). An dieser Stelle tritt zudem die zu klärende Frage im Umgang mit fehlenden Werten auf. Briesch verweist hierzu lediglich auf die Frage, ob eine graphische Darstellung fehlender Werte für die Fragestellung relevant ist, falls ja sollte man fehlende Werte kenntlich machen, falls nicht, auslassen. (vgl. ebd.) Für die vorliegende Arbeit wird gefolgert, dass wenn möglich fehlende Werte als solche durch den Graphen gekennzeichnet werden sollen, da so mögliche Sprünge im Verhalten durch Krankheit, Fehlzeiten, nicht stattfindende Interventionen erkennbar und bei der Graphenauswertung berücksichtigt werden können. Es wird festgelegt, dass fehlende Werte -wenn möglich- in Schritt 3 gedoppelt werden sollen und anschließend im Graphen durch einen gestrichelten Graphenverlauf gekennzeichnet werden. Die Y-Achse repräsentiert die Punktewerte, die erlangt wurden. Möglichst sollen Minimal- und Maximalwerte auf der Y-Achse angegeben werden. Hier zeigte sich die Erstellung von Graphen in Excel mit Hilfe von Summenscores als schwierig. Die Graphenerstellung wäre mit einem erheblichen zeitlichen Mehraufwand verbunden gewesen, hätten die Maximalwerte der einzelnen Multi-Item-Skalen auf der Y-Achse angezeigt werden sollen. Es wurde schließlich darauf verzichtet, diese graphisch darzustellen. Im Bereich der Auswertung der Graphen durch die Lehrkräfte ist das Fehlen der Maximalwerte auf der Y-Achse jedoch zu berücksichtigen.

Bei der Darstellung der Graphen wurde immer der Gesamtgraph eines Verhaltensbereichs dargelegt, darunter jedoch auch die Graphen der angekreuzten Subitems. Dem Autor ist bewusst, dass die einzelnen Items eventuell als Single-Item-Skalen wahrgenommen werden, nichtsdestotrotz bietet die Darstellung gegebenenfalls Anhaltspunkte über Fördererfolge innerhalb einzelner Verhaltensbereiche, die ansonsten unberücksichtigt blieben. Ein solches Vorgehen bietet jedoch kritisches Potenzial, da so nicht mehr auf eine reine Multi-Item-Skala verwiesen wird.



(Abb. 6: „Graphische Darstellung „Trotzverhalten“ nach 22 möglichen Messzeitpunkten (B=Baselineerhebung, I=Interventionsphase“)

5.4 Verschriftlichung der Interviewdaten

Wie schon dargelegt, erfolgt die Erhebung der Lehreraussagen über leitfadengestützte Interviews, beziehungsweise „Planungsgespräche“. Der Grundgedanke ist, dass sich im Rahmen dieser Gespräche das zu beobachtende Phänomen – in diesem Fall die Handhabbarkeit oder Probleme im Umgang und Einstellungen der Lehrkräfte zum DBR „PUTSIE“ – niederschlagen. (vgl. Fuß und Karbach 2014) Die Gespräche werden daher mittels eines Audioaufnahmegeräts aufgenommen (Anhang H auf der beigefügten CD). Die Lehrkräfte wurden hierüber informiert und aufgeklärt. Um die Gesprächsdaten verarbeiten zu können, ist eine Transkription der Gespräche notwendig. Wissenschaftliche Transkripte geben Wort für Wort sämtliche Inhalte eines Gesprächs wieder. Der dramaturgische Aufbau einer Gesprächssituation wird sichtbar. Gedankensprünge und nebensächliche Aussagen werden mittranskribiert, um den Gesprächsverlauf vollständig erfassen zu können. Es finden keine Auslassungen und Zusammenfassungen von Passagen seitens des Transkribierenden statt. In welcher Detailgenauigkeit und Art der Sprechakt transkribiert wird, wird durch Transkriptionsregeln festgelegt. Transkriptionen ermöglichen eine wissenschaftliche Analyse von Gesprächen, indem die Gesprächssituation in schriftliche Daten transformiert wird (vgl. ebd.). Die Ausgestaltung von Transkripten kann dabei sehr unterschiedlich sein und ist stark abhängig von Auswertungsfokus und dem Detaillierungsgrad sowie den angewendeten Notationszeichen. Es lassen sich „Module“ unterscheiden, für die es bei der schriftlichen Transkription Regeln oder deren Berücksichtigung festzulegen gilt (vgl. Fuß und

Karbach 2014). So ist zu entscheiden, inwieweit die Module Sprachglättung (die Umwandlung von Dialekt in Hochdeutsch), Pausen, Sprachklang, Lautäußerungen, Wortabbrüche, Verschleifungen, nicht-sprachliche Ereignisse, Interaktion (gleichzeitiges Sprechen, Sprechunterstützung), Unsicherheit, Unterbrechungen, Auslassungen und Zeichensetzung berücksichtigt werden sollen. Die Autoren verweisen beispielsweise darauf, dass insbesondere bei inhaltlichem Fokus der Transkription, Regeln und Module der Transkription reduziert werden können (vgl. ebd.).

Um nicht immer wieder jede Regel neu festlegen zu müssen, haben Wissenschaftler für unterschiedliche Anlässe Transkriptionsregeln verfasst, welche sich als solche übernehmen lassen. So gibt es beispielsweise das Grundtranskript und das Regelsystem nach Kuckartz für vorrangig inhaltliche Fragestellungen, wie sie dieser Arbeit zugrunde liegen (vgl. Fuß und Karbach 2014). Die Transkription der erhobenen Audiodaten soll in Anlehnung an diese Regelsysteme geschehen. Im Folgenden werden daher die einzelnen Module (siehe oben) und deren Berücksichtigung für diese Arbeit dargelegt.

Sprachglättung: Die gesprochene Sprache wird in Standardsprache beziehungsweise Schriftsprache überführt. Dabei wird jedes Wort ins Hochdeutsche übertragen und verschriftlicht, wobei syntaktische Fehler im Text beibehalten werden.

Pausen: Pausen sollen nach Kuckartz und dem Grundtranskript zwar berücksichtigt werden, bleiben aber durch den inhaltlichen Fokus der Fragestellung bei der Transkription unberücksichtigt und werden nicht mittranskribiert, da diese später in der Datenanalyse unberücksichtigt bleiben. Dies hat vor allem ökonomische Gründe beim transkribieren.

Sprachklang: Hinsichtlich des Sprachklangs müssten laut Grundtranskript Auffälligkeiten in der Gesprächslautstärke markiert werden. Auch diese Markierung bleibt im Transkript unberücksichtigt, da eine detaillierte Auswertung der Sprechweise in der Analyse nicht verfolgt werden soll.

Lautäußerungen: Zuhörersignale des Interviewers müssten nach Kuckartz (2014) und dem Grundtranskript nicht berücksichtigt werden. Da Zuhörersignale jedoch Zustimmung oder Ablehnung zu Äußerungen beinhalten und somit inhaltlich relevant sein können sollen diese berücksichtigt werden. So wird insbesondere „Mhm.“(zustimmend/ablehnend) bei der Transkription berücksichtigt.

Wortabbrüche: Wortabbrüche werden in Anlehnung an Kuckartz und das Grundtranskript mittranskribiert.

Verschleifungen: Verschleifungen können in Anlehnung an beide Transkriptionsregularien unberücksichtigt bleiben.

Nicht-sprachliche Ereignisse: Nichtsprachliche Ereignisse, die zu einer Unterbrechung der Gesprächssituation führen werden in Anlehnung an das Grundtranskript berücksichtigt.

Interaktion: Interaktion, insbesondere gleichzeitiges Sprechen und Sprechunterstützung muss in Anlehnung an beide Transkriptionsregularien nicht berücksichtigt werden. Dies ermöglicht auch die Vernachlässigung der sonst für Transkriptionen häufig üblichen „Partiturschreibweise".

Unsicherheit: Unsicherheit wird von Seiten der beiden berücksichtigten Regelsysteme nicht als Modul aufgegriffen und bleibt daher unberücksichtigt.

Unterbrechungen: Unterbrechungen können ebenfalls unberücksichtigt bleiben, da sonst die zuvor erwähnte „Partiturschreibweise" Verwendung finden müsste, welche mit einem zeitlichen Mehraufwand verbunden wäre. Abbrüche sind in den Gesprächen dennoch erkennbar.

Auslassungen: Treten Auslassungen von Wörtern auf, werden diese ähnlich wie bei Wortabbrüchen beibehalten und dementsprechend transkribiert.

Zeichensetzung: Die Zeichensetzung orientiert sich an den deutschen Rechtschreibregeln und wird zugunsten besserer Lesbarkeit beibehalten. Da gesprochene Sprache sich jedoch Schriftsprache unterscheidet, werden auf diese Weise entstehende syntaktische Fehler beibehalten.

Die oben dargelegten Regeln stellen die für die Aufbereitung der Audioaufnahmen notwendigen Transkriptionsregeln dar. Des Weiteren wird darauf verwiesen, dass jegliche personenbezogenen Daten in den Transkripten anonymisiert wurden. Hierbei wurden die Schülernamen der einfachen Lesbarkeit halber in andere Vornamen umgeändert. Den Lehrkräften wurden die Kürzel „L1" bis „L9" zugewiesen, der Autor und Leiter der Gespräche erhält das Kürzel „M" für „Moderator". Die Transkripte werden so sortiert, dass je beobachtetem Kind Kind eine Transkriptnummer von 1 bis 11 zu-

gewiesen wird. Da je Schüler vier Gespräche geführt wurden, werden die unterschiedlichen Gespräche durch nachgeordnete Nummern X.1 (erstes Gespräch) bis X.4 (viertes Gespräch) kenntlich gemacht. Den einzelnen Transkripten wurden der Vollständigkeit halber zusätzlich die dem jeweiligen Gespräch zugrundeliegenden Verlaufsgraphen angefügt. Die anonymisierten Transkripte und Verlaufsgraphen sind in Anhang F hinterlegt.

Da das Textmaterial nun aufbereitet wurde, muss dargelegt werden, wie mit den gewonnenen Transkripten verfahren und diese ausgewertet werden sollen.

5.5 Transkriptauswertung

Durch die Fragestellung, ob das DBR „PUTSIE“ für die Lehrkräfte handhabbar erscheint und welche Einstellungen sie gegenüber dem Instrument zeigen, ist eine inhaltliche Analyse der Transkripte notwendig. Sollen Textmengen, beziehungsweise Transkripte ausgewertet werden, bieten sich hierzu inhaltsanalytische Methoden zu deren Bearbeitung an. Grundannahme dieser ist, dass in Texten kulturelle Formen und Ausdrucksweisen ausgedrückt werden können und die so erhobenen Daten mit der sozialen Realität korrespondieren (vgl. Bos und Tarnai 1999). Die Inhaltsanalyse „ist eine empirische Methode zur systematischen, intersubjektiv nachvollziehbaren Beschreibung inhaltlicher und formaler Merkmale von Mitteilungen, meist mit dem Ziel einer darauf gestützten interpretativen Inferenz auf mitteilungsexterne Sachverhalte.“ (Früh 2017, S. 29) Durch die Definition wird klar, dass sich die Inhaltsanalyse für die inhaltliche Analyse von Textmengen, wie sie für die Auswertung der Transkripte notwendig ist, eignet.

Inhaltsanalysen zeigen methodisch einige Unterschiede, lassen sich aber auf einige Gemeinsamkeiten reduzieren: Der Sinn jeder Inhaltsanalyse besteht darin, die Komplexität des Materials durch analytische Methoden zu reduzieren. Dies geschieht, indem Textmengen hinsichtlich interessierender Merkmale, die in Klassifikationssystemen beschrieben werden, analysiert werden. Diese Reduktion führt dazu, dass nicht notwendige Informationen verloren gehen. Nach Kriterien werden die übrig gebliebenen Textmerkmale in den Kategoriensystemen festgelegten Merkmalsklassen oder Merkmalstypen zugeordnet (vgl. ebd.).

Die klassisch-hermeneutische Inhaltsanalyse betrachtet textliche Inhalte hierbei streng getrennt voneinander. Dies nimmt die Möglichkeit, Ergebnisse intersubjektiv darzustellen, da jeweils nur ein Text oder Transkript betrachtet werden darf (vgl. Bos und Tarnai 1999). Ein Vorgehen für die Fragestellung dieser Arbeit ist demnach schwierig, da möglichst Gemeinsamkeiten und Unterschiede zwischen den Transkripten dargelegt werden sollen, um die Aussagen der Lehrkräfte miteinander vergleichen zu können.

Die quantifizierende Inhaltsanalyse hingegen ermöglicht den Vergleich von Aussagen verschiedener Quellen, indem sie die Frequenz bestimmter textlicher Einheiten auszählt und so quantifiziert. Auftretenshäufigkeiten textlicher Merkmale können auf diese Weise verglichen werden. Hierzu ist die Schaffung eines Kategoriensystems im Vorhinein wichtig, welches notwendige Analysekategorien in Hypothesenform aus der Theorie herleitet (vgl. ebd.).

Früh (2017) führt in seinem Ansatz der integrativen Inhaltsanalyse Aspekte der qualitativen Inhaltsanalyse mit Aspekten der quantitativen Inhaltsanalyse zusammen. Zu Beginn der Analyse steht nicht die Frage, „was steht in den Texten?“, sondern „sind die Merkmale X, Y, und Z in bestimmten Textmengen vorhanden?“ und „in welchem Umfang, welcher Verteilung“ liegen diese vor. Dies führt dazu, dass keine offene Forschungsfrage, sondern überprüfbare Hypothesen formuliert werden müssen, welche aus der Theorie oder, und hier kommt der qualitative Aspekt ins Spiel, aus dem Material selbst hergeleitet werden können (vgl. ebd.).

Methodisch sieht das Vorgehen nach Früh folgendermaßen aus: Aus den Inhalten der Forschungsfrage werden Hauptkategorien extrahiert, das heißt für jeden theoretischen Hauptaspekt der mit der Forschungsfrage verbunden ist, wird eine Hauptkategorie generiert. Oft lassen sich dabei schon einige evidente Unterkategorien festlegen. Anschließend wird durch eine induktive Analyse jede Hauptkategorie systematisch in Unterkategorien aufgeteilt. Das Ausmaß dieser Diversifizierung ist durch die Zahl der in den Texten gefundenen und gemäß der Forschungsfrage interessierenden Inhaltsaspekte limitiert: Nicht alles ist wichtig, und vieles muss nicht in alle denkbaren Unteraspekte ausdifferenziert werden (hierzu kann die Voranalyse einer kleinen Stichprobe des Untersuchungsmaterials genutzt werden) (vgl. ebd.).

Wichtig ist, dass die Bedeutung einer jeden Unterkategorie durch eine Definition festgelegt wird, sodass die Frage „steht XY sinngemäß da?“ beantwortet werden kann.

Deshalb muss die Definition aufzeigen, welche Mitteilungsaspekte wie interpretiert werden sollen, hierzu können zudem Indikatoren herangezogen werden. Um zu untersuchen, ob ein Kategoriensystem tragfähig ist, also ob die Definitionen eindeutig sind, lässt sich eine Probecodierung und ein Reliabilitätstest durchführen, bei dem jeweils untersucht wird, inwiefern Textstücke der Textmenge von unterschiedlichen Codierern gleichen Kategorien zugeordnet werden oder nicht (vgl. Früh 2017).

Ist ein reliables Kategoriensystem vorhanden, kann die Codierung beginnen. Im besten Fall analysieren mehrere Codierer das Material und weisen dem Kategoriensystem entsprechend zugehörige Textabschnitte zu den Unterkategorien zu. Im Anschluss hieran können die Ergebnisse gemessen, also quantifiziert werden, indem zum Beispiel die Anzahl der Codierungen je Kategorien gezählt wird (vgl. ebd.).

5.5.1 Darlegung des Kategoriensystems

In Bezug auf die Forschungsfragen zur Implementierung des Direct Behavior Rating „PUTSIE“ gilt es demnach die Forschungsfragen in ein Kategoriensystem zu überführen. Hierzu können die in Kapitel 3.2 dargelegten bisherigen Forschungsergebnisse zur Handhabbarkeit von Direct Behavior Ratings dienen.

Grundsätzlich werden durch die Forschungsfragen (Kapitel 3.3) Fragen nach der Handhabbarkeit und den Einstellungen der Lehrkräfte zu dem Ratinginstrument „PUTSIE“ gestellt. Diese sind nicht für die Auswertung aus inhaltsanalytischer Methode geeignet, da die Quantifizierung von Aussagen, wie erläutert, nur mittels Hypothesen gelingen kann. Daher müssen in einer Planungsphase aus der Theorie Annahmen gewonnen werden, die sich direkt auf die Anwendbarkeit oder Nichtanwendbarkeit der Methode in der Praxis beziehen und sich als Hypothese formulieren lassen. In einem zweiten Schritt können anhand einer oberflächlichen Sichtung des Materials Hypothesen angefügt werden, die sich aus dem Textmaterial herleiten lassen (vgl. Früh 2017).

Es wurden Annahmen aus der Theorie gesammelt, die sich direkt auf die Handhabbarkeit und Einstellungen zu Direct Behavior Ratings in der Praxis bezogen. Nach kurzer Zeit ließen sich drei Oberkategorien herausarbeiten, die im Folgenden mit ihren Theoriebezügen und jeweiligen Unterkategorien dargelegt werden sollen. Wichtig ist hierbei, dass die Kategorienbildung nicht auf Vollständigkeit ausgelegt werden kann und

gerade der Reduktion von Komplexität und Quantität der erhobenen Daten dienen muss:

1. Auswertung: Hierunter fallen Annahmen zur Auswertbarkeit der Graphen, sowie der Interpretation der Ratingergebnisse.

Insgesamt wurden der Oberkategorie drei Unterkategorien zugeordnet, „Graphenanalyse“ (1), „Grapheninterpretation“ (2) und „Berücksichtigung von Verhaltensmotiven“ (3).

Graphenanalyse (1): Die aus den Ratings gewonnenen Daten werden, wie in Kapitel 5.2 dargelegt als Verlaufsgraph dargestellt. Es wird vielmals angenommen, dass sich aus den gewonnenen Graphen die Entwicklung des Verhaltens ablesen lässt. So folgern Huber und Rietz: „Stellt man die Anzahl der erreichten Punkte pro Woche zum Beispiel als Kurve dar, könnten professionelle Helferinnen und Helfer, aber auch Eltern und ein Schüler oder eine Schülerin selbst ohne großen Aufwand die Verhaltensentwicklung (…) erkennen“ (ebd. 2015, S. 93). Casale, Hennemann et al. schließen: „Aus diesen Trends, die sich anschaulich mit Liniendiagrammen darstellen lassen, können Rückschlüsse auf die Wirksamkeit der Förderung (…) gezogen werden.“ (ebd. 2015a, S. 330) Es stellt sich jedoch die Frage, ob die Graphen tatsächlich so leicht von den Lehrkräften abgelesen werden können, oder ob sich in der Praxis das Ablesen und korrekte Deuten der Graphen als schwierig erweist. Aus diesem Grund wurde die Hypothese *„Die Lehrkräfte können anhand des Graphen den Entwicklungsverlauf des Verhaltens beschreiben“* abgeleitet. Mögliche Indikatoren für passende Aussagen im Material sind: Der Einbezug der Graphen bei der Beschreibung des Verhaltens und Graphenbeschreibungen oder die Beschreibung der Graphen an sich.

Grapheninterpretation (2): Nach dem Ablesen der Graphen folgt zumeist die Entscheidung über eine geeignete Fördermethode. Wie schon im vorigen Zitat angedeutet, soll diese Entscheidung unter Einbezug der Graphen erfolgen, was auch als Data based decision making bezeichnet wird. Autoren folgern für die Anwendung der Methode folgendes: „Aus diesen Trends, die sich anschaulich mit Liniendiagrammen darstellen lassen, können Rückschlüsse auf die Wirksamkeit der Förderung (…) gezogen werden. Verbessert sich etwa das Verhalten (…), wird die Förderung beibehalten. Ist keine positive Verhaltensänderung feststellbar, wird die Förderung entsprechend angepasst“ (Casale et al. 2015a, S.330) und „Insbesondere die Auswertung eines Entwicklungsverlaufs als „Response“ auf zuvor eingeleitete pädagogische Maßnahmen in Schule

oder Umfeld bietet dabei neue Ansatzpunkte für die weitere pädagogische Arbeit." (Huber und Rietz 2015, S. 93) Auch hier stellt sich die Frage, ob Lehrkräften die Planung des weiteren Vorgehen im Sinne eines data-based-decision-makings gelingt oder nicht. Die hieraus folgende Hypothese lautet daher: *„Die Lehrkräfte können anhand der Graphen Rückschlüsse über das weitere Vorgehen zu ihrer Förderung ziehen."* Indikatoren für Aussagen zu der Hypothese sind: Die Lehrkräfte beziehen sich bei Beibehaltung oder Änderung ihrer Fördermethoden auf die Verlaufsgraphen; Lehrkräfte beziehen sich bei Förderentscheidungen auf die Verlaufsgraphen; Lehrkräfte beziehen den Einsatz von Fördermaßnahmen in die Beschreibung der Verlaufsgraphen ein.

Berücksichtigung von Verhaltensmotiven (3): Die Methode des DBR steht in der Kritik, Verhalten nur als rein oberflächlich wahrnehmbares Verhalten zu betrachten. Willmann (2018) kritisiert die Verwendung der Methode, da subjektiver Sinn, individuelle Bedeutung und emotionale Hintergründe des Verhaltens ausgeblendet würden. Es stellt sich die Frage, ob die Lehrkräfte bei Anwendung der Methode, dem Verhalten zugrundeliegende Motive tatsächlich unberücksichtigt lassen, oder dennoch mit in die Förderentscheidungen einbeziehen. Wäre solch eine Verknüpfung möglich, würde das der Kritik Willmanns zumindest in Teilen widersprechen und die Methode für ein ganzheitliches Verständnis von Verhalten offenhalten. Daher wird die Hypothese „*die Lehrkräfte nutzen neben den beobachteten Symptomen auch andere Begründungen zur Erklärung des Verhaltens bei der Förderentscheidung"* in die Erhebung einbezogen. Mögliche Indikatoren sind hier die Berücksichtigung besonderer schulischer oder häuslicher Situationen bei der Graphenanalyse und Förderplanung sowie die Berücksichtigung besonderer Charaktermerkmale bei der Förderplanung.

2. Anwendung: Hierunter fallen Annahmen zur Ratingsituation an sich. Insbesondere der Ablauf und der flexible Umgang mit dem Ratingbogen sind hier anzusiedeln. Insgesamt wurden vier Unterkategorien gebildet: Direkte Beobachtung (4); Itemformulierung (5); Flexible Wahl der Zeiträume (6); Flexible Wahl der Verhaltensbereiche (7) und die Wahl des Raters (8).

Direkte Beobachtung (4): Die Direktheit der Methode gilt als eines der wichtigsten Merkmale von Direct Behavior Ratings (siehe Kapitel 3.1). Aus diesem Grund soll die Bewertung des Verhaltens direkt nach dem Beobachtungsintervall erfolgen (vgl. Briesch 2016). Es stellt sich die Frage, ob die Lehrkräfte dieser Anforderung gerecht

werden können und es im Alltag tatsächlich schaffen, das Verhalten direkt im Anschluss an die Beobachtungssituation mittels Ratingbogen zu bewerten. Die hieraus abgeleitete Hypothese lautet: *„Die Lehrkräfte beurteilen das Verhalten direkt im Anschluss an die Beobachtungssituation.“* Indikatoren, welche auf geeignete Textstellen verweisen sind beispielsweise Äußerungen, die sich auf den Bewertungszeitpunkt beziehen.

Itemformulierung (5): Die geeignete Formulierung der Items ist eine Notwendigkeit, um einen Ratingbogen anwendbar und vergleichbar zu machen. Aus diesem Grund müssen die Items so gestaltet werden, dass keine eigenen Deutungen für das Verständnis notwendig sind (vgl. Briesch 2016). Daher gilt es die Handhabbarkeit des Ratings auch in Bezug auf dessen Items zu untersuchen. Die hierfür generierte Hypothese lautet*: „Die Lehrkräfte können die Items eindeutig bewerten.“* Indikatoren für Stellen, die Hinweise auf diese Hypothese geben, sind: Fragen, wie Items zu verstehen sind, Hinweise auf die einfache Bewertung mittels der Items, Probleme oder Leichtigkeit in der Zuordnung von Items, Schilderungen vom Ankreuzen von Items.

Geeignete Wahl der Zeiträume (6): Insgesamt geht es in den Unterkategorien 6-8 um die Flexibilität des Ratinginstruments. Die Lehrkräfte sollen Zeiträume auswählen, in denen sie das Verhalten bewerten. In der Theorie wird davon ausgegangen, dass zum einen die Häufigkeit der Ratings variieren kann (vgl. Briesch 2016), zum anderen aber auch die Zeiträume der Bewertungssituationen an sich (vgl. Casale et al. 2015a). Hier stellt sich die Frage, ob die Lehrkräfte für sich geeignete Zeiträume für die Ratings festlegen können oder ob sie hierbei Schwierigkeiten haben. Die Hypothese: „*Die Lehrkräfte wählen für sie passende Bewertungszeiträume“* kann hier Aufschluss geben. Indikatoren sind: Äußerungen zum Wechsel von Bewertungszeiträumen, Aussagen zu Problemen, das Verhalten im Bewertungszeitraum zu erfassen, Zufriedenheit mit den Bewertungszeiträumen.

Flexible Wahl der Verhaltensbereiche (7): Eine weitere Entscheidung, die die Lehrkräfte treffen ist die Wahl von Verhaltensbereichen. Auch diese soll flexibel geschehen. Die Lehrkräfte sollen die Verhaltensbereiche flexibel auswählen können (vgl. Briesch 2016). Zudem sollen die Ratinginstrumente prinzipiell für jedes beobachtbare Verhalten genutzt werden können (vgl. Casale et al. 2015a). Es stellt sich auch hier die Frage, ob die Lehrkräfte geeignete Verhaltensbereiche für das Schülerverhalten finden

können. Die Hypothese lautet daher: „*Die Lehrkräfte verwenden für sie geeignete Verhaltensbereiche.“* Indikatoren für diesen Bereich betreffende Aussagen sind: Probleme bei der Auswahl von Verhaltensbereichen, Aussagen von Zufriedenheit mit den Verhaltensbereichen, Wechsel von Verhaltensbereichen.

Wahl des Raters (8): Ratings sollten möglichst von einer Person durchgeführt werden. Dies zeigen insbesondere die Ergebnisse zur Interraterreliabilität in Kapitel 3.2. Es wird darauf verwiesen, dass die Bewertungen des Verhaltens immer durch die gleiche Person erfolgen sollen. An dieser Stelle stellt sich die Frage, ob die Zuordnung einzelner Rater in inklusiven Schulsystemen möglich ist, wobei zu berücksichtigen ist, dass die Durchführung an einem eher „klassischen“ Schulsystem mit Klassenlehrkräften durchgeführt wurde. Daher wird die Hypothese „*Die Lehrkräfte können eine Person bestimmen, die das Verhalten bewertet“* formuliert. Mögliche Indikatoren sind, Probleme bei der Bewertung des Verhaltens, Zufriedenheit mit der Bewertungssituation, Probleme bei der Wahl eines geeigneten Raters, Probleme bei der kontinuierlichen Bewertung des Verhaltens.

3. Einstellungen: Einstellungen von Lehrkräften zu einem Instrument können einen großen Einfluss auf dessen Handhabung im schulischen Alltag haben. Stehen die Lehrkräfte einer Methode skeptisch gegenüber oder empfinden diese als nicht für den schulischen Alltag geeignet, kann dies dazu führen, dass die Implementierung bzw. Anwendung abgebrochen wird. Es gilt daher auch die Einstellungen der Lehrkräfte zu bestimmten Aspekten des Instruments zu berücksichtigen und in die Ergebnisse miteinzubeziehen. Dies gilt insbesondere, da das Instrument häufig mit Attributen wie Effizienz und Ökonomie verbunden wird, die subjektiv äußerst unterschiedlich eingeschätzt werden können. Für die Erfassung der Einstellungen wurden daher fünf Unterkategorien hergeleitet: Ökonomie; Eignung für den schulischen Alltag; Zufriedenheit; Kommunikation; Erwartungen an eine bessere Diagnostik.

Ökonomie (9): Unter dem Begriff Ökonomie wird im Folgenden die Umsetzbarkeit der Methode mit den Ressourcen, die den Lehrkräften zur Verfügung stehen verstanden. In der Theorie wird stets betont, dass DBRs als effizient gelten, so betont Briesch (2016), dass DBRs in ein paar Sekunden ausgefüllt werden könnten. Casale und Hennemann (2015a) äußern sich hingegen vorsichtiger, indem sie betonen, dass Förderungen finanziell, strukturell und personell umsetzbar bleiben müssen und DBRs hierzu

geeignet sein könnten. Es lohnt sich daher zu überprüfen, wie die Lehrkräfte die Ökonomie des Ratings „PUTSIE“ einschätzen. Daher wurde die Hypothese *„Die Lehrkräfte empfinden das Ratinginstrument „PUTSIE“ ökonomisch“* abgeleitet. Indikatoren sind Beschreibungen der Handhabbarkeit des Bogens; Beschreibungen zur Schnelligkeit bei der Bewertung der Items; Äußerungen von Ressourcenproblemen (zu wenig Zeit, zu wenig Personal).

Eignung für den schulischen Alltag (10): Es macht einen Unterschied, ob ein Instrument in den schulischen Alltag integrierbar scheint (siehe vorheriger Abschnitt) und ob das Kosten-Nutzen-Verhältnis der Methode positiv eingeschätzt wird. DBRs sind als dauerhaftes Diagnostikinstrument für die Verhaltensverlaufsdiagnostik gedacht und sollen als dauerhaftes Instrument auf der Förderebene 2 und 3 in RTI-Modellen Anwendung finden (vgl. Briesch 2016). Es muss daher untersucht werden, ob das Ratinginstrument von den Lehrkräften so wahrgenommen wird, dass diese sich vorstellen können damit auch über den Erhebungszeitraum hinaus zu arbeiten. Die zu überprüfende Hypothese lautet: „*Die Lehrkräfte äußern, das Ratinginstrument „PUTSIE“ weiter nutzen zu wollen“*. Indikatoren für die Textanalyse sind insbesondere positive oder negative Antworten auf die Frage, ob eine Weiterarbeit mit dem Instrument vorstellbar ist.

Zufriedenheit mit der diagnostischen Qualität (11): Ein weiterer Aspekt, den es zu berücksichtigen gilt ist die Zufriedenheit der Lehrkräfte mit der diagnostischen Qualität der Methode. So postuliert Briesch (2016), dass sowohl Lehrer und Schulpsychologen Umfragen zufolge DBRs als verlaufsdiagnostische Instrumente akzeptieren. Demzufolge müssten sich die Lehrkräfte auch positiv gegenüber der diagnostischen Güte des Ratinginstruments „PUTSIE“ äußern. Daher wird die Hypothese „*Die Lehrkräfte äußern sich positiv gegenüber „PUTSIE“ als Diagnostikinstrument“* aufgestellt. Mögliche Indikatoren hierfür sind Äußerungen inwiefern die Verwendung des Ratingbogens hilfreich/nicht hilfreich für die Betrachtung des Schülerverhaltens war; Äußerungen zum Einbezug/Nichteinbezug des Instruments in die Förderplanung; Äußerungen zu Verbesserungen des Instruments und Antworten auf die Frage, ob die Lehrkräfte „PUTSIE“ als diagnostisches Instrument des Verhaltens geeignet empfinden.

Kommunikation (12): Ein großer Inhaltsbereich, welcher auch schon in Kapitel 3.1 dargestellt wurde ist die Vereinfachung der Kommunikation über das Schülerverhalten

mittels Verlaufsgraphen. Es wird davon ausgegangen, dass die Darstellung des Verhaltens mittels Verlaufsgraphen die Kommunikation mit an der Förderung beteiligten Personengruppen vereinfacht (vgl. Huber und Rietz 2015). Es stellt sich jedoch die Frage, ob die Lehrkräfte die Verlaufsgraphen tatsächlich als geeignet für die Kommunikation über das Verhalten einschätzen. Hierzu passt die Hypothese *„Die Lehrkräfte können sich vorstellen die Graphen für die Kommunikation über das Verhalten der Schüler_Innen zu nutzen."* Indikatoren sind positive oder negative Äußerungen über die Eignung der Graphen zur Darstellung des Verhaltens; Positive oder negative Antworten auf die Frage des vierten Gesprächs, ob die Lehrkräfte sich vorstellen können die Ratingergebnisse zur Kommunikation über das Verhalten hinzuzuziehen.

Erwartungen (13): Lehrkräfte können unterschiedliche Erwartungen an Ratinginstrumente haben. Es ist wichtig zu untersuchen, welche Hoffnungen die Lehrkräfte mit dem Ratinginstrument verknüpfen und ob diese durch die Theorie gedeckt werden können. Direct Behavior Ratings dienen vorerst der Diagnostik des Schülerverhaltens über die Zeit. Daran anschließen können sich Erwartungen an eine bessere Diagnostik oder Evaluation von Fördermaßnahmen. Es soll daher untersucht werden, ob die Lehrkräfte mit dem Instrument tatsächlich Erwartungen an eine bessere Diagnostik verknüpfen. Daher wurde die Hypothese „*Die Lehrkräfte verknüpfen die Methode mit Erwartungen an eine bessere Diagnostik"* aufgestellt. Als Indikator dient hier vorrangig die Antwort auf die im ersten Gespräch gestellte Frage, was die Lehrkräfte von dem Instrument erwarten. Diese Unterkategorie wurde anhand des gesichteten Materials und nicht aus der Theorie hergeleitet.

5.5.2 Codierung

Bei der Auswertung der Ergebnisse sollen in einem ersten Schritt Textpassagen passend zu den Unterkategorien gesammelt werden. Da sich diese auf die Lehreraussagen beziehen, wurde entschieden das Textmaterial zwar in Gespräche aufgeteilt zu lassen, jedoch nach an den Gesprächen teilnehmenden Lehrkräften codiert, sodass den einzelnen Lehrkräften zu den Hypothesen passende Aussagen zugeordnet werden. Bei der Codierung dient die Untergliederung nach Lehrkräften zudem als Strukturierungshilfe. Die Aussagen, werden in der jeweiligen Hypothese der Unterkategorie widersprechende oder befürwortende Aussagen eingeordnet. So kann für jede Lehrkraft untersucht werden, wie häufig sich diese positiv gegenüber einer Hypothese äußert und ein

prozentualer Anteil hypothesenbestätigender Aussagen gegenüber der Gesamtaussagen als Indikator für die Zustimmung/Ablehnung einer Unterkategorie genutzt werden. So lässt sich untersuchen, wie viele Lehrkräfte einer Hypothese eher positiv oder eher negativ gegenüberstehen, sodass die Aussagen der Lehrkräfte untereinander verglichen werden. Diese Werte können und müssen anschließend gedeutet werden, wobei Belege im Textmaterial zur qualitativen Begründung herangezogen werden können.

Für die Analyse des Textmaterials sollen „Codierregeln" formuliert werden, die darlegen, auf welche Art und Weise Belege für die einzelnen Hypothesen des Kategoriensystems ausgewählt werden: Gesucht wird nach Textstellen in den Lehreraussagen, die den einzelnen Hypothesen zustimmen oder diese ablehnen. Dieses Vorgehen ist stark mit der Sinngebung des Codierers (in diesem Falle des Autors verbunden). Wichtig ist daher mögliche Indikatoren festzulegen, die auf die Textpassagen zutreffen sollen, um aufzuzeigen, welche möglichen Merkmale passende Textpassagen haben sollen. Diese wurden schon bei der Darlegung des Kategoriensystems aufgezeigt. Zudem soll darauf geachtet werden, möglichst einzelne Lehreraussagen zu codieren, um die Textbelege kurz zu halten. Es kann jedoch notwendig für das Verständnis der Aussagen sein, die Textbelege zu erweitern, da beispielsweise die Bedeutung der Aussage erst aus den Gesprächsanteilen des Interviewers mit der Lehrkraft deutlich werden. Hierbei können für das Verständnis nicht notwendige Aussagen gekürzt werden. Auslassungen werden durch „(…)" gekennzeichnet. Jeder Aussage wird zudem der jeweilige Gesprächscode, wie er in Kapitel 5.3 dargelegt wird angefügt, um die Textstellen in den Transkripten schnell wiederzufinden. Bei der Auswahl der Lehreraussagen wird zudem darauf geachtet, dass nur Aussagen von Lehrkräften codiert werden, die an allen vier Gesprächsterminen teilnehmen konnten, da nur so die Entwicklung von Aussagen über die Zeit wahrgenommen werden kann. Auf diese Weise bleiben die Aussagen der Lehrkräfte L1 und L7 unberücksichtigt.

In einer ersten Probecodierung soll nach der Festlegung der Codierregeln anschließend das Kategoriensystem auf dessen Handhabbarkeit untersucht werden. Dies geschieht, indem ein kleiner Teil der Gesamtstichprobe des Textmaterials mit dem Kategoriensystem ausgewertet wird (vgl. Früh 2017). Auf diese Weise lässt sich untersuchen, ob geeignete Aussagen für das Kategoriensystem gefunden wurden und sich dieses für die Analyse des Gesamtmaterials eignet. Festgestellt wurde bei der Probecodierung

der Transkripte 1 bis 4, dass nicht nur die Hypothese bestätigende oder ablehnende Aussagen gefunden werden konnten. Es wurde daher neben die Merkmalskategorien „Hypothese zustimmend“ und „Hypothese ablehnend“ noch eine Kategorie „nicht eindeutig“ hinzugefügt. Diese gilt es bei der Auswertung kurz zu berücksichtigen, um Zuordnungsprobleme offenzulegen.

In diesem Kapitel wurde das Vorgehen bei der Erhebung dargelegt. Insbesondere die Erstellung des Ratinginstruments „PUTSIE“, die Gewinnung von für die Fragestellung relevanten Daten mittels der Implementierung der Methode an einer Schule mit regelmäßigen leitfadengestützten Planungsgesprächen, sowie die Transkription und Auswertung der Gesprächsdaten mittels der integrativen Inhaltsanalyse wurden dargelegt. Nun soll die Darstellung der Ergebnisse folgen.

Teil C: Ergebnisse

Der folgende Teil dient der Darstellung und Diskussion der Ergebnisse der inhaltsanalytischen Auswertung. In Kapitel 6 werden zunächst die Ergebnisse zu den einzelnen Kategorien dargelegt und kurz diskutiert. In Kapitel 7 werden Stärken und Schwächen der Vorgehensweise ergänzt. Kapitel 8 dient dem abschließenden Fazit.

6. Ergebnisse und Diskussion

Die vollständige Sammlung der zugeordneten Transkriptteile zu den einzelnen Kategorien befindet sich im Anhang (Anhang G). Für die Darstellung der Ergebnisse wird folgendes Vorgehen verfolgt. Jede Kategorie wird in einem Unterkapitel dargestellt, indem die Ergebnisse der Unterkategorien einzeln dargelegt werden. Dabei werden zunächst tabellarisch die quantitativen Ergebnisse je Unterkategorie dargestellt und kurz erläutert. Anschließend werden einzelne Textstellen beispielhaft dargelegt, um auf Gelingendes und Probleme in der Handhabbarkeit/ bei Einstellungen der Lehrkräfte zum Instrument während der Implementierung hinzuweisen. Anschließend werden die Ergebnisse und die Auswertung kurz kritisch hinterfragt und diskutiert. Ergebnisdarstellung, Ergebnisinterpretation und Ergebniskritik werden auf diese Weise je Unterkategorie behandelt und zusammengeführt. Dies lässt sich durch die Anzahl der ausgewerteten Unterkategorien begründen: Eine getrennte Darstellung der Ergebnisse und ihrer Diskussion wäre weniger übersichtlich.

6.1 Kategorie „Auswertung“

Die Darstellung der Ergebnisse zum Teil „Auswertung“ erweist sich als umfangreich. So wurde den einzelnen Unterkategorien eine Vielzahl an Transkriptstellen zugewiesen. Dies betrifft insbesondere die Unterkategorien „Graphenanalyse“ und „Grapheninterpretation“.

(1) Graphenanalyse: Das Material wurde nach Transkriptstellen durchsucht, die die Hypothese „Die Lehrkräfte können anhand des Graphen den Entwicklungsverlauf des Verhaltens beschreiben“ betreffen. Die quantitativen Ergebnisse werden in Tabelle 2 dargelegt.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2 - L6, L8, L9)	n=7	n=6	n=6
$L2_{Anzahl\ Aussagen}$	n=15	n=5	n=1
$L3_{Anzahl\ Aussagen}$	n=4	n=2	n=1
$L4_{Anzahl\ Aussagen}$	n=15	n=1	n=0
$L5_{Anzahl\ Aussagen}$	n=5	n=2	n=1
$L6_{Anzahl\ Aussagen}$	n=10	n=4	n=4
$L8_{Anzahl\ Aussagen}$	n=14	n=0	n=2
$L9_{Anzahl\ Aussagen}$	n=7	n=1	n=3

Tabelle 2: „Graphenanalyse“

Es ist zu erkennen, dass alle sieben Lehrkräfte Äußerungen tätigen, die die Hypothese bestätigen. Sechs Lehrkräften können zudem die Hypothese ablehnende Äußerungen zugewiesen werden. Bei ebenfalls sechs Lehrkräften sind Äußerungen, die in Bezug auf die Hypothese uneindeutig sind zu finden, sich jedoch auf diese beziehen lassen. Betrachtet man die Lehrkräfte einzeln lässt sich das Verhältnis von Hypothesen bestätigenden gegenüber Hypothesen ablehnenden Aussagen und uneindeutigen Hypothesen gegenüber stellen: Bei Lehrkraft L2 zeigt sich ein Verhältnis von 15:6 von Hypothesen bestätigenden Aussagen gegenüber Hypothesen ablehnenden und uneindeutigen Aussagen. Das heißt, dass die Aussagen in etwa 71% der Fälle der Hypothese zu entsprechen. Lehrkraft 2 scheint also wenig Probleme bei der Beschreibung des Verhaltens mittels der Verlaufsgraphen zu zeigen. Bei Lehrkraft L3 scheinen in etwa ein Drittel der Aussagen der Hypothese zu entsprechen, bei L4 treffen 94 % der Aussagen auf die Hypothese zu. Bei L5 etwa 63%, bei L6 64%, L8 zeigt eine Übereinstimmung von 87,5% und L9 von 64%. Insgesamt lassen sich die Ergebnisse so deuten, dass die Lehrkräfte die Graphenanalyse in den meisten Fällen bewältigen können und es schaffen, die Graphen in etwa zwei Drittel der Fälle korrekt auf das Verhalten der Schüler zu beziehen, was die Hypothese aus quantitativer Sicht eher bestätigt.

Nichtsdestotrotz muss genauer untersucht werden, wo die Schwierigkeiten für die Lehrkräfte bei der Analyse der Graphen gelegen haben. Hierzu sollen insbesondere die Hypothesen ablehnenden und uneindeutigen Aussagen noch einmal genauer betrachtet und einige Ergebnisse anhand beispielhafter Transkriptstellen beschrieben werden.

Insgesamt lässt sich festhalten, dass nur an einer Stelle eine Aussage zur Graphenanalyse nicht möglich ist und weitere Hilfe eingefordert wird.

> „Hier unten sind die einzelnen Items und hier oben der Gesamtgraph. Also ob du ohne Erläuterung erst mal weiterhin etwas damit anfangen kannst. L4: Nein. (1.2)

Die meisten der den Hypothesen widersprechenden Lehreraussagen lassen sich auf Probleme mit der Skalierung zurückführen. Dies kann beispielsweise durch folgende Aussagen begründet werden:

> „L2: Ja, also in der Grundlagenbewertung, diese vier Bewertungen, da steigt der Graph ja an, jetzt weiß ich nicht, ob die Skala „Null“ bis „Dreißig“, ob die erst mal positiv oder negativ zu bewerten ist. (3.3)“

> „L3: (...) und deswegen fällt es mir immer noch schwer, umzudenken – zu wissen, bei welchem Unterpunkt es gut ist, wenn die Kurve nach oben geht und bei welchem es schlecht ist, wenn die Kurve nach oben geht. Diese Umdenkerei ist für mich immer noch schwierig. Da muss ich immer noch den Zettel holen und gucken. (5.4)“

Eine Vielzahl von Ablesefehlern ist in der Skalierung der zu bewertenden Verhaltensbereiche begründet. Hier zeigt sich explizit ein Problem in der Handhabbarkeit mit dem Ratinginstrument „PUTSIE“ dadurch, dass die Skalierung im Bereich schulbezogenen Verhaltens positive Items abfragt, sodass ein Ansteigen des Verlaufsgraphen eine positive Auswirkung auf das Schülerverhalten aufweist, während die anderen Verhaltensbereiche negativ formulierte Items erfragen, sodass dort ein Ansteigen des Graphen mit einer negativen Verhaltensänderung verbunden ist.

Ein weiteres Problem zeigt sich bei schwankenden Graphen. Dieses kann durch folgende Aussage von L6 dargelegt werden.

> „L6: Naja, also hier war es gut, und dann nicht mehr und dann geht es wieder nach oben. Ja, es ist schwankend oder nicht? (8.3)“

Einzelnen Lehrkräften scheint es Schwierigkeiten zu bereiten, bei Graphenschwankungen in längerfristige Verhaltensentwicklungen erkennen. Schwankungen werden zudem teilweise als „schlechtestes“ Verhalten wahrgenommen.

Insgesamt ließe sich die Skalierung gegebenenfalls durch Indikatoren an der Y-Achse (zum Beispiel Smileys) beheben, die eine klare Zuordnung der Graphenhöhe mit po-

sitivem oder negativem Verhalten ermöglichen. Ablesefehler bei Graphenschwankungen ließen sich möglicherweise durch das Einzeichnen einer Trendlinie beheben, welche Schwankungen ausgleicht und den durchschnittlichen Verlauf des Graphen angibt. Insgesamt zeigt sich jedoch, dass die Hypothese zu großen Teilen bestätigt werden kann.

Bei dieser Unterkategorie ist darauf hinzuweisen, dass sich die Äußerungen die der Hypothese zustimmen in ihrer Qualität deutlich unterscheiden. So äußern beispielsweise einige Lehrkräfte ein „steigen" oder „sinken" der Graphen, während andere von einem „der geht nach unten" oder der „geht nach oben" sprechen. Dies erschwerte teilweise auch die Codierung der Kategorie. An dieser Stelle wurde lediglich darauf geachtet, dass die Darstellungen der Lehrkräfte den zugrundeliegenden Graphen des Gesprächs entsprechen.

(2) Grapheninterpretation: Die der Analyse zugrundeliegende Hypothese lautet: „Die Lehrkräfte können anhand der Graphen Rückschlüsse über ihre Förderung ziehen." Die quantitative Ergebnisdarstellung folgt in Tabelle 3.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2 - L6, L8, L9)	n=7	n=6	n=5
$L2_{Anzahl\ Aussagen}$	n=12	n=0	n=0
$L3_{Anzahl\ Aussagen}$	n=11	n=2	n=4
$L4_{Anzahl\ Aussagen}$	n=9	n=1	n=2
$L5_{Anzahl\ Aussagen}$	n=3	n=3	n=4
$L6_{Anzahl\ Aussagen}$	n=9	n=3	n=2
$L8_{Anzahl\ Aussagen}$	n=4	n=1	n=2
$L9_{Anzahl\ Aussagen}$	n=4	n=2	n=2

Tabelle 3: „Grapheninterpretation"

Es ist zu erkennen, dass ähnlich zu Unterkategorie 1 wieder alle in die Analysen einbezogenen Lehrkräfte die Hypothese bestätigende Aussagen tätigen, jedoch auch sechs Lehrkräften die Hypothese ablehnende Aussagen zugeordnet werden konnten. 5 Lehrkräfte äußern sich zudem uneindeutig. Prozentual lässt sich der Anteil hypothesenbestätigender Aussagen je Lehrkraft wie folgt darstellen. 100% der Aussagen von L2, 69% der Aussagen von L3, 64% von L6 und 65% der Aussagen von L4 bestätigen die Hypothese. Demgegenüber zeigen nur 30% der Aussagen von L5 bestätigende Übereinstimmungen. 57% an hypothesenübereinstimmenden Aussagen von L8 und 50% solcher Aussagen von L9 vervollständigen das Bild. Insgesamt zeigt sich, dass einem Großteil der Lehrkräfte hypothesenübereinstimmende Aussagen über 50% zugeordnet wurde, die Hypothese demnach eher bestätigt werden kann. Eine Lehrkraft zeigt dagegen einen niedrigen Anteil hypothesenübereinstimmender Aussagen, wobei

hier ein hoher Anteil uneindeutiger Aussagen aufzufinden ist (40%). Demnach muss auch hier genauer untersucht werden, wo Probleme bei der Nutzung der Graphenergebnisse für die Förderplanung lagen.

Eine sehr große Gruppe an Hypothesenablehnenden und -uneindeutigen Aussagen betrifft den Einbezug des subjektiven Empfindens und anderer Informationsquellen der Lehrkraft, die neben dem Diagnoseinstrument einbezogen werden. Wahrnehmungen der Lehrkraft werden für die Beibehaltung, Installation oder Änderung von Fördermaßnahmen genutzt. Dabei wird häufig nicht mehr auf die Graphenergebnisse rekurriert, sodass keine datenbasierte Entscheidung verfolgt wird. Folgende Aussagen können als Beispiele dienen:

> „M: (…) am Anfang eurer Interventionsphase hat er noch mal zwei Hochs irgendwie gehabt, aber das waren zwei Tage und nach diesen zwei Tagen ist er ja rapide noch mal unter die Baseline gerückt. Und da wäre jetzt die Frage, wie ihr da jetzt weitermachen würdet. Wenn du sagst, es ist sowieso gerade schwierig bei ihm müsste man jetzt überlegen, ob man jetzt trotzdem sagt, man fördert so weiter oder man setzt die Ziele trotzdem ein bisschen höher oder niedriger, da wäre jetzt die Frage, wie ihr da weiter fördern würdet.
> L5: Also ich würde das auf jeden Fall beibehalten wollen, weil er hat schon Tage an denen er daran arbeiten will und ich sage mal, zumindest an solchen Tagen wäre es für ihn auch positiv zu sehen, wenn er das gut gemacht hat. (6.4)“
> „L8: Ja man könnte es in dem Sinne modifizieren, dass sie es nicht für wichtig gehalten hat, und, dass es ein oder zwei Mal nur bekommen hat. Vielleicht muss sie auch die Erfahrung erst mal machen, dass sie es auch häufiger bekommt, wenn es dann halt klappt. Also vielleicht war das für sie kein Anreiz, der groß genug war, um es überhaupt zu machen. (10.3)“

> „L9: Das ist immer dieses „lässt sich leicht ärgern“, das ist so tagesformabhängig, manchmal kommt er wegen jedes Pups quasi und
> L3: beschwert sich über die anderen, ne?
> M: Ja, dann
> L9: Es ist aber halt total schwierig, wenn das jetzt nicht mehr auftaucht, dass wir es jetzt anwenden und er es nicht mehr macht, verstehst du wie ich das meine? (11.2)“
> „L8: Das er sich melden soll. Also beim ersten Mal, dass er sich drei Mal meldet in einer, also in einer bestimmten Phase und beim zweiten Mal hatten wir gesagt, dass sich das ausdehnt auf den ganzen Tag, so sechs Mal am Tag, also dass er sich den ganzen Tag melden soll.
> M: Ok.
> L8: Wobei das hat er ja auch, das ist ja auch gut gewesen.
> M: Also da bist du auch vom Eindruck her zufrieden mit dem Ziel, das hat er von deinem Eindruck her auch erreicht?
> L8: Ja. (9.4)“

Hieran gliedern sich zudem Aussagen an, die zwar die Graphen bei der Förderplanung miteinbeziehen, jedoch durch das subjektive Empfinden und ökonomische Entscheidungen überlagert werden:

> „L4: Also ich hatte jetzt das Gefühl, dass er wirklich nicht mehr so oft unnötig aufsteht, dass das wirklich, auch wenn der Graph das vielleicht nicht so zeigt, aber das war zumindest mein persönliches Empfinden, dass das wirklich besser geworden ist und, dass man jetzt auch überlegen könnte, ob man das auf einen anderen Item jetzt macht. (1.4)“

> „L6: Das würde dann ja vielleicht hier passen. Dann hat er gerade hier sofort das mit der Kontrollkarte sofort ausgeführt, wobei ich ihn immer daran erinnern muss. Also ich lege ihm quasi

> das morgens bevor die Kinder kommen liegt das auf seinem Platz, aber ich muss das einfordern. Also er macht das nicht alleine. Ich muss halt sagen: „Denk an deine Karte!“ und das ist halt für mich nicht wirklich ne Arbeitserleichterung. So gesehen ist das vollkommen tagesabhängig wie er arbeitet, ob er überhaupt arbeitet, wie viel er schafft. (8.4)“

Zudem zeigen sich teilweise Unstimmigkeiten in der Rückführung von Graphenergebnissen auf die verbundenen Verhaltensziele, wie sich am folgenden Beispiel zeigen lässt.

> „L8: Ich gucke grad, wir hatten ja bei ihr drauf geachtet, ob sie schnell mit der Arbeit beginnt und dann auch bei der Arbeit bleibt. Ja, ist hier ein bisschen schwer abzufragen. Weil hier „bringt Schularbeiten zu Ende“, das schafft sie nie, aber das hat nicht unbedingt damit zu tun, dass sie sich nicht auf die Sache konzentriert, also da kann ich jetzt nichts genau meinem Förderziel zuordnen. (10.4)“

L8 hat scheinbar Schwierigkeiten, die Verlaufsgraphen und einzelne Items auf ihre Förderung zu beziehen. Sie sieht scheinbar keinen Zusammenhang ihrer Diagnose mit ihrer Intervention. Auf diese Weise geht die wichtige Funktion des Ratinginstruments, die eigene Intervention zu evaluieren verloren.

Abschließend soll auf das Problem hingewiesen werden, dass Lehrkräfte einige Verhaltensziele für das Rating ausschließen, da diese ihrer Meinung nach nicht häufig genug auftreten. Sie schließen die Möglichkeit, Ergebnisse in einzelnen Bereichen im Rahmen des Erhebungszeitraums zu erlangen von vornherein aus:

> „L9: Ja, aber das ist, dieses „lässt sich leicht ärgern“ zum Beispiel, das hätte ich Montag sagen können, aber die letzten zwei Tage hätte ich da nichts zu, also das ist eben mein Problem. (11.2)“

> „L4: Ich glaube, ich sage das jetzt einmal. Ich glaube, dass wir Impulsivität, vielleicht in einem kürzeren Zeitraum etwas bewirken können. Ich glaube, dass wir, wenn wir am Trotzverhalten arbeiten, dass das eine längerfristige Geschichte ist. (…) Das man vielleicht wenn man jetzt sagt, wir suchen uns den Punkt Impulsivität jetzt raus, dass wir dann in kürzeren Abständen Ergebnisse auch sehen. Was wir glaube ich beim Trotzverhalten nicht in diesem Zeitraum, die dieser Test hier läuft, ich glaube das wäre schwierig. Ich glaube das muss länger laufen. (1.2)“

Positiv muss jedoch erwähnt werden, dass die Lehrkräfte viele Interventionen an einzelnen Items ansetzen die Intervention anhand der Graphen evaluieren. Auf diese Weise kommen die meisten Hypothesen zustimmenden Aussagen zustande. Beispielhaft soll die Förderung von L2 dargestellt werden.

> „L2: Ja, also, genau er braucht auf jeden Fall nen Fokus, weil so ist es viel zu breit gefächert. Ähm, das stach gerade nicht so raus, dieses „Steht in Situationen auf, in denen Sitzenbleiben erwartet wird, aber das ist im Grunde das, wenn er in Arbeitsphasen anfängt, Plätze zu wechseln. Item 17, das wäre dann, das ist orangefarbend das wäre auch gleichbleibend schwach. (3.2)“

> „L2: Das heißt, da war ein Ansteigen zu beobachten, das heißt das Verhalten ist durchaus besser geworden, da war am Anfang noch mal so ein Hin und her, aber jetzt seitdem wir die Interventionsmaßnahme haben, ist es im Grunde konstant auf einem höheren Niveau, das heißt also da ist eine Verhaltensverbesserung zu erkennen. (3.3)“

„L2: Ja, auch bei Luca war es wieder, als das Ziel dann höher gesteckt war, auf du musst den und den Teil des Blattes auch bewältigt haben, es reicht nicht mehr nur, dass du am Platz sitzt, ich will auch wissen, dass du das und das geschafft hast, da wurde ihm der Anspruch zu hoch. Da hat er schnell dann das Blatt liegen lassen, oder nichts gemacht. (...)
M: Ok, ja, erkennst du das im Graphen wieder?
L2: Ja, an den zackigen Ausschlägen. Wenn das hier die Intervention ist, da war am Anfang war es wieder unruhiger mit Auf- und Abbewegungen und jetzt scheint es sich wieder auf weniger starke Verhaltenssprünge sich einzuordnen wieder. (3.4)“

Kritisch lässt sich an dieser Stelle bemerken, dass letztlich die Förderentscheidungen auch in diesem Gespräch wieder von subjektiven Faktoren mitbeeinflusst werden. So nimmt die Lehrkraft zwar den Einbruch der Förderung wahr und will diese verändern, begründet ihr weiteres Vorgehen jedoch wie folgt.

„L2: Also ich glaube, das würde ich nicht ab der nächsten Woche machen, weil dafür sind die beide gerade viel zu sehr darauf bedacht auch diese Rückmeldung aus der ersten Phase zu kriegen, wenn ich das jetzt weglassen würde, glaube ich wäre wieder dieses „ja toll, dann habe ich ja gar keine Chance einen Muggelstein zu kriegen.“ Und über drei geschaffte Aufgaben sich einen Stein zu verdienen schaffen die beiden eh schlecht, weil sie eh nicht so viel schaffen wie andere Kinder und dementsprechend müsste das jetzt eh ne Zeit so laufen, weiter.
M: Also würdest du das jetzt ne Zeit weiterlaufen lassen, dann würdest du wahrscheinlich schauen, ob du den Fokus auf die zweite Arbeitsphase legst.
L2: Ja, oder Fokus auf die zweite und wenn die zweite Arbeitsphase auch super läuft, kriegst du da sogar zwei Muggelsteine für. (3.4)“

Der Einbruch der Förderung, welcher durch die Lehrkraft unter anderem zuvor durch die Graphenergebnisse dargelegt wird, wird zwar thematisiert, das weitere Vorgehen wird jedoch nicht auf den Graphen, sondern das wahrgenommene Verhalten der Schüler bezogen. So werden die Graphenergebnisse implizit und nicht explizit berücksichtigt.

Insgesamt lässt sich festhalten, dass die meisten Äußerungen der Lehrkräfte, die die Graphenergebnisse auf die Förderung beziehen, die Evaluation der vorhergegangenen Förderung betreffen. Zudem ist zu beobachten, dass eine Evaluation Probleme bereitet, wenn Förderziele nicht nah an einzelne Items des Ratingbogens angelehnt sind. Zudem wird für die Entscheidung weiterer Förderung lediglich selten auf die Graphen und stärker auf die Wahrnehmung des Verhaltens in der Klasse Bezug genommen, sodass zwar die Grundtendenz in einem Sinne von „das Verhalten wird besser oder schlechter“ über die Graphen abgelesen wird, die Anpassung der Förderung sich jedoch auf direkten Wahrnehmungen der Intervention basiert.

An dieser Stelle lässt sich diskutieren, ob zukünftige Förderentscheidungen im Sinne eines Data Based Decision Making tatsächlich die explizite Berücksichtigung der Graphen für die weitere Förderung einbeziehen muss, oder ob die Evaluation der Intervention genügt. So könnte in der Berücksichtigung subjektiver Wahrnehmung in Kombination mit den Graphen ein „gesunder Mittelweg“ einer Förderung liegen, bei dem

Verhalten aus einer ganzheitlichen Perspektive betrachtet werden kann. Um das data based decision making zu fördern könnteüberlegt werden, den „Decision-making tree for determining next steps based on progress monitoring data“ (Briesch 2016, S. 230), welcher den Lehrkräften zwar im Handbuch verfügbar war, in der Erhebung jedoch nicht direkt oder zu selten angesprochen wurde (was zugleich kritisch zu betrachten ist) stärker zu implementieren. Möglicherweise könnte dies gewährleistet werden, indem bei einer Onlineversion den Verlaufsgraphen eine automatische Förderempfehlung im Sinne der Urteile „Intervention anpassen/ändern“ bei negativen Verhaltensentwicklungen oder „Intervention beibehalten/ausblenden/erweitern“ bei positiven Verhaltensentwicklungen angeschlossen wird.

Kritisch ist an dieser Stelle zu berücksichtigen, dass bei der Unterkategorie „Graphen-interpretation“ ein Teil der Unterkategorie „Graphenanalyse“ miteinspielt. So stellt die Unterkategorie „Grapheninterpretation“ lediglich eine Erweiterung um den Aspekt des Einbezugs der Fördermaßnahme in die „Graphenanalyse“ dar. Die beiden Kategorien erscheinen so nicht ganz trennscharf. Dies wird noch verstärkt, indem nahezu alle Aussagen, die sich in irgendeiner Weise auf die Fördermaßnahmen beziehen einbezogen wurden. Es lässt sich hinterfragen, ob die Unterkategorie differenziert genug für die Auswertung gewählt und codiert wurde, beispielsweise eine Hypothese: „Die Lehrkräfte orientieren sich in ihrer Entscheidung am „Decision-making tree for determining next steps based on progress monitoring data“ (Briesch 2016, S. 230)“ besser gewählt wäre. Hierfür hätte dieser jedoch während der Implementierung stärker betont werden müssen.

Abschließend muss daher geurteilt werden, dass der Hypothese „Die Lehrkräfte können anhand der Graphen Rückschlüsse über ihre Förderung ziehen“ Grundsätzlich zwar zugestimmt werden kann, wenn es um die Evaluation der Interventionsmaßnahmen geht. Fast nie werden jedoch zukünftige Förderentscheidungen anhand der Graphen abgeleitet.

(3) Berücksichtigung von Verhaltensmotiven: Wie schon in Kategorie (2) angedeutet scheinen die Lehrkräfte neben den Graphenanalysen weitere Indikatoren in ihre Förderentscheidungen miteinzubeziehen. Durch die Hypothese „Die Lehrkräfte nutzen neben den beobachteten Symptomen auch andere Begründungen für die Erklärung des Verhaltens bei der Förderentscheidung“ soll daher untersucht werden, ob die Lehr-

kräfte ihre Förderentscheidungen rein von beobachtbaren Symptomen abhängig machen, oder das Verhalten durch weitere Komponenten in einen ganzheitlicheren Kontext setzen.

Lehrkräfte	Hypothese bestätigend	Hypothese ablehnend	Aussage uneindeutig
N_{gesamt}=7 (L2 - L6, L8, L9)	n=5	n=7	n=3
$L2_{Anzahl\ Aussagen}$	n=1	n=3	n=1
$L3_{Anzahl\ Aussagen}$	n=3	n=3	n=1
$L4_{Anzahl\ Aussagen}$	n=0	n=3	n=0
$L5_{Anzahl\ Aussagen}$	n=1	n=2	n=0
$L6_{Anzahl\ Aussagen}$	n=8	n=2	n=0
$L8_{Anzahl\ Aussagen}$	n=0	n=5	n=1
$L9_{Anzahl\ Aussagen}$	n=1	n=2	n=0

Tabelle 4: Ergebnisse „Berücksichtigung von Verhaltensmotiven“

Insgesamt konnten der Hypothese weniger Aussagen zugeordnet werden als zu den ersten beiden Unterkategorien. Rein quantitativ lässt sich feststellen, dass fünf Lehrkräfte Förderentscheidungen nicht nur durch die Verhaltenssymptome erklären und weitere Komponenten zur Erklärung des Verhaltens einbeziehen. Es lassen sich jedoch von sieben Lehrkräften Äußerungen finden, in denen das Verhalten allein anhand der durch das Rating erfassten Verhaltenssymptome erläutert wird. Drei Lehrkräften wurden uneindeutige Aussagen zugeordnet. Prozentual äußert sich der Anteil hypothesenbestätigender Aussagen zu den Gesamtaussagen wie folgt: L2 konnten 20% hypothesenbestätigende Aussagen zugeordnet werden. L3: 43%, L4: 0%, L5: 34%, L6: 80%, L8: 0%, L9: 34% hypothesenbestätigende Aussagen. Hierdurch würde die Hypothese abgelehnt. Durch die geringe Anzahl an Aussagen (teilweise lediglich 3 Aussagen je Lehrkraft) müssen die Werte jedoch vorsichtig betrachtet werden. Es soll außerdem darauf hingewiesen werden, dass es so scheint, dass die Werte von Lehrkraft zu Lehrkraft in Quantität und Verhältnis von hypothesenbestätigenden zu hypothesenablehnenden und uneindeutigen Aussagen stark zu schwanken scheinen.

Die codierten Textstellen beziehen sich größtenteils auf zwei Situationen. Zum einen die Festlegung von neuen Förderzielen, zum anderen die Evaluation der Intervention. Bei der Festlegung von Förderzielen ist vor allem eine Aussage zu benennen, die weitere Ursachen zur Erklärung von Verhaltenssymptomen miteinbezieht:

> „L9: Der redet ja oft ununterbrochen, können wir aber auch „er singt viel“ daraus machen?
> M: Das könnt ihr auch.
> L3: Das ist aber so eher, bei dem Singen, das nimmt er überhaupt nicht wahr, dass er da singt, also wenn er sich konzentriert und in sich zurückzieht, dann fängt er an zu singen. Also das ist ne Frage, ob man da überhaupt rangehen will, das zu verhindern im Moment, ich weiß, dass das nervt, weil das so für ihn im Moment zum Konzentrieren dazugehört und ob man ihn da vielleicht erst mal lässt, wenn er da nicht wer-weiß-wie-laut singt. Redet viel ist was anderes, weil da, Singen, das glaube ich macht er dann auch wirklich nicht bewusst. Also er

> singt dann nicht um zu singen, aber er redet dann schon um zu reden. Also deswegen würde ich, „redet viel“ finde ich in Ordnung. (11.2)“

Bei der Besprechung des Ziels wird eine Förderung im Bereich „singen“ durch L3 abgelehnt, da das von L9 als störendes empfundene Verhalten von L3 als „unterbewusster Konzentrationshelfer“ wahrgenommen wird. Hier wird das als störend empfundene Verhalten mit seinem subjektiven Sinn für den Schüler verknüpft und so für die schulische Arbeit als individuell förderlich dargelegt.

Eine weitere Lehrkraft begründet die Verhaltenssymptome ihrer Schüler und darauf aufbauenden Förderentscheidungen häufig mit dem Einbezug des häuslichen Umfeldes. Folgende Aussagen sind beispielhaft:

> „L6: Ich muss bei David wieder, wie ich das eigentlich auch immer mache, jetzt ist wieder der Zeitpunkt, wo ich bei der Mama anrufen muss und sagen muss es läuft nicht und es muss von zu Hause mehr kommen. Dann wird sie mir sagen: „Ja, machen wir!“ Dann werden die Hausaufgaben wieder zusammen gemacht oder mit Oma, das läuft wieder für zwei Wochen und dann geht das Ganze wieder von vorne los. (7.4)“

> „L6: Es war auch, also Oma und Opa sind seit dem Wochenende da und bleiben auch die ganze Woche, das heißt Patrick ist die ganze Woche über nicht in der OGS und denkt quasi den ganzen Morgen schon darüber nach, was er nachmittags alles Tolles mit Oma und Opa unternimmt. Sonst ist Patricks Tagesablauf so: Er geht nach der Schule in die OGS, wenn er von der OGS nach Hause kommt, dann geht er nach Hause und sitzt an der Playstation, am Handy oder sonst irgendwas, das heißt er hat da nicht viel. Und jetzt gerade ist natürlich richtig toll: Die waren Schlittschuhlaufen, die waren im Kino, die waren spazieren und es wird richtig was unternommen zu Hause und das erzählt er morgens immer in der Schule und deswegen kann ich das verstehen, wenn er nicht so konzentriert ist, wie er eigentlich sollte, weil er einfach total aufgeregt ist. (8.4)“

Widersprechende Aussagen zeigen häufig folgende Charakteristiken. Förderentscheidungen und Verhaltensbeschreibungen bleiben zum einen häufig allein auf der beobachtbaren Symptomebene, also dem beobachtbaren Verhalten. Zum anderen häufen sich Aussagen, die die subjektive Einschätzung der Eignung der Intervention durch die Lehrkraft beziehen.

Zwei Beispiele für die Förderentscheidung auf der Symptomebene:

> L3: Ja, ich denke wir müssen da weiter dran arbeiten, ich denke es hat sich noch nicht automatisiert das Verhalten. (1.3)

> L8: Ne, glaube ich nicht, weil er so darauf fokussiert ist, dann eben, also er meldet sich jetzt, ja, aber er kommt dann nachher direkt, also ich habe mich jetzt ja zwei Mal gemeldet, also ich habe das jetzt schon hochgesetzt, also er muss sich jetzt drei Mal melden, aber er zählt das genau ab, aber es ist noch nicht übergegangen. (9.3)

Zwei Beispiele für die Verhaltensbeschreibung durch subjektive Einschätzung der Eignung der Intervention:

> L2: Also auch bei ihm ist der Vertrag im Grunde gut angekommen, er kommt auch und fordert seine Belohnung und fordert Rücksprache darüber, wie er gearbeitet hat und auch bei ihm habe

> ich den Eindruck, dass ihn das auch bestärkt darin, dass er auch weiterhin gut am Platz arbeitet und sich nicht so ablenken lässt. (4.3)
>
> L9: Also ich finde das zieht, das zieht ganz langsam, aber es zieht. Also das, ich habe so langsam den Eindruck, dass das Früchte trägt, dass das langsam besser wird und er auf jeden Fall was tut und auch wenn das nicht immer so ist, wie wir das haben wollen aber er sitzt schon mal da und macht was und sagt sich das auch vor. Ich würde sagen, dass das wir das so weitermachen. (11.3)

Die Ergebnisse der Unterkategorie erweisen sich als äußerst divers und schwer zu deuten. Dies zeigte sich auch schon bei der Codierung, sodass eventuell die Eignung der Kategorie in Frage gestellt werden kann, zumal die Ergebnisse nicht direkt auf Probleme und Möglichkeiten der Handhabbarkeit von DBRs verweist. Es zeigt sich, dass es von Lehrkraft zu Lehrkraft unterschiedlich ist, ob und in welcher Häufigkeit Verhaltensmotive berücksichtigt werden. Um den Fokus stärker auf ein ganzheitliches Bild von Verhalten zu richten und damit der in Kapitel 3 dargelegten Kritik an der Methode zu entgehen, könnte es sinnvoll sein im Bereich der Formulierung der Förderziele nach der Baselineerhebung auch auf den möglichst vorhandenen Einbezug subjektiver Sinngebung des Verhaltens zu verweisen.

Insgesamt zeigte sich bei der Codierung der Aussagen der Kategorien im Bereich „Auswerten“ folgendes Problem. Häufig lassen sich Aussagen, die sich auf die Förderung beziehen nicht von Aussagen, die sich nur auf die Beschreibung des Verhaltens beziehen, trennen. Auf diese Weise werden in Kategorie 3, welche explizit die Berücksichtigung von Verhaltensmotiven bei der Förderplanung untersucht auch Aussagen codiert, welche sich nicht direkt auf die Förderplanung beziehen. Gegebenenfalls treten durch mangelnde Trennschärfe der Kategorien und Ungenauigkeiten der Hypothesen Codierfehler auf, sodass die quantitativen Ergebnisse mit Vorsicht zu betrachten sind. Es lassen sich jedoch Tendenzen zu Problemen in der Handhabbarkeit erkennen, welche insbesondere durch die Ergebnisse der Kategorie „Graphenanalyse“ dargelegt werden konnten.

6.2 Kategorie „Bewertung“

Die Ergebnisse der Unterkategorien der Kategorie „Bewertung“, weisen insgesamt weniger zugeordnete Aussagen auf. Aus diesem Grund scheint eine Darlegung der Ergebnisse hier zunächst einfacher möglich. Es muss jedoch darauf hingewiesen werden, dass die quantitativen Ergebnisse durch teilweise nur eine Aussage je Lehrkraft vorsichtig zu betrachten sind.

(4) Direktheit. Die betrachtete Hypothese lautet: „Die Lehrkräfte können das Verhalten direkt im Anschluss an die Beobachtungssituation beurteilen." Auf diese Weise kann eingeschätzt werden, ob die direkte Beobachtung im schulischen Alltag möglich ist.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=6 (L2 , L4 - L6, L8, L9)	n=5	n=2	n=1
$L2_{Anzahl\ Aussagen}$	n=1	n=1	n=0
$L4_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L5_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L6_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L8_{Anzahl\ Aussagen}$	n=1	n=3	n=0
$L9_{Anzahl\ Aussagen}$	n=1	n=0	n=0

Tabelle 5 Ergebnisse „Direktheit der Beobachtung"

Quantitativ (Tabelle 5) Iassen sich von sechs Lehrkräften Aussagen zu der Hypothese gewinnen (dies lässt sich darin begründen, dass eine der sieben Lehrkräfte, die bei der Codierung berücksichtigt wurde, keine eigene Beobachtung durchführte).

Fünf Lehrkräften konnten die Hypothese bestätigende Aussagen zugeordnet werden, zwei die Hypothese ablehnende Aussagen, einer Lehrkraft nur eine uneindeutige Aussage. L2 wurden zu 50% übereinstimmende Aussagen mit der Hypothese zugeordnet. Den Lehrkäften L4 bis L7 und L9 konnte jeweils eine Aussage zugeordnet werden (jeweils 100% Übereinstimmung). L8 zeigt zu 25% hypothesenbestätigende Aussagen. Einem Großteil der Lehrkräfte gelingt also die direkte Beobachtung und Bewertung des Verhaltens im Anschluss an den Beobachtungszeitraum, sodass die Hypothese aus quantitativer Sicht bestätigt wird.

Im Folgenden soll insbesondere dargelegt werden, ob aus Hypothesen ablehnenden Aussagen von L2 und L8 Gründe für das Nichtgelingen direkter Beobachtungen erkennbar werden, aus denen Rückschlüsse für eine besser gelingende Implementierung abgeleitet werden können:

> „L2: Gut, also ähm, den Fragebogen auszufüllen das hat geklappt, dadurch, dass am Montag wir beide im Unterricht waren, Dienstag nur L1 und heute wieder wir beide, zum Teil zusammen und Zeitversetzt, ist das halt ähm wahrscheinlich nen bisschen schwierig, weil vor der Pause zum Beispiel sind die Kinder generell noch aufmerksamer und können sich besser an Regeln halten als nach der Pause und je nachdem zu welchen Zeitpunkten ich dann die Beobachtungen schwerpunktmäßig mache, denke ich, ist dieser Beobachtungszeitraum den wir gewählt haben vielleicht zu groß über den gesamten Schultag (…) (4.2)"

> „L2: Also dadurch, dass wir bei der Bewertung den zu beobachtenden Bereich runtergeschraubt haben auf den Bereich Unaufmerksamkeit und auf die Arbeitsphasen, die am Platz stattfinden sollen, war es viel übersichtlicher und noch flotter und dadurch war ein klar gesetzter Zeitrahmen da, wann ich wirklich gucken muss, was macht er da. Und der Fokus war einfach noch mehr gegeben. (4.4)"

Lehrkraft L2 betont Probleme im Ausfüllen des Fragebogens in zu großen Zeiträumen. Sie kann diese nicht überblicken, da sie nicht über den gesamten Beobachtungszeitraum in der Klasse ist. Nach einer Anpassung des Beobachtungszeitraums auf „Arbeitsphasen“ gelingt der Lehrkraft jedoch die Beobachtung und wird als „flott“ beschrieben. Die Lehrkraft äußert sich jedoch nicht eindeutig, wann genau sie das Schülerverhalten bewertet.

> „L8: Also geklappt hat es schon, es war nur so, dass in der letzten Schulwoche es generell sehr unruhig war und wir ein bisschen Probleme hatten, hier in der Klasse mit einigen Kindern und da ist einfach viel Zeit für draufgegangen das wieder so in den Griff zu kriegen und darunter hat ein bisschen der Fokus auf den einzelnen Schüler gelitten (9.3)“

> „M: (...) war es da auch das gleiche, wie bei den anderen, dass du es letzte Woche nicht so gut bewerten konntest, weil es so trubelig war?
> L8: Ja, das gleiche. (10.3)“

> „L8: Ja, da musste ich erst wieder umdenken, mehr Punkte sind schlecht, ne, also die Skalierung ist ja andersherum, was mich zwischendurch, wenn ich das „mal eben“ machen wollte, nach so Arbeitsphasen oder so, hat mich das rausgebracht, also musste ich umdenken und das hat mich Zeit gekostet. Hört sich doof an, weil wusste ich ja, aber trotzdem musste ich da noch mal lesen, und noch mal, jetzt also ja. (10.4)“

Bei L8 äußern sich die Probleme ebenfalls nicht im direkten Ausfüllen des Bogens im Anschluss an die Beobachtungssituation, sondern in der Problematik, den Fokus in der Beobachtungssituation auf dem Schüler zu behalten. Die Lehrkraft äußert zudem Probleme beim Ausfüllen des Bogens, sodass die Auswertungssituation sich zeitlich ausdehnte und so das kurze Ausfüllen des Ratingbogens im schulischen Alltag erschwert wird.

Es lassen sich einerseits Schwierigkeiten in der Direkten Beobachtung im Anschluss an die Beobachtungszeiträume darin erkennen, dass der Beobachtungszeitraum falsch gewählt war (zu trubelig, nicht überschaubar). Da die Wahl der Zeiträume in Kategorie 6 genauer dargelegt wird, wird an dieser Stelle nicht genauer darauf Bezug genommen. Möglicherweise genügt bei der Implementierung ein Verweis darauf, bei Problemen mit der Beobachtung den Beobachtungszeitraum zu überdenken. Des Weiteren lassen sich Problematiken in der Skalierung erkennen. Die Lehrkraft äußert Ökonomieprobleme beim Ankreuzen der Items dadurch, dass nicht immer positives Verhalten mit „nie“ und negatives Verhalten mit „immer“ bewertet werden muss. Das Problem ließe sich möglicherweise bei der Darstellung der Items im Ratingmaterial verändern, indem den Minimalwerten „Nie“ und Maximalwerten „Immer“ jeweils ein lachender oder trauriger Smiley angefügt wird, je nachdem ob ein hoher oder niedriger Wert im Verhaltensbereich positives oder negatives Verhalten erfasst.

Insgesamt kann die Hypothese bestätigt werden. Die Lehrkräfte können das Verhalten direkt im Anschluss an die Beobachtungssituation bewerten. Falls nicht, kann dies gegebenenfalls durch Anpassung der Bewertungszeiträume oder Vereinfachungen des Ratingbogens behoben werden.

An dieser Stelle ist darauf zu verweisen, dass sich die hypothesenablehnenden Aussagen eher auf Schwierigkeiten bei der Auswertungssituation an sich, als der direkten Bewertung im Anschluss an den Beobachtungszeitraum beziehen. Dies lässt sich dadurch begründen, dass die direkte Beurteilung durch die codierten Aussagen zumindest eingeschränkt wurde.

(5) Itemformulierung. Die untersuchte Hypothese lautet „Die Lehrkräfte können Items eindeutig bewerten."

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2 , L3 - L6, L8, L9)	n=5	n=6	n=4
$L2_{Anzahl\ Aussagen}$	n=0	n=3	n=1
$L3_{Anzahl\ Aussagen}$	n=3	n=4	n=0
$L4_{Anzahl\ Aussagen}$	n=2	n=3	n=1
$L5_{Anzahl\ Aussagen}$	n=4	n=2	n=0
$L6_{Anzahl\ Aussagen}$	n=1	n=2	n=1
$L8_{Anzahl\ Aussagen}$	n=0	n=5	n=0
$L9_{Anzahl\ Aussagen}$	n=2	n=0	n=0

Tabelle 6 „Itemformulierung"

Zu der Hypothese ließen sich von allen sieben Lehrkräften Äußerungen zuordnen. Insgesamt wurden fünf Lehrkräften hypothesenbestätigende Aussagen, sechs Lehrkräften hypothesenablehnende Aussagen und vier Lehrkräften uneindeutige Aussagen zugewiesen. Der prozentuale Anteil hypothesenbestätigender Aussagen an den Gesamtaussagen je Lehrkraft stellt sich wie folgt dar: L2: 0%, L3: 43%, L4: 33%, L5: 66%, L6: 25%, L8:0%, L9:100%. Es zeigt sich also ein äußerst diverses Bild in den Codierungen je Lehrkraft. Wobei die Hypothese insgesamt eher abgelehnt werden müsste, da von über der Hälfte der Lehrkräfte weniger als die Hälfte der Aussagen die Hypothese bestätigen. Es ist jedoch darauf hinzuweisen, dass im Leitfaden keine Frage zur direkten Handhabbarkeit der einzelnen Items gestellt wurde und so die Anzahl der codierten Aussagen sehr gering ist.

Hypothesenbestätigende Aussagen verweisen vor allem auf die Passung einzelner Items oder Items bestimmter Verhaltensbereiche zum gezeigten Verhalten. Beispielhaft sind folgende Aussagen:

> „L5: Also das hier „ärgert andere absichtlich", das passt ja schon, ne

L3: Das passt schon. Genau. (5.1)“

„L5: Ja, also ich finde ganz besonders „steht in Situationen auf, in denen Sitzenbleiben erwartet wird“, und „kann nur schwer warten, bis er/sie an der Reihe ist“, das ist wirklich auch
L3: Nen Problem.
L5: Ja, nen Problem (…). (6.2)“

„L4: Ich könnte mir vorstellen, dass wir „bringt Schularbeiten nicht zu Ende“, dass wir da, wenn wir da ansetzen, gute Erfolge, dass wir, dass du auch vielleicht noch mal mit ihr rausgehst, noch mal in Ruhe mit ihr sprichst. Ihr noch einmal sagst, weil sie kommt halt ganz oft, hat die Aufgaben nicht zu Ende weil sie es kann. Wir wissen, dass sie es kann und einfach immer wieder diese Bestätigung sucht. (1.2)“

Diesen Aussagen schließen sich Aussagen über die gute Bewältigung der Bewertung an:

„L9: Nein, ich hatte keine Schwierigkeiten, das hat alles gut geklappt. Was mir aufgefallen ist, beim Durchblättern, dass ich gedacht habe, ich sehe eine Veränderung. Ansonsten ist mir nichts aufgefallen. (11.3)“

„L6: Nein, Besonderheiten nicht, ich hatte auch keine Probleme beim Ausfüllen (8.3)“

Im Bereich der hypothesenablehnenden Aussagen lassen sich Verständnisprobleme mit einzelnen Items feststellen. So lässt beispielsweise L4 die Bewertung des Items „handelt als wäre er, bzw. sie getrieben“ aus.

„L4: Ja, mit diesem Begriff „handelt als wäre er, bzw. sie getrieben“ hatte ich Schwierigkeiten, deswegen habe ich das hinterher auch einfach nicht mehr angekreuzt. (1.2)“

Ähnlich zeigen sich Probleme im Bewerten einzelner Items darin, dass einige Items bestimmte Voraussetzungen einfordern, die nicht gegeben sein müssen, wie folgende Lehreraussagen belegen:

„L4: Was mir beim Ankreuzen noch mal aufgefallen ist, es gab Tage, an denen Gregor gar nicht oder sehr wenig gearbeitet hat und dann habe ich ein bisschen über den Punkt geschmunzelt „macht Flüchtigkeitsfehler bei den Schularbeiten“, weil wenn man keine macht, kann man auch keine Flüchtigkeitsfehler machen, ich habe das dann dementsprechend angepasst und habe natürlich den Item weiter hinten angekreuzt, also, dass er halt viele Flüchtigkeitsfehler macht, damit das passt, aber eigentlich hat er keine gemacht, weil er auch nicht gearbeitet hat. (2.3)“

„L8: Also, ich bin damit schon zurechtgekommen, nur bei einem Item da wusste ich nicht, wie ich das bewerten sollte, bei „arbeitet ruhig am Platz“, es macht einen Unterschied, ob jemand ruhig dasitzt, oder ob jemand arbeitet. (9.4)“

„L8: Ja, bei einem Item „Lässt sich durch äußere Reize ablenken“, war es manchmal schwer das zu beurteilen, weil wenn es keine äußeren Reize gibt, also in Arbeitsphasen, wenn alles ruhig ist und nichts drumherum passiert, lässt sie sich natürlich auch nicht ablenken, da passiert ja gerade nichts. Das kann man glaube ich auch nicht anders abfragen. Aber ich weiß genau, wäre draußen etwas gewesen, oder wäre irgendwas gekommen, hätte das anders ausgesehen. (10.4)“

Die größte Problematik, welche sich jedoch bei der Bewertung von Items zeigt bezieht sich auf die Formulierung des Verhaltensbereichs des schulbezogenen Verhaltens. Die Items wurden so formuliert, das jeweils das Vorhandensein eines bestimmten positiven

oder negativen Verhaltens abgefragt wird. Nur im Verhaltensbereich „schulbezogenes Verhalten“ wird positiv vorhandenes Verhalten erfragt. Dies führt zu Verwirrung beim Bewerten mittels Ratingbogen:

> L5: Ja, bei Erdem musste ich immer wieder komplett umdenken, weil ich auf einmal wieder auf der anderen Seite ankreuzen musste. (6.3)
>
> L6: Ich hab das falsch, das ist natürlich „immer“ ich hab das falsch, ja das ist falsch. Ich hab das einfach umgedreht. (7.2)
>
> L8: Ja, also höchstens das mit der Skalierung. Das war mir halt nicht klar, dass in dem Bereich (…) es war mir schon klar, wenn man lesen kann, dann weiß man es, aber wenn man es so macht und es schnell macht, dann muss man eben einfach gucken. Für mich wäre es einfacher gewesen, es wäre durchgängig (…) (10.4)

Insgesamt zeigen sich also drei Probleme in der Itemformulierung, die zu der abschließenden Einschätzung führen, dass die Hypothese eher abgelehnt werden muss. Im Folgenden sollen mögliche Lösungsansätze für diese dargelegt werden.

Verständnisprobleme von Items: Den Items könnten mögliche Verhaltensbeispiele angeschlossen werden, die durch das Item erfasst werden. Das Item „handelt als wäre er bzw. sie getrieben“ könnte beispielsweise durch „überstürztes Handeln“, „hastige Bewegungen“ genauer umschrieben werden.

Voraussetzungen von Items: Die Items, welche bestimmte Voraussetzungen erfüllen, könnten überarbeitet werden. Eine vollständige Lösung des Problems erscheint jedoch schwierig. Möglicherweise muss hier darauf verwiesen werden, die Beobachtungszeiträume oder -bereiche bei einer nichtmöglichen Beobachtung anzupassen. Es kann zudem die Möglichkeit bestehen von vornherein mehrere Ratingbögen zur Erfassung des Verhaltens zur Verfügung zu stellen, sodass eine Vielzahl vorhandener Items zu einer geeigneteren Auswahl führen könnte.

Skalierung „schulbezogenes Verhalten“: Hier könnte ein zusätzlicher visueller Hinweis auf der Likert-Skala dazu führen, die Bewertung zu vereinfachen. Auf diese Weise könnte einer Umformulierung der Items umgangen werden. Diese wäre eine zweite Möglichkeit, sodass die Abwesenheit des Verhaltens im Verhaltensbereich erfragt würde und sich die Frage stellt, ob die Lehrkräfte diese genauso gut erfassen könnten. Hier wären weitere Untersuchungen nötig.

(6) Geeignete Wahl der Zeiträume. In den Ergebnissen wurde schon darauf hingewiesen, dass gelingende oder scheiternde Ratings unter anderem mit der passenden Wahl der Beobachtungszeiträume zusammenhängen. Mit der Hypothese „Die Lehrkräfte

wählen für sie passende Bewertungszeiträume“ soll der Zusammenhang genauer untersucht werden.

Lehrkräfte	Hypothese bestätigend	Hypothese ablehnend	Aussage uneindeutig
N_{gesamt}=7 (L2, L3, L4 - L6, L8, L9)	n=7	n=3	n=2
$L2_{Anzahl\ Aussagen}$	n=2	n=2	n=0
$L3_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L4_{Anzahl\ Aussagen}$	n=7	n=0	n=0
$L5_{Anzahl\ Aussagen}$	n=2	n=1	n=0
$L6_{Anzahl\ Aussagen}$	n=2	n=0	n=1
$L8_{Anzahl\ Aussagen}$	n=3	n=2	n=0
$L9_{Anzahl\ Aussagen}$	n=3	n=0	n=1

Tabelle 7: „Geeignete Beobachtungszeiträume“

Quantitativ (siehe Tabelle 7) konnten allen Lehrkräften Aussagen bezogen auf die Hypothese zugeordnet werden, zudem allen Lehrkräften hypothesenbestätigende Aussagen. Nur zwei Lehrkräften zeigten zudem hypothesenablehnende Aussagen, ebenfalls zwei Lehrkräften uneindeutige Aussagen. Die Prozentualen Anteile hypothesenbestätigender Aussagen je Lehrkraft stellen sich wie folgt dar: „L2: 50%, L3: 100%, L4: 100%, L5: 67%, L6: 67%, L8: 60%, L9: 75%“. Insgesamt kann der Hypothese also eher zugestimmt werden, die Lehrkräfte scheinen größtenteils geeignete Beobachtungszeiträume wählen zu können.

Betrachtet man die Aussagen inhaltlicher Ebene lässt sich feststellen, dass hypothesenbestätigende Aussagen zunächst die reine Zufriedenheit mit der Wahl der Zeiträume ausdrücken, beziehungsweise sich auf die Fortführung der gewählten Zeiträume beziehen:

> „M: und die Zeiträume bleiben auch die gleichen, wie bisher?
> L4: Genau, nach dem Unterricht. (2.2)“

> „L5: Mit den Zeiträumen bin ich gut klargekommen. Ja. (6.2)“

Hierbei ist zu betrachten, dass weitere, schon installierte Instrumente zur Einschätzung des Verhaltens in die Wahl der Zeiträume miteinbezogen werden:

> „L4: Ja, also wir haben ja zum Beispiel bei Nariel, also sie hat schon einen Sonnenplan am laufen, wo ich dann auch wirklich am Ende des Tages auf der Ampel gucken kann, ist irgendwas passiert, sodass man auch denke ich nen guten Überblick hat, sonst ist man ja leicht so, dass man was vergisst, wenn irgendwie in der ersten Stunde was war, aber ich denke, dass man das, dass ich das auch am Ende gut machen kann (…).
> M: Gut, das heißt du koppelst das im Grunde mit eurem Verstärkerplan dann.
> L4: Ja. (1.1)“

Weitere Aussagen beziehen sich auf die positiv wahrgenommene Anpassung von Beobachtungszeiträumen.

> „L2: Bezogen auf Luca war es so, dass es schon mal generell aber von Vorteil ist, dass wir die Beobachtungszeiträume heruntergeschraubt haben, also wirklich erst mal nur Arbeitsphasen, dass wir wirklich uns kleinere Ziele gesteckt haben. (3.3)“

> L6: Also ich habe sowohl bei Patrick, als auch bei David immer den Freitag rausgelassen, da habe ich das nicht bewertet, einfach weil wir freitags, wir kommen in die Schule, machen Erzählkreis, dann Frühstücken wir und dann haben wir Faustlos. Also es ist keinerlei Arbeitsphase, die haben dann noch Kunst und ich wollte Lehrerin XY das nicht aufs Auge drücken und deswegen habe ich gesagt, der Freitag ist einfach -da wären die Kreuzchen immer total positiv, weil ich da nichts bewerten könnte eigentlich. (8.3)

Die hypothesenablehnenden Aussagen beziehen sich auf zu große oder falschgewählte Beobachtungzeiträume, in denen das Verhalten nicht ersichtlich oder die Gesamtsituation zu unübersichtlich zum Bewerten ist. L2 äußert sich hinsichtlich des zu groß gewähltem Zeitraum wie folgt:

> „L2: je nachdem zu welchen Zeitpunkten ich dann die Beobachtungen schwerpunktmäßig mache, denke ich, ist dieser Beobachtungszeitraum, den wir gewählt haben vielleicht zu groß über den gesamten Schultag (…). (4.2)“

L8 beschreibt ein Problem, das Verhalten im gewählten Zeitraum zu beobachten.

> L8: Joa, also ich bin prinzipiell schon gut damit zurechtgekommen, ich hatte dann manchmal ein Problem damit, da ich mich auf die einzelnen Stunden festgelegt hatte, dass ich nicht immer, je nachdem was wir dann gemacht haben, es so ankreuzen konnte, wie es eigentlich in Arbeitsphasen ist. Also wenn wir jetzt gerade Kunst hatten wegen des Laternenumzugs, oder jetzt gerade im Sachunterricht experimentiert haben, oder so, dann war das nicht unbedingt der Schwerpunkt, den ich eigentlich beobachten will. (...) (9.2)

Es ist jedoch darauf hinzuweisen, dass alle Lehrkräfte es schaffen die Beobachtungszeiträume so anzupassen, dass sie mit ihrer Förderung im Folgenden zufrieden sind, wie beispielsweise das Folgegespräch mit L8 zeigt.

> L8: Ja. Wenn ich das in der Arbeitsphase mache, das ist für mich immer noch am besten, einen ganzen Tag kann ich das nicht im Blick behalten. (9.3)

Auf diese Weise kann die Hypothese im Sinne der quantitativen und qualitativen Analyse bestätigt werden. Bei der Implementierung scheint es jedoch sinnvoll darauf hinzuweisen, möglicherweise zunächst verschiedene Beobachtungszeiträume auszuprobieren, um geeignete Zeiträume für die Beobachtung herauszufinden.

Kritisch muss hier auf die Codierung der Unterkategorie hingewiesen werden. Es ist schwierig, die Erstwahl der Beobachtungszeiträume als hypothesenbestätigende Aussage zu codieren, da diese keine „Zufriedenheit“ ausdrücken kann. Da jedoch der gewählte Zeitraum in Folgeaussagen nicht unbedingt ersichtlich wird, wurden Aussagen der Wahl des Zeitraums mitcodiert und in Abhängigkeit davon, ob dieser in Folgegesprächen als „Hypothese bestätigend“ oder „Hypothese ablehnend“ codiert. Zudem ist

darauf hinzuweisen, dass die Lehrkräfte das Verhalten jeweils einmal täglich einschätzen sollten, was nur von einer Lehrkraft als ungeeignet eingeschätzt wurde. Diese Voraussetzung schränkt jedoch die Flexibilität der Auswahl an Beobachtungszeiträumen unter Umständen ein, da ein Beobachtungszeitraum „in den Sportstunden" so beispielsweise nicht gewählt werden konnte.

(7) Geeignete Wahl der Verhaltensbereiche. Die Hypothese zu der Unterkategorie lautet: „Die Lehrkräfte wählen für sie geeignete Verhaltensbereiche".

Quantitativ lassen sich die Ergebnisse (Tabelle 8) wie folgt darlegen. Wieder konnten allen sieben Lehrkräften Aussagen zu der Hypothese zugeordnet werden. Dabei wurden von allen Lehrkräften hypothesenbestätigende Aussagen codiert, von zwei Lehrkräften zusätzlich hypothesenablehnende Aussagen und von vier Lehrkräften uneindeutige Aussagen. Die Anteile hypothesenbestätigender Aussagen je Lehrkraft zeigen sich in dieser Unterkategorie durch folgende Werte: L2: 50%; L3: 100%, L4: 80%, L5: 50%, L6: 100%, L8: 67%, L9: 67%. Insgesamt liegen die Anteile hypothesenbestätigender Anteile bei über 50%, zwei Lehrkräfte erreichen sogar 100% Anteile hypothesenbestätigender Anteile. Dies führt zu einer Bestätigung der Hypothese aus quantitativer Sicht. Es ist jedoch darauf zu verweisen, dass die Anzahl der codierten Aussagen je Lehrkraft wieder höchst unterschiedlich sind (minimal 3, maximal 10 Aussagen).

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2, L3 - L6, L8, L9)	n=7	n=2	n=4
$L2_{Anzahl\ Aussagen}$	n=2	n=0	n=2
$L3_{Anzahl\ Aussagen}$	n=3	n=0	n=0
$L4_{Anzahl\ Aussagen}$	n=8	n=2	n=0
$L5_{Anzahl\ Aussagen}$	n=4	n=2	n=1
$L6_{Anzahl\ Aussagen}$	n=3	n=0	n=0
$L8_{Anzahl\ Aussagen}$	n=2	n=0	n=1
$L9_{Anzahl\ Aussagen}$	n=2	n=0	n=1

Tabelle 8 „Geeignete Verhaltensbereiche"

Hypothesenbestätigende Aussagen zeigen häufig die Charakteristik, dass Lehrkräfte den Kindern ohne Probleme Verhaltensbereiche für die Beobachtung zuordnen können oder zugeordnete Verhaltensbereiche beibehalten wollen. Beispielhaft wird dies an den Aussagen von L4 dargelegt.

> „L4: Also auf jeden Fall Impulsivität, klar haben auch beide Kinder in anderen Bereichen, aber ich denke, das ist, ja. (2.1)"
>
> „L4: Ja, ich würde auch hier beides weiter bewerten. (2.2)"

„M: Gut, dann ist die zweite Frage, möchtest du beide Beobachtungsbereiche weiter bewerten, ist das gut für dich?
L4: Ja, das ist gut machbar. (2.3)“

Zudem wurden Aussagen codiert, in denen die Lehrkräfte äußern, durch die Auswahl mehrerer Verhaltensbereiche mögliche unpassende Bereiche später wegfallen lassen zu können, um so den Umgang mit den Verhaltensbereichen flexibler zu gestalten:

„L3: Und dann können wir ja eventuell sagen, oder du sagst, du reduzierst
L4: Genau, Mhm. (zustimmend)
M:Genau (2.1)“

„M: Also fokussierst du dich dann jetzt auf Impulsivität und lässt das andere weg?
L5: Ja. (6.2)“

„L6: Nein, war alles in Ordnung, ich habe ja jetzt dieses Mal nicht mehr das schulbezogene Verhalten oder das Trotzverhalten, ja Trotzverhalten war es bei Patrick, das habe ich ja nicht mehr gemacht, sondern wirklich nur noch die Unaufmerksamkeit habe ich bewertet. (8.4)“

Die codierten Aussagen beziehen sich zumeist darauf, dass die Lehrkräfte keine für das Verhalten der Schüler passenden Verhaltensbereiche finden können. Dies machen sie dabei an den zu beobachtenden Einzelitems fest, sodass in einem Einzelfall der Interviewer einen Verhaltensbereich vorschlägt.

„L4: Es passt aber nichts wirklich (…) Also zu der Problematik, die Nariel zeigt passt kein Überitem (1.1)“

„L5: Ja, ich gucke jetzt gerade, weil bei beidem, „schulbezogenem Verhalten“, als auch „Impulsivität“ passen zwei Sachen total gut und zwei Sachen total schlecht. (6.3)“

„M: Ja, also sonst würde ich euch empfehlen, vielleicht erst mal das Trotzverhalten zu probieren (5.2)“

Die als uneindeutig codierten Transkriptausschnitte zeigen zumeist Probleme bei der Entscheidung für einzelne Verhaltensbereiche, welche entweder durch viele passende Items oder durch wahrgenommene inhaltliche Doppelungen in den Items entstehen:

„L2: Also ich sehe eine gewisse Doppelung in den Inhalten. Also der Bereich, ob er sich ja zum Beispiel im Unterricht meldet, oder mitarbeitet, das ist im Grunde genommen, passt der nicht zu dem Bereich „richtet die Konzentration auf die Bearbeitung der Aufgabe“ und „bleibt ruhig am Platz“, weil das nimmt ja im Grunde im Bereich Unaufmerksamkeit noch mal weiter ab. Das heißt, ich würde vermuten, dass es reicht, wenn ich nur noch den Bereich „Unaufmerksamkeit“ ankreuze. (3.3)“

„M: Mhm. (zustimmend) Gut. Jetzt noch einmal ein Schritt zurück im Grunde, jetzt noch mal zu den Graphen, denn da sind wiederum mir ein paar Sachen aufgefallen. Erst mal hatten wir glaube ich im Gespräch festgelegt, wolltest du erst Impulsivität beobachten. Das war der Beobachtungsschwerpunkt, den wir am Ende besprochen hatten, das ist überhaupt nicht schlimm, dass du jetzt schulbezogenes Verhalten genutzt hast, weil…

L3: Du hattest, „steht in Situationen auf, in denen Sitzen bleiben erwartet wird.“
M: Genau, das wird nämlich durch beide Items abgedeckt. (6.3)“

L8: Mhm (zustimmend). Also Emotionalität würde ich auf jeden Fall bei ihr nehmen, und Probleme in der Gruppe (...) Aber eigentlich auch schulbezogenes Verhalten. (...) Wenn ich hier die Items noch mal überflogen habe, ich glaube ich nehme Unaufmerksamkeit (...) Und schulbezogenes Verhalten (...). (10.1)

Bei einer Lehrkraft zeigten sich zudem Probleme in der Beantwortung der Multi-Item-Skalen. Sie beantwortete in der ersten Interventionsphase lediglich eines der dem Verhaltensbereich zugeordneten Items.

> „M: Doch wir haben die Baseline „schulbezogenes Verhalten“ da drunter. Aber eins ist natürlich mir aufgefallen, dass ihr euch jetzt auf ein Item fokussiert habt. Weil eigentlich war es ja so, ich weiß nicht, ob ich mich da richtig ausgedrückt hatte, dass ihr ja eigentlich trotzdem die anderen mitankreuzen sollt. Nicht nur ein Item. (6.3)“

Um die Handhabbarkeit des Ratings zu erhöhen und die Entscheidung der Lehrkräfte für oder gegen einen Verhaltensbereich zu erleichtern, könnten Items, welche in zwei Verhaltensbereichen erfragt werden aus einem Verhaltensbereich des Ratingbogens entfernt werden. Zudem könnte ein zweiter Ratingbogen mit anderen Items dazu führen, dass die Lehrkräfte für ihre Schüler geeignetere Verhaltensbereiche finden können. Zu bedenken ist jedoch hierbei, dass eine breitere Auswahl auch zu einer längeren Einarbeitungszeit in das Instrument führt.

Kritisch ist bei dieser Kategorie auch zu erkennen, dass sie nicht ganz trennscharf insbesondere zu der Unterkategorie „Itemformulierung“ ist, da die Lehrkräfte die Auswahl der Verhaltensbereiche meist mit einzelnen Formulierungen verknüpfen. Dies zeigt sich auch in der Auswahl von einzelnen Items, die für die Förderzielplanung berücksichtigt werden. Insgesamt konnte durch die vereinzelt auftretenden Probleme der Hypothese nicht vollständig zugestimmt werden. In diesem Bereich sind noch Nachbesserungen in Bezug auf die Handhabbarkeit notwendig.

(8) Geeignete Wahl der Rater. Um reliable Ergebnisse zu erlangen, müssen Direct Behavior Ratings von einer Person durchgeführt werden. In Bezug auf die Handhabbarkeit des Ratings muss es den Lehrkräften daher gelingen, eine Person zu bestimmen, die die Bewertung mittels Ratingbogen übernimmt. Hierzu wurde die Hypothese „die Lehrkräfte können eine Person bestimmen, die das Verhalten bewertet“ untersucht.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2, L3 - L6, L8, L9)	n=5	n=2	n=2
$L2_{Anzahl\ Aussagen}$	n=0	n=2	n=0
$L3_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L4_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L5_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L6_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L8_{Anzahl\ Aussagen}$	n=2	n=0	n=0
$L9_{Anzahl\ Aussagen}$	n=0	n=0	n=1

Tabelle 9 „Wahl der Rater“

Die quantitativen Auswertungsergebnisse können Tabelle 9 entnommen werden. Insgesamt wurden den Transkripten aller Lehrkräfte Aussagen entnommen, die sich auf die Hypothese beziehen. Dabei wurden von fünf Lehrkräften hypothesenbestätigende Aussagen, von einer Lehrkraft eine hypothesenablehnende Aussage und von drei Lehrkräften uneindeutige Aussagen codiert. Insgesamt sind jedoch je Lehrkraft nur ein bis zwei Aussagen codiert worden, sodass die Berücksichtigung der prozentualen Anteile hypothesenbestätigender Aussagen am Anteil der Gesamtaussagen nur eingeschränkte Aussagekraft hat. Dennoch stellen sich die Werte wie folgt dar: L2: 0%, L3: 100%, L4: 100%, L5: 100%, L6: 100%, L8: 100%, L9 0%. Ein Großteil der Lehrkräfte scheint keine Probleme damit zu haben eine Person für das Rating zu bestimmen, sodass aus quantitativer Sicht der Hypothese zugestimmt werden kann.

Hypothesenbestätigende Aussagen beinhalten Aussagen zur Festlegung des Raters. Beispielhaft sind folgende Aussagen:

> „M: Ok. Und du beobachtest dann auch das Verhalten wieder.
> L5: Ja. (5.1)“

> „M: Ok. Sehr gut. Ähm. Dann bewertest du eben auch das Verhalten, ne?
> L6: Mhm. (zustimmend) (7.1)“

Betrachtet man die hypothesenablehnenden Aussagen lässt sich festhalten, dass diese Aussagen enthalten, bei denen das gemeinsame Ausfüllen des Bogens bei täglichen Ratings unumgänglich ist. So verweist L2 darauf, dass tägliche Beobachtungen aus Termingründen nur durch zwei Lehrkräfte erfolgen können.

> „L2: Du auf jeden Fall Dienstags und ich bin an den anderen Tagen dann wahrscheinlich eher drin in nächster Zeit, weil dann die Lernanfängeranmeldungen sind, wie oft L1 dann besetzt und da ist müssen wir dann gucken. (3.1)“

Ein Großteil der uneindeutigen Aussagen bezieht sich auf die Lehrkräfte, die regelmäßig im Team mit einer sonderpädagogischen Lehrkraft arbeiten. Diese wünschen neben der eigenen Einschätzung die Berücksichtigung der Einschätzung der sonderpädagogischen Lehrkraft, welche jedoch bei dem Rating nicht vorgesehen ist. Folgende Aussagen sind beispielhaft:

> „L4: Ja, ich würde das bewerten, auch in Rücksprache mit L3, aber hauptsächlich ich, weil ich ja die meisten Stunden da bin. (1.1)“

> „L9: Dürfen wir das zusammen machen? Muss ich das alleine machen? Ich muss das alleine machen? (…) Ich bin ja da.
> L3: Wir könnens ja auch zusammen machen, nur ich seh ihn ja immer nur eine Stunde, ich seh ihn nicht den ganzen Tag. Wir könnens, also ich komme gerne, um das mit dir zu machen, aber – wir könnens auch auf jeden Fall die ersten Male zusammen machen, oder ist das schlecht? (...)

> L9: Ich kann ja auch, könnte ja theoretisch auch äh ankreuzen und sagen, hör mal, siehst du das genauso, oder kann man das nicht? (11.1)“

Bei der Festlegung des Raters scheinen also die umgebenden Faktoren eine große Rolle zu spielen. Es ist daher wichtig, bei der Implementierung eine größere Flexibilität der Beobachtungszeiträume aufzuzeigen, sodass das Verhalten nicht unbedingt täglich, sondern beispielsweise zwei bis drei Mal wöchentlich beurteilt werden kann. Wichtig ist jedoch eine Intervallskalierung auf der X-Achse bei der Graphenauswertung beizubehalten (tageweise, etc.) und eine Mindestanzahl an Messungen für die Baselineerhebung festzuhalten. Es scheint jedoch nicht immer möglich das Intervall einer Bewertung je Tag einhalten zu können. Des Weiteren ist darauf hinzuweisen, dass die Ratings immer subjektive Einschätzungen darstellen dürfen und sollen. Den Lehrkräften kann so die Angst vor „falschen“ Einschätzungen, wie sie den uneindeutigen Aussagen zu entnehmen ist, gegebenenfalls genommen werden. Insgesamt zeigt jedoch auch die qualitative Analyse keine großen Probleme bei der Festlegung von Ratern, sodass der Hypothese in Bezug auf die Handhabbarkeit zugestimmt werden kann.

Allgemein zeigt sich bei der Kategorie „Bewertung“ ein uneinheitlicheres Bild als bei der Kategorie „Auswerten“. So wurden die Hypothesen der Unterkategorien „Direktheit“, „geeignete Wahl der Zeiträume“ und „Wahl der Rater“ bestätigt. Probleme zeigen sich in der Handhabbarkeit des Ratings in den Unterkategorien „Itemformulierung“ und „Wahl der Verhaltensbereiche“. Diese ließen sich zum Teil durch möglicherweise ungeeignete Items und Doppelungen im Ratingmaterial erschließen, zudem zeigt sich eine enge Verknüpfung von Fördermaßnahmen zu einzelnen Items, anstatt zu Multi-Item-Skalen. Auf diese Weise ist die Eignung des Skalenformats der Multi-Item-Skalen für die Kombination mit Fördermaßnahmen kritisch zu hinterfragen.

6.3 Kategorie „Einstellungen“

Die Kategorie „Einstellungen“ bezieht sich auf Äußerungen, die die Eignung des Instruments aus subjektiver Lehrerperspektive betreffen. Es wurden die fünf Unterkategorien: (9) Ökonomie, (10) Eignung für den schulischen Alltag, (11) Zufriedenheit mit diagnostischer Qualität, (12) Kommunikation und (13) Erwartungen gebildet. Die Ergebnisse werden nun dargelegt.

(9) Ökonomie. Die der Unterkategorie zugeordnete Hypothese lautet: „Die Lehrkräfte empfinden die Methode als ökonomisch." Grundlegend ist die Annahme, dass das Instrument nicht dauerhaft Anwendung findet, wenn es für die Lehrkräfte in keinem guten Kosten-Nutzen-Verhältnis steht. In die Analyse wurden Aussagen einbezogen, die sich auf die Handhabbarkeit des Instrumentes beziehen.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2 - L6, L8, L9)	n=2	n=3	n=5
$L2_{Anzahl\ Aussagen}$	n=3	n=1	n=0
$L3_{Anzahl\ Aussagen}$	n=0	n=2	n=2
$L4_{Anzahl\ Aussagen}$	n=0	n=0	n=3
$L5_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L6_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L8_{Anzahl\ Aussagen}$	n=0	n=0	n=2
$L9_{Anzahl\ Aussagen}$	n=0	n=0	n=2

Tabelle 10 „Ökonomie"

Die quantitativen Ergebnisse (Tabelle 10) zeigen, dass wieder von allen Lehrkräften Aussagen zu der Hypothese gewonnen werden konnten. Die Anzahl der Aussagen schwankt dabei zwischen einer und vier Aussagen je Lehrkraft. Insgesamt konnten von zwei Lehrkräften hypothesenbestätigende, von drei Lehrkräften hypothesenablehnende und von fünf Lehrkräften uneindeutige Aussagen gewonnen werden. Von einer Lehrkraft stammen zugleich hypothesenbestätigende und -ablehnende Aussagen. Dies deutet auf sehr schwankende oder differenzierte Einschätzungen von Lehrkräften zu einzelnen Aspekten des Ratings hin, was die genauere qualitative Analyse interessant macht. Die Anteile hypothesenbestätigender Aussagen an den Gesamtaussagen ist dementsprechend gering: L2: 75%, L3: 0%, L4: 0%, L5: 0%, L6: 50%, L8: 0%, L9 0%. Den Anteilen entsprechend müsste die Hypothese für einen Großteil der Lehrkräfte eindeutig abgelehnt werden. Das Ratinginstrument wird entgegen theoretischer Annahmen nicht als ökonomisch eingeschätzt.

Betrachtet man die hypothesenbestätigenden Aussagen, beziehen sich diese zum einen auf die Handhabbarkeit nach dem Weglassen zuvor gewählter Verhaltensbereiche:

> „L2: Und dadurch, dass wir das immer weiter heruntergeschraubt haben und den Fokus auf einzelne Bereiche gelegt haben, ist das jetzt besser zu handhaben, deutlich besser zu überschauen, wie die Veränderung ist und wie sich auch Veränderungen in der Methode, die wir gerade anwenden auf das Verhalten auswirken. (3.4)"

Zum anderen wird von einer Lehrkraft das Bewerten der Items als ökonomisch eingeschätzt.

> „L6: (…) es ging immer sehr, sehr schnell und man konnte das mal eben in der Frühstückspause machen. (8.4)"

Schon in diesen Aussagen wird jedoch klar, dass sie sich lediglich auf einzelne Aspekte des Ratings beziehen (Bewertungssituation), oder zunächst ein „Herunterarbeiten“ notwendig war, um das Instrument handhabbar zu machen.

Die uneindeutigen und hypothesenablehnenden Lehreraussagen sind nicht ganz trennscharf zu betrachten und beziehen sich auf verschiedene Einzelaspekte, die als nicht ökonomisch dargelegt werden. Zum einen wird die Arbeit mit individualisierten Fördermethoden und -zielen für verschiedene Kinder einer Klasse als nicht leistbar angesehen. Dies zeigen insbesondere folgende Aussagen:

> „L3: Ja, es muss ja auch noch machbar sein. Wir haben hier ja auch andere Kinder, die irgendwie noch gefördert werden und du kannst ja nicht noch zehn Methoden da mischen. (2.2)“
>
> „L4: Genau. Nur halt mit diesem hin- und herswitchen der Punkte, ich fands einfach schwierig, verschiedene Methoden gleichzeitig durchzuführen. (2.3)“
>
> „L9: Ja, es geht auf jeden Fall nicht, wenn man alleine ist. Wenn ich wirklich den ganzen Tag alleine wäre, ginge das nicht, das alles durchzuführen. Also immer die Smileys zu kleben und also das geht nicht. (11.4)“

Des Weiteren zeigt sich eine Argumentation, die sich allgemein auf die Implementierung von Methoden in der Schule bezieht, indem auf Schwierigkeiten, die Methode kontinuierlich weiterzuführen verwiesen wird.

> „L5: Ja, das sind Dinge, die sowieso immer schwierig sind, egal welches Instrument man nimmt. Dass man halt immer wirklich dabeibleiben muss und sich die Zeit nehmen muss und sich hinsetzen muss und das konsequent durchziehen muss und ja, das ist halt die Schwierigkeit, weil immer so viel noch Nebenbei ist, aber das ist egal, welches Instrument das ist, weil die Schwierigkeit besteht ja immer. (5.4)“

Die Lehrkräfte verweisen auch auf die quantitative Begrenzung der Methode, eine Durchführung der Ratings mit mehr als zwei Kindern scheint für sie nicht durchführbar.

> „L8: Also es ist schon sehr, ich habe es ja erst versucht mit zwei Schülern, für zwei Teilbereiche, das war zu viel und dann für zwei mit einem, das geht. Aber dann hörts eben auch schon auf. Also die Zeit, die man sich eben genommen hat, da musste man sich schon auch erinnern, das eben zu tun, schleift sich bestimmt eben auch ein, aber wenn es jetzt noch mehr gewesen wäre, hätte ich es schwierig gefunden. (10.4)“
>
> „L9: Was ich allerdings schwierig finden würde, wenn ich das mit 24 Kindern jetzt machen müsste. (11.4)“

Insgesamt kann der Hypothese „Die Lehrkräfte empfinden die Methode als ökonomisch“ auf diese Weise nicht zugestimmt werden. Bezieht man die Aussagen auf die Annahmen in der Theorie, dass die individuelle Förderdiagnostik mittels Direct Behavior Ratings in einem Response-to-Intervention-Modell für etwa 20% einer Klasse geeignet sei, wird dies durch unzureichende personellen und zeitlichen Ressourcen mit

diesem Ratingbogen und an diesem Schulsystem aus Sicht der Lehrkräfte nicht möglich sein. Mögliche Verbesserungen für Implementierungen des Ratings bestünden beispielsweise in einer stärkeren Verknüpfung der für die Förderung gewählten Interventionen mit leicht individualisierbaren vorhandenen Fördermethoden wie dem Verstärkerplan, der sich leicht mit individuellen Verhaltensverträgen verknüpfen lässt. Zudem könnten die Lehrkräfte darauf hingewiesen werden, sich einen „Erinnerungstimer" für die Bewertung des Verhaltens zu stellen, um die Bewertung nicht zu vergessen.

Kritisch zu betrachten ist, dass die differenzierten Aussagen der Lehrkräfte nicht eindeutig in hypothesenablehnende und uneindeutige Aussagen eingeordnet werden können. Es zeigt sich, dass die Unterkategorie möglicherweise noch in weitere Teilaspekte, die sich genauer auf die Einschätzung einzelner Bereiche des Ratings bezieht aufgeteilt hätte werden müssen.

(10) Eignung für die Weiterarbeit. Den Lehrkräften wurde im letzten „Planungsgespräch" die Frage gestellt, ob sie sich vorstellen könnten, die Methode für sich weiter nutzen zu wollen. Die Frage soll Aufschluss darüber geben, ob die Lehrkräfte das Instrument für sich als hilfreich empfinden. Die Auswertung der Frage wurde durch die Hypothese „Die Lehrkräfte äußern, die Methode weiter nutzen zu wollen" ermöglicht.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=6 (L2 – L4, L6, L8, L9)	n=4	n=1	n=3
$L2_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L3_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L4_{Anzahl\ Aussagen}$	n=1	n=1	n=0
$L6_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L8_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L9_{Anzahl\ Aussagen}$	n=1	n=0	n=1

Tabelle 11: „Eignung für die Weiterarbeit"

Tabelle 11 repräsentiert die quantitativen Ergebnisse der Analyse: Sechs Lehrkräfte äußerten sich zu der Hypothese. Hierbei ist festzuhalten, dass aus den Transkripten mit L5 ersichtlich ist, dass die Frage des Leitfadens, ob man sich vorstellen könne, mit der Methode weiterzuarbeiten, vom Interviewer (M) nicht gestellt wurde, sodass hier ein Fehler in der Durchführung erkennbar wird. Zu vier Lehrkräften lassen sich Aussagen, die die Hypothese bestätigen, einer Lehrkraft eine Aussage, die die Hypothese ablehnt und drei Lehrkräften Aussagen, die uneindeutig sind zuordnen. Die Anzahl der Aussagen schwankt zwischen einer und zwei Aussagen. Die geringe Anzahl der Aussagen

lässt sich dadurch erklären, dass sie sich auf die Beantwortung einer Frage des Leitfadens beziehen. Die Anteile hypothesenbestätigender Aussagen liegen bei den einzelnen Lehrkräften bei L2: 100%, L3: 0%, L4: 50%, L6: 0%, L8: 100%, L9: 100%. Demnach zeigt sich eine leichte Tendenz zur Bestätigung der Hypothese, wobei das Ergebnis insgesamt eher uneindeutig scheint.

Die Lehrkräfte, die der Hypothese zustimmen äußern, dass sie die Methode für sie und ihre Weiterarbeit durchaus in Betracht ziehen, was folgende Aussagen zeigen.

> „M: Also es ist halt die Frage, wenn du längerfristig damit arbeitest, also auch in deinem normalen Alltag, wenn ich weg wäre, so in Anführungsstrichen, ob dir das zu viel Arbeit wäre oder nicht?
> L9: Ne, also man sieht ja selber schon was, also ich habe ja selber gesehen, dass sich da was geändert hat, als ich das angekreuzt habe.
> M: ja. Ok, das heißt, du kannst es für dich schon vorher, auch ohne diese Visualisierung ein Stück weit als Feedback für dich nutzen?
> L9: Ja. (11.3)“

> „L8: Ich kann mir das schon vorstellen, vielleicht für einen, wenn ich irgendwelche Punkte habe, die mich sehr stören, oder die es dem Kind nicht möglich machen, vernünftig zu arbeiten, dann kann ich mir das schon vorstellen. Ja. (10.4)“

In der Aussage von L8 wird jedoch ersichtlich, dass die Anlässe und die Anzahl der Ratings stark eingegrenzt werden müssten. Dies führt auch zu der weiteren hypothesenablehnenden Aussage von L8:

> „L4: Es ist auf Dauer nicht leistbar und vor allem haben wir es jetzt für zwei Kinder gemacht, für mich wäre es nicht leistbar, das für mehrere Kinder auch durchzuziehen. (1.4)“

Die uneindeutigen Aussagen betonen ebenfalls die quantitative Eingrenzung der Methode auf Einzelfälle, die für eine Weiterarbeit notwendig wäre, was folgende Beispiele zeigen.

> „L3: Ja, also diese Beobachtungen lagen ja in erster Linie in eurer Hand, weil wir die Intervalle ja auch so gewählt haben, dass das für mich so fast nicht leistbar war, von daher ich könnte mir vorstellen das für die einzelnen Kinder, oder die schwierigen Kinder in den einzelnen Klassen auch wohl einzusetzen, oder wenn ich das Gefühl habe, es ist vielleicht auch gut, dass man ne Argumentationshilfe für die Eltern hat, dann könnte ich es mir auch gut vorstellen, ich müsste es ausprobieren, ich kann im Moment noch nicht so viel dazu sagen, ob ich es mir für alle Schüler vorstellen könnte, wenn das eben alles ein bisschen vereinfachter läuft, weil man es eben online machen kann. (11.4)“

> „L6: Ja, also wenn ich das – für höchstens zwei Kinder in der Klasse würde ich das machen. Ich würde nicht alles ausdrucken und aufschreiben, ich würde es dann direkt am Laptop machen, aber dafür müsste ich den Laptop direkt in der Klasse haben. (8.4)“

Die Betrachtung der Ergebnisse zeigt, dass die Hypothese nur bedingt bestätigt werden kann, wobei Gründe für die nichtmögliche Weiterarbeit in der Ökonomie angesiedelt werden. Es zeigt sich, dass die Methode den Lehrkräften als Instrument für die Verlaufsdiagnostik des Verhaltens einzelner Kinder geeignet erscheint, sie diese jedoch

nicht für mehr als zwei Schüler mit Verhaltensschwierigkeiten nutzen möchten. Dementsprechend kann das Instrument „PUTSIE“ nicht vollständig den theoretischen Ansprüchen, für die Diagnostik von etwa 20% der Kinder einer Klasse geeignet zu sein, gerecht werden.

Es muss jedoch kritisch berücksichtigt werden, dass die Durchführung der Erhebung für die Lehrkräfte sehr zeitintensiv gewesen sein mag, da je Schüler alle zwei Wochen ein Gespräch geführt und täglich das Verhalten bewertet werden musste. Wenn Ratings nicht täglich, sondern in weiteren zeitlichen Abständen und die Evaluation der Maßnahmen nicht so engmaschig implementiert werden, wären eventuell andere Aussagen möglich.

(11) Zufriedenheit mit diagnostischer Qualität. Die zugehörige Hypothese lautet „Die Lehrkräfte äußern sich positiv gegenüber „PUTSIE“ als Diagnostikinstrument. Es soll die Zufriedenheit mit den Ergebnissen des Ratinginstruments aus Sicht der Lehrkräfte erhoben werden.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2- L6, L8, L9)	n=7	n=2	n=3
$L2_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L3_{Anzahl\ Aussagen}$	n=1	n=1	n=1
$L4_{Anzahl\ Aussagen}$	n=2	n=0	n=0
$L5_{Anzahl\ Aussagen}$	n=1	n=1	n=0
$L6_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L8_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L9_{Anzahl\ Aussagen}$	n=1	n=0	n=0

Tabelle 12: „Zufriedenheit mit diagnostischer Qualität“

Betrachtet man die Ergebnisse der Analyse aus quantitativer Perspektive (Tabelle 12) zeigt sich, dass allen Lehrkräften Äußerungen zu der Hypothese zugewiesen wurden, ebenfalls allen Lehrkräften hypothesenbestätigende Aussagen. Dagegen zeigen nur zwei Lehrkräfte Äußerungen, die die Hypothese ablehnen und drei Lehrkräfte uneindeutige Aussagen. Je Lehrkraft wurden zwischen einer und drei Aussagen codiert. Die Anteile an hypothesenbestätigenden Aussagen im Vergleich zu den codierten Aussagen insgesamt zeigen sich wie folgt: L2: 100%, L3: 33%, L4: 100%, L5: 50%, L6: 50%, L8: 100%, L9 100%. Auf dieser Grundlage müsste der Hypothese eher zugestimmt werden. Die Betrachtung der codierten Aussagen soll genauer zeigen, ob dieser Ersteinschätzung zugestimmt werden kann.

Der Großteil codierter Aussagen bezieht sich auf die Frage „Welche Möglichkeiten und Grenzen birgt das Ratinginstrument für Sie?“. Zudem wurde einigen Lehrkräften

die Frage gestellt, ob sie das Instrument als geeignet für die Diagnostik von Schülerverhalten über die Zeit empfinden. Die Fragen wurden im vierten Planungsgespräch zum Ende der Erhebung gestellt.

Insgesamt zeigen die hypothesenbestätigenden Aussagen vielmals die Charakteristik, dass das Instrument von den Lehrkräften als Möglichkeit zur Fokussierung auf einzelne Verhaltensbereiche und deren Förderung zu betrachten. Dies zeigen folgende Aussagen:

> „Um einfach wirklich einen Fokus zu finden, weil pauschale Aussagen wirklich nicht weiterhelfen. Und in der Flut der Dinge, die man beobachten, fördern, machen muss, habe ich glaube ich auch durch dieses Runterfahren des zu beobachtenden Verhaltens und der Zeiträume gemerkt, dass man viel weiter unten und viel kleiner anfangen muss, das zu machen. (3.4)“

> „L6: Ja, ich sehe auf jeden Fall den Schüler. Weil sonst habe ich die Kinder ja irgendwie so im Blick. Aber so weiß ich genau, ich kann mich auf ein Kind, kann mich auf eine Sache konzentrieren und weiß genau das und werte das, oder du hast das ausgewertet, und sehe das einfach mal. Ich glaube das macht mir einfach mal bewusst, wie schwankend das einfach ist, oder jetzt einfach bei Patrick, wie sehr das hilft, was wir da gemacht haben. Man sieht es einfach auf den ersten Blick. (8.4)“

> „L8: Also ich finde schon gut, dass man den Fokus auf eine Sache legt, also sonst sagt man ja so was wie „Oh, der hat sich ja heute wieder total blöd benommen und hat wieder nicht gearbeitet.“ Und so pickt man sich schon ein Teilverhalten heraus, dass man ganz gezielt ansehen kann und auch gezielt beobachtet, dadurch, dass man es eben immer notiert und was man eben auch gezielt verändern kann, wie wir ja gesehen haben. (10.4)“

Des Weiteren wird die Belegbarkeit pädagogischer Diagnostik mittels der Verlaufsgraphen als positiv empfunden:

> „L4: Ja, ich finde anhand dieser Graphen oder dieses Veranschaulichen, hat man einfach etwas in der Hand, was man auch den Eltern zeigen kann. Etwas Messbares und nicht etwas, was wir subjektiv erzählen. (1.4)“

> „L5: Also die Möglichkeiten sehe ich schon darin, dass man das schon auch einfach noch mal schwarz auf weiß hat, was man sonst nur intuitiv vermutet und ja auch zu gucken, sich Gedanken darüber zu machen, was kann dem Kind auch helfen, sein Verhalten zu verbessern. Dass man halt auch immer wieder darüber nachdenkt. (5.4)“

Dennoch zeigen sich Schwierigkeiten durch hypothesenablehnende und uneindeutige Aussagen, die aus Lehrersicht die diagnostische Güte begrenzen. Zum einen betont ein Teil der Lehrkräfte, dass die Einschätzungen des Ratings letztendlich immer noch auf subjektiven Wahrnehmungen beruhen, was einer Diagnostik entgegenstünde. Dies belegt vor allem folgende Aussage:

> M: Gut und bei der Beurteilung jetzt, würdest du sagen, das ist wirklich ein Instrument wo du sagst, das ist wirklich geeignet um das Verhalten eines Schülers über die Zeit hinweg zu erfassen (…)? (…) L5: Es bleibt subjektiv und deshalb weiß ich nicht, inwiefern das für eine richtige Diagnose dann geeignet ist. (5.4)

Zudem bezieht sich eine Lehreraussage explizit auf die Formulierung der Items. Für sie ist das Rating erst dann geeignet, wenn passendere Items für die Erfassung des Schülerverhaltens gefunden würden.

> L3: Was ich mir eben wünschen würde, wäre auch, was mir eben denke ich auch, was ist, was euch Schwierigkeiten macht, ist die Zuordnung zu den Items, also das wäre dann aber ne Sache die dann sich eher auf das Instrument – eine Hoffnung, dass man treffendere Items noch entwickeln kann auf Dauer daraus. (2.1)

Zwei Lehreraussagen beinhalten zudem die Einschätzung, dass das Diagnostikinstrument zwar für spezifische Verhaltensweisen (die Verhaltensbereiche) geeignet sei, jedoch nicht für einen Überblick des Gesamtkonstrukts Verhalten:

> „M: Ok. Und wie schätzt du, es ist ja ein Diagnostikinstrument eigentlich, ist es für dich ein ausreichendes Diagnostikinstrument um das Verhalten der Schüler zu messen oder nicht?
>
> L8: Nein, dafür ist es viel zu spezifisch. Also wenn ich mir überlege, was ich beobachten kann, habe ich ja gerade gesagt, dann kann ich mir einen Teilbereich des Verhaltens rausnehmen, beobachten und fördern, aber das ist ja nur ein Teilbereich dessen, also das Große und Ganze muss ich ja trotzdem sehen. (10.4)“

Es zeigt sich insgesamt, dass die Lehrkräfte es schaffen, die Grenzen des Diagnostikinstrumentes zu erkennen, was durch die ablehnenden und uneindeutigen Aussagen repräsentiert wird. Dennoch ist zu erkennen, dass jede Lehrkraft Zufriedenheit mit der diagnostischen Güte für die individuelle Förderung zeigt. Aus diesem Grund wird die Hypothese als bestätigt für das Ratinginstrument angesehen. Es ließe sich bei einer weiteren Implementierung jedoch darauf hinweisen, dass jedes diagnostische Instrument im Bereich des Verhaltens letztlich auch auf subjektiven Einschätzungen und Wahrnehmungen beruht. Dies wird auch aus der theoretischen Darstellung von Instrumenten zur Verhaltensdiagnostik in Kapitel 2.3 deutlich.

(12) Kommunikation. Mit der Unterkategorie wurde die Hypothese „Die Lehrkräfte können sich vorstellen die Graphen für die Kommunikation über das Verhalten der Schülerinnen und Schüler zu nutzen“ verknüpft.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2- L6, L8, L9)	n=5	n=0	n=3
$L2_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L3_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L4_{Anzahl\ Aussagen}$	n=1	n=0	n=1
$L5_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L6_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L8_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L9_{Anzahl\ Aussagen}$	n=1	n=0	n=0

Tabelle 13: „Kommunikation“

Alle Lehrkräfte haben sich zu der Hypothese geäußert. Hiervon zeigen fünf Lehrkräfte hypothesenbestätigende Äußerungen. Von keiner Lehrkraft wurden hypothesenwiderlegende Äußerungen codiert. Drei Lehrkräfte zeigen uneindeutige Aussagen. Zu jeder Lehrkraft wurden ein bis zwei Äußerungen codiert, was sich darauf zurückführen lässt, dass sich die Äußerungen auf die sinngemäße Frage, ob man sich vorstellen könne, die Ratingergebnisse für die Kommunikation mit an der Förderung beteiligten Partnern zu nutzen bezieht. Die Frage wurde ebenfalls im letzten Gespräch der Erhebung gestellt. Die Prozentualen Anteile hypothesenbestätigender Aussagen an den Gesamtaussagen der Lehrkräfte zeigt sich wie folgt: L2: 100%, L3: 100%, L4: 50%, L5: 0%, L6: 0%, L8: 100%, L9: 100%. Laut den prozentualen Anteilen wird die Hypothese eher bestätigt werden.

Die codierten Äußerungen, die als die Hypothese bestätigend zugeordnet wurden zeigen alle die Charakteristik, dass mit dem Rating die Möglichkeit verbunden wird, die eigenen subjektiven Aussagen mittels Daten zu fundieren. Beispielhaft wird dies an folgenden Aussagen von L8 und L9 deutlich:

> „L8: Ja doch, vielleicht unterstützt es schon das, was ich sage, wenn ich über einen Schüler spreche. Ja, doch könnte ich mir schon vorstellen. (10.4)“

> „L9: Nein, ich würde das mit reinnehmen, ich würde das auch zeigen. Daran lässt es sich ja auch erkennen. Das ist noch mal etwas ganz anderes, als wenn ich das sage. Wenn ich das zeigen kann, dass das so ist, untermauert man ja seine Aussage. (11.4)“

Als primäre Zielgruppe für die Kommunikation mittels Ratingergebnissen werden vor allem Eltern genannt:

> „L3: Und es ist eben ein Argument, dass man irgendwann den Eltern gegenüber bei Schwierigkeiten noch mal aufmachen kann. (11.4)“

> „L4: Ja, ich finde anhand dieser Graphen oder dieses Veranschaulichen, hat man einfach etwas in der Hand, was man auch den Eltern zeigen kann. Etwas Messbares und nicht etwas, was wir subjektiv erzählen. (1.4)“

Bei einigen Lehrkräften wird diese Einschätzung jedoch eingeschränkt. Dies leitet die Betrachtung der uneindeutigen Aussagen ein. So wird darauf verwiesen, dass das Verstehen der Verhaltensgraphen von einigen Eltern unter Umständen nicht möglich ist. Hier scheinen unter anderem sprachliche Barrieren eine Rolle zu spielen, wie folgende Aussagen zeigen:

> „L6: Also es kommt darauf an, bei welchem Kind. Bei den Eltern, die ich jetzt habe, würde das gehen, weil die Eltern meiner Meinung nach diesen Graphen verstehen würden. Es gibt aber auch leider genug Eltern bei uns an der Schule jetzt, die das einfach nicht verstehen würden, was wir damit jetzt gemacht haben. (8.4)“

> „L4: Ähm, jetzt unabhängig von den beiden Kindern könnte ich es mir bei einem gewissen Teil der Elternschaft vorstellen. Weil, wir haben hier ja eine Elternschaft, wo die Deutschkenntnisse auch sehr gering sind und ich finde man muss halt einfach ein Verständnis für diese Graphen zeigen. Ich könnte jetzt nicht jemandem diesen Graphen zeigen, der einfach nichts damit anzufangen hat. Deswegen finde ich das, wie ich die einsetzen kann bei der Elternschaft, die wir hier haben, gering. Aber unser Hintergedanke war auch schon, wenn jetzt zum Beispiel AO-SF-Anträge gestellt werden, dass man so etwas dann auch hat, also nicht nur als Bericht, sondern, dass man das dann auch für die AO-SF-Anträge nutzen kann. (1.4)“

Die Aussage von L4 schränkt die Nutzbarkeit der Methode zwar für Eltern ein, verweist aber auf die Möglichkeit, die Ergebnisse für die Erstellung von AO-SF-Anträgen zu nutzen. So wird die Methode auch für die Belegbarkeit sonderpädagogischer Förderbedarfe bei schulischen Entscheidungsträgern als möglich dargestellt. Insgesamt schätzen alle Lehrkräfte das Instrument zumindest als bedingt geeignet zur Kommunikation des Schülerverhaltens ein. Insbesondere wird die Belegbarkeit bestimmter Verhaltensweisen betont. Die Hypothese wird daher durch die Ergebnisse bestätigt. Es ist jedoch schon bei Beginn einer Implementierung stärker auf die Kommunikationsmöglichkeiten mittels des Instruments hinzuweisen, wie diese auch für Berichte zu nutzen.

Während der Erhebung zeigte sich zudem, dass die Möglichkeit der Kommunikation von den Ratingergebnissen mit außerschulischen Partnern zwar geäußert, eine tatsächliche Kommunikation der Ergebnisse mit Eltern oder anderen außerschulischen Partnern während des Erhebungszeitraums jedoch bei keinem beobachteten Schüler stattfand.

(13) Erwartungen. Die letzte untersuchte Kategorie geht zurück an den Anfang der Erhebung. Es sollte die Hypothese untersucht werden, ob die Lehrkräfte das Ratinginstrument von Anfang an mit der Erwartung einer besseren Diagnostik des Schülerverhaltens verknüpfen. Zur Erfassung der Erwartungen diente im Leitfaden die Frage, welche Erwartungen die Lehrkräfte mit dem Instrument am Ende des ersten Planungsgesprächs verknüpfen.

Lehrkräfte	**Hypothese bestätigend**	**Hypothese ablehnend**	**Aussage uneindeutig**
N_{gesamt}=7 (L2 - L6, L8, L9)	n=5	n=1	n=1
$L2_{Anzahl\ Aussagen}$	n=0	n=0	n=1
$L3_{Anzahl\ Aussagen}$	n=3	n=0	n=0
$L4_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L5_{Anzahl\ Aussagen}$	n=1	n=0	n=0
$L6_{Anzahl\ Aussagen}$	n=0	n=1	n=0
$L8_{Anzahl\ Aussagen}$	n=2	n=0	n=0
$L9_{Anzahl\ Aussagen}$	n=1	n=0	n=0

Tabelle 14: „Erwartungen“

Die codierten Aussagen zeigen die in Tabelle 14 dargelegte Struktur. Insgesamt wurden auch zu dieser Kategorie Aussagen jeder Lehrkraft zu der Hypothese codiert. Hiervon bestätigen 5 Lehrkräfte mit Aussagen die Hypothese, einer Lehrkraft wurde eine die Hypothese ablehnende Aussage zugewiesen und ebenfalls einer Lehrkraft eine uneindeutige Aussage zur Hypothese. Je Lehrkraft wurden ein bis drei Aussagen zugeordnet. Die prozentualen Anteile an hypothesenbestätigenden Aussagen zu den Gesamtaussagen liegen je Lehrkraft bei: L2: 0%, L3: 100%, L4: 100%, L5: 100%, L6: 0%, L8: 100%, L9: 100%.

Die hypothesenbestätigenden Aussagen formulieren zumeist Erwartungen der Lehrkräfte an eine bessere objektive Wahrnehmung des Verhaltens:

> „L3: Ja, also bei mir ist es eben so, dass ich auch denke, wie du das eben so gesagt hast, dass das es gut ist, das Ganze nen bisschen objektiver zu betrachten, dass nen bisschen abzulösen von den Gefühlen, oder Eindrücken (2.1)“

> „L5: Ja, also ich erhoffe mir natürlich schon einmal für mich son bisschen Klarheit zu kriegen, ähm oder klarer zu sehen, in welchen Situationen es tatsächlich dann so aufgetaucht ist, oder wie oft es dann aufgetaucht ist, weil man hat ja nur so ne Wahrnehmung, (5.1)“

> „L8: Also, da es zwei Schüler sind, mit denen ich ja Bereiche sehe, in denen eben Schwierigkeiten bestehen, dass aber eben Bereiche sind, die man oft so nebenbei wahrnimmt, also ich weiß, dass es da Schwierigkeiten gibt, aber ich erhoffe mir da durch die Dokumentation vielleicht ne, ne gewisse Gesetzmäßigkeit oder Struktur zu entdecken, um da dann eben ansetzen zu können, also es ist sonst eben oft so, es fällt einem auf und ich weiß es auch aber wirklich ansetzen kann man im Schulalltag nicht und da denk ich mir wenn man hier so was aber ganz genau dokumentiert, bestimmte vielleicht Muster deutlich werden und ich weiß irgendwie in der und der Situation passiert das, weil… (10.1)“

> „L9: Und auch vielleicht so, ich habe jetzt das Gefühl das trifft alles zu, aber vielleicht ist das auch was, das ganz doll heraussticht und vielleicht irgendwas wo ich denke, das macht er auch nicht, das ist aber gar nich, also so. (11.1)“

Zudem erhoffen einige Lehrkräfte, wie auch L8 oben, Gesetzmäßigkeiten oder Strukturen des Verhaltens genauer wahrnehmen zu können. Es ist in Frage zu stellen, ob das Instrument dies durch die Graphen ermöglichen kann, da hier lediglich die wahrgenommene Intensität des Verhaltens über die Zeit abgebildet wird.

Die uneindeutige Aussage bezieht sich auf eine allgemeine Verbesserung des Verhaltens, wobei hier in einem Nachsatz differenziert wird, dass diese durch die Interventionsmaßnahmen erreicht werden sollen, die durch das Rating evaluiert werden:

> „L2: Ich habe ja natürlich schon die Hoffnung, dass sich da was ändert (…) Durch die Kreuze nicht, aber durch die Interventionsmaßnahmen. Die sind ja da um zu gucken, inwiefern er drauf anspringt. (4.1)“

Eine Aussage richtet sich wiederum nur auf die Erwartung besserer Lernergebnisse. Diese wurde als hypothesenablehnende Aussage codiert, da hier kein Bezug auf das Rating als Diagnostikinstrument zu erkennen war.

> „L6: Also was ich mir jetzt so wünsche, was da jetzt so rauskommen soll, also bei David wünsch ich mir einfach, das, dass es irgendwie einfach eine Möglichkeit gibt, wie ich dieses Kind mit gutem Gewissen an die weiterführende Schule schicken kann, dass er lernt, selbstständig zu arbeiten und nicht, dass ich alle zehn Minuten sagen muss, David arbeite bitte! (8.1)“

Insgesamt wird gefolgert, dass von einem Großteil der Lehrkräfte das Instrument mit realistischen Erwartungen begegnet wurde. Daher kann der Hypothese abschließend zugestimmt werden. Es ist dennoch wichtig, den Lehrkräften bei der Implementierung die diagnostischen Funktionen des Instruments (Erfassung des Problemverhaltens über die Zeit, Evaluation der eigenen Intervention) genau aufzuzeigen, um falsche und zu hohe Erwartungen an eine gelingende Förderung oder dem Erkennen von „Verhaltensgesetzmäßigkeiten“ zu verhindern. Kritisch ist bei der Unterkategorie zu bedenken, dass die Erwartungen zwar zu Beginn abgefragt wurden, jedoch darauf verzichtet wurde, am Ende der Erhebung Rückbezug auf die Erwartungen zu nehmen und zu fragen, ob sich diese erfüllt haben.

Abschließend lässt sich festhalten, dass die Unterkategorien zu „Einstellungen“ nahezu alle bestätigt werden können. Es zeigt sich lediglich bei der Kategorie „Eignung für die Weiterarbeit“ ein uneinheitliches Bild, indem darauf verwiesen wird, dass das Ratinginstrument sich nicht für den Einsatz mit vielen Schülern eignet und zudem das Ratingmaterial besser aufbereitet werden könnte.

Vergleicht man die Kategorien miteinander lassen sich daher grob einige Schwächen des Ratings identifizieren, welche sich über die Kategorien hinweg finden lassen. Diese betreffen zum einen Aspekte der Ökonomie, da immer wieder darauf hingewiesen wird, dass das Rating im Aufwand während der Durchführung nicht durchgeführt werden könne und zu viel Zeit und Ressourcen koste. Hier muss bei der Implementierung insbesondere darauf geachtet werden, schon vorhandene Interventionen der Schule zu nutzen und auf ökonomische Beobachtungszeitpunkte und -zeiträume zu achten. Zum anderen zeigen sich Schwächen im Ratingmaterial, insbesondere bei der Graphendarstellung und der Itemformulierung. Die Graphendarstellung muss einfacher erfolgen. Es muss klar ersichtlich sein, ob ein steigender Graph eine positive oder negative Verhaltensentwicklung bedeutet. Dies gilt ebenso für den Ratingbogen. Bei der Itemformulierung wünschen sich die Lehrkräfte geeignetere Items und benötigen

insbesondere Hinweise, wie mit den Multi-Item-Skalen verfahren werden soll. Die klinischen Items der DSM-V scheinen teilweise nur begrenzt geeignet. Genauere Vorschläge können den einzelnen Unterkategorien entnommen werden.

7. Schwächen und Stärken

Nach der Darstellung der Ergebnisse sollen an dieser Stelle kurz zunächst die Schwächen dann die Stärken in der Durchführung und Planung der Erhebung dargelegt werden.

Allgemein wurde die Erhebung nicht im Kontrollgruppendesign durchgeführt. Auf diese Weise kann kein Vergleich der Ergebnisse mit anderen Schulsystemen erfolgen. Zudem wird davon ausgegangen, dass durch die Voraussetzungen des Schulsystems auch die Ergebnisse eingeschränkte Gültigkeit besitzen und immer vor dem Hintergrund der in Kapitel 4 dargelegten Rahmenbedingungen betrachtet werden müssen.

Weitere Schwächen zeigen sich im Material für die Durchführung. Es muss darauf hingewiesen werden, dass noch keine Überprüfung der Testgütekriterien zum Ratinginstrument „PUTSIE“ durchgeführt wurde, sodass keine Aussage zu dessen Eignung aus psychometrischer Sicht gemacht werden kann. Zusätzlich sind in den Leitfadengesprächen (Planungsgesprächen) viele Erläuterungen, Erklärungen und Tipps durch den Moderator (M) zu erkennen, die die Aussagen der Lehrkräfte möglicherweise beeinflussen. Dies lässt sich jedoch durch das Verständnis der Interviewsituation als „Planungsgespräch“ erklären, in dem auf Fragen seitens der teilnehmenden Lehrkräfte eingegangen wird und die Gesprächssituationen ebenfalls zur Implementierung der Methode dient. Ebenfalls ist darauf zu verweisen, dass durch den Interviewer im Sinne der Stichprobenpflege eine Darstellung möglicher evidenzbasierter Fördermöglichkeiten erfolgen sollte, sodass erhöhte Redeanteile, gerade auch bei der Förderplanung vertretbar sind.

Durch die Kenntlichmachung der Erstellung der Ratingmaterialien durch den Autor wird zudem darauf verwiesen, dass ein Antwortverhalten der Lehrkräfte im Sinne sozialer Erwünschtheit nicht ausgeschlossen ist.

Einschränkungen betreffen zudem die Auswertung der Transkripte mittels der integrativen Inhaltsanalyse. Zum einen muss festgestellt werden, dass das erarbeitete Kategoriensystem nicht trennscharf ist und sich verschiedene Kategorien inhaltlich überlappen. Dies führte teilweise zu Problemen bei der Codierung. Zum anderen zeigt sich

ein Problem bei der Auswahl geeigneter Textstellen häufig darin, dass der Sinn von Äußerungen erst aus dem Gesamtkontext des Gesprächs ersichtlich wird. Es konnte hierbei die Reliabilität der Codierung nicht untersucht werden, da das Material nur von einer Person codiert wurde. Eine Probecodierung des Materials mit mehreren Codierern, um zu untersuchen ob gleiche Transkriptstellen den Kategorien zugeordnet werden wäre notwendig.

Auch die Auswertung muss kritisch hinterfragt werden. Die integrative Inhaltsanalyse dient dazu Textmaterial zu analysieren und auf Merkmale im Sinne der Bestätigung oder Ablehnung von Hypothesen zu untersuchen. Durch die Einteilung von Aussagen in die Auswertungskategorien „Hypothese zustimmend", „Hypothese ablehnend" und „uneindeutige Aussagen" gehen Unterschiede in der Qualität der einzelnen Aussage verloren. Zudem sind die Redeanteile der einzelnen Lehrkräfte in der Erhebung ungleich verteilt, da Lehrkräfte ein bis zwei Schüler beobachten konnten und beispielsweise eine sonderpädagogische Lehrkraft (L3) an einer Vielzahl von Gesprächen teilnahm. Dies führt dazu, dass eine Deutung der quantitativen Ergebnisse nur eingeschränkte Gültigkeit besitzt. Außerdem muss festgehalten werden, dass differenzierte Stärken und Schwächen der Implementierung nur durch die qualitative Untersuchung der einzelnen Aussagen gewährleistet werden kann und aus quantitativen Vergleichen nur schwer ersichtlich wird, weshalb die Beschreibung der Aussagen auf qualitativer Ebene zu jeder Kategorie angefügt wurde. Die integrative Inhaltsanalyse allein ist nur eingeschränkt für die Beantwortung der Fragestellung geeignet.

Dennoch ist darauf zu verweisen, dass auch diese Darstellung der Schwächen für kommende eventuell größer angelegte Studien zur Untersuchung der Implementation von Direct Behavior Ratings genutzt werden kann.

Zudem zeigen sich auch einige Stärken der Erhebung. Die Erhebung erfolgte im Feld und ist durch offene, persönliche Gespräche geprägt gewesen, die den Transkripten entnommen werden können. Auch zeigt sich für eine qualitative Datenerhebung im Rahmen dieser Masterarbeit eine relativ große Stichprobe mit sieben Lehrkräften und einer großen analysierten Datenmenge (siehe Anhang F). Durch die Erhebung über einen langen Zeitraum von etwa zwei Monaten mit vier Gesprächen über die Zeit zeigt sich außerdem ein langer Zeitraum, in dem das Ratinginstrument erprobt wurde.

Der Interviewleitfaden und die zuvor festgelegten Rahmenbedingungen schränkten zwar die Flexibilität des Instruments ein, ermöglichten aber die Vergleichbarkeit der Implementierung und Gespräche.

Bei der Auswertung zeigt die Berücksichtigung von dreizehn Kategorien zwar teilweise Überschneidungen, dennoch konnte ein inhaltlich breites Spektrum zur Implementierung betrachtet werden. Insbesondere die qualitative Analyse der Kategorien ermöglichte Erkenntnisgewinne zu Verbesserungsmöglichkeiten und Schwächen im Ratingmaterial.

Abschließend ist darauf hinzuweisen, dass mögliche, zu berücksichtigende Schwächen und Stärken der Implementierung des Direct Behavior Ratings „PUTSIE“ durch die Betrachtung der Kategorien dargelegt werden konnten. Auf diese Weise wurden Erkenntnisse zur Handhabbarkeit und Einstellungen zum Ratinginstrument erkennbar, sodass die Forschungsfragen beantwortet werden können.

8. Fazit

Insgesamt zeigt sich, dass die Betrachtung der Implementierung des Direct Behavior Ratings „PUTSIE“ viele Erkenntnisse produzierte. Die Erhebung ist jedoch in ihrer bisherigen Form nur bedingt geeignet, um tatsächlich vollständig wissenschaftlichen Gütekriterien zu entsprechen. In diesem Sinne muss Prenzels Aufruf zu größer angelegten Implementationsstudien und der Untersuchung weiterer Modellprogramme (Kapitel 3.3) auch für den Bereich von Direct Behavior Ratings gefolgt werden. Hierzu bedarf es einer engeren Verzahnung von Wissenschaft mit Schule und einem erleichterten Zugang zum schulischen Feld. Nur über eine Vielzahl von Studien an verschiedenen Schulsystemen kann erkannt werden, welche Arten der Implementierung von Direct Behavior Ratings erfolgsversprechend sind und wo mögliche Stolpersteine liegen. Das Projekt LEVUMI bietet sich hierfür ausgezeichnet an.

Es müssen jedoch noch bessere Forschungsmethoden für qualitative Fragestellungen, welche die Implementierung von Methoden betreffen, erarbeitet werden. Die integrative Inhaltsanalyse scheint in der verwendeten Form zwar nicht völlig ungeeignet, die vielfältigen Aussagen der Lehrkräfte können jedoch nur in äußerst differenzierten Kategoriensystemen ausgewertet werden. Es gilt daher, Methoden für die Betrachtung der Verlaufsdiagnostik weiterzuentwickeln, um genauere Aussagen treffen zu können.

Das Ratinginstrument „PUTSIE“ konnte die Lehrkräfte leider noch nicht vollständig überzeugen, wobei Hinweise auf mögliche Probleme in den Ergebnissen vorgestellt und mögliche Lösungswege erläutert werden konnten. Zudem konnte gezeigt werden, dass die Arbeitsweise mit Direct Behavior Ratings kein „Selbstläufer“ ist und von den Lehrkräften einen nicht unerheblichen Aufwand an zeitlichen und personellen Ressourcen einfordert. Es wäre daher zu überprüfen, ob insbesondere die Annahmen zur Ökonomie des Ratinginstruments mit einer Onlineversion womöglich besser verlaufen kann.

Positiv ist abschließend auf die in der Einleitung angesprochene Förderung von Schülerinnen und Schülern mit Verhaltensschwierigkeiten zu verweisen. Direct Behavior Ratings ermöglichen auch aus Sicht der Lehrkräfte eine begleitende Diagnostik des Verhaltens und von Interventionsmaßnahmen. Es scheint daher lohnend das Instrument für die Diagnostik gerade auch im inklusiven Setting weiterzuentwickeln. Die hier dargestellten Ergebnisse können hierzu Anregungen bieten.

Literaturverzeichnis

Blumenthal, Yvonne; Kuhlmann, Kristin; Hartke, Bodo (2014): Diagnostik und Prävention von Lernschwierigkeiten im Aptitude Treatment Interaction-(ATI-) und Response to Intervention-(RTI-)Ansatz. In: Marcus Hasselhorn, Wolfgang Schneider und Ulrich Trautwein (Hg.): Lernverlaufsdiagnostik. 1. Aufl. Göttingen, Niedersachs: Hogrefe Verlag (Tests und Trends, 12), S. 61–81.

Blumenthal, Yvonne; Voß, Stefan (2016): Effekte des Response to Intervention-Ansatzes auf die emotionale und soziale Situation von Grundschülern der vierten Jahrgangsstufe. In: *Empirische Pädagogik* 30 (1), S. 81–97.

Bos, Wilfried; Tarnai, Christian (1999): Content analysis in empirical social research. In: *International Journal of Educational Research* 31, S. 659–671.

Briesch, Amy M. (Hg.) (2016): Direct behavior rating. Linking assessment, communication, and intervention. New York, NY: The Guilford Press.

Casale, Gino; Grosche, Michael; Volpe, Robert, J.; Hennemann, Thomas (2017): Zuverlässigkeit von Verhaltensverlaufsdiagnostik über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen. In: *Empirische Sonderpädagogik* (2), S. 30–42. Online verfügbar unter http://www.wiso-net.de/document/ESP__D426F976973A86483E26384B61C33E39.

Casale, Gino; Hennemann, Thomas; Grosche, Michael (2015a): Zum Beitrag der Verlaufsdiagnostik für eine evidenzbasierte sonderpädagogische Praxis am Beispiel des Förderschwerpunkts der emotionalen und sozialen Entwicklung. In: *Zeitschrift für Heilpädagogik* 66, S. 325–334.

Casale, Gino; Hennemann, Thomas; Huber, Christian; Grosche, Michael (2015b): Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. In: *Heilpädagogische Forschung* 41, S. 37–54.

Falkai, Peter; Wittchen, Hans-Ulrich; Döpfner, Manfred; Gaebel, Wolfgang; Maier, Wolfgang; Rief, Winfried et al. (Hg.) (2018): Diagnostisches und statistisches Manual psychischer Störungen DSM-5®. American Psychiatric Association; Hogrefe-Verlag. 2. korrigierte Auflage, deutsche Ausgabe. Göttingen: Hogrefe.

Faßnacht, Gerhard (2000): Verhalten. Online verfügbar unter https://www.spektrum.de/lexikon/psychologie/verhalten/16243, zuletzt aktualisiert am 9.01.19.

Fröhlich, Werner D. (2015): Wörterbuch Psychologie. 4., unveränderte Nachaufl. München: Dt. Taschenbuch-Verl. (Dtv, 34625).

Früh, Werner (2017): Inhaltsanalyse. Theorie und Praxis. 8. Aufl. Konstanz: UTB; UVK (utb, 2501. Medien- und Kommunikationswissenschaft, Psychologie, Soziologie).

Fuß, Susanne; Karbach, Ute (2014): Grundlagen der Transkription. Eine praktische Einführung. 1. Aufl. Leverkusen: UTB; Budrich, Barbara (utb, 4185 : Sozialwissenschaften).

Gebhardt, Markus; Diehl, Kirsten; Mühling, Andreas (2015): Online-Lernverlaufsmessung für alle Schülerinnen und Schüler in inklusiven Klassen. In: *Zeitschrift für Heilpädagogik* 67, S. 444–453.

Gebhardt, Markus; Diehl, Kirsten; Mühling, Andreas (2016): Lern-Verlaufs-Monitoring. LEVUMI Lehrerhandbuch. Unter Mitarbeit von Christine Engert-Seitz, Ute Haid, Romanna Heinz, Katrin Scheler und Sabine Thoma. Online verfügbar unter https://www.levumi.de/assets/LEVUMI_Lehrerhandbuch-700b60144761e0b1f305dc47846561ce6d47ad108b38aca8dc7f8a87616ceff7.pdf.

Gebhardt, M., DeVries, J.M., Jungjohann, J., Casale, G., Gegenfurtner, A., Kuhn. T. J. (2019). Measurement Invariance of a Direct Behavior Rating Multi Item Scale across Occasions. *Social Sciences, 8*(2), 46. https://doi.org/10.3390/socsci8020046

Hartke, Bodo; Vrban, Robert (2010): Schwierige Schüler - was kann ich tun? 49 Handlungsmöglichkeiten bei Verhaltensauffälligkeiten. 4. Aufl. Buxtehude: Persen (Bergedorfer Grundsteine Schulalltag).

Hennemann, Thomas; Ricking, Heinrich; Huber, Christian (2015): Organisationsformen inklusiver Förderung im Bereich emotional-sozialer Entwicklung. In: Roland Stein, Thomas Müller, Erhard Fischer, Ulrich Heimlich, Joachim Kahlert und Reinhard Lelgemann (Hg.): Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung. 1. Auflage. Stuttgart: Verlag W. Kohlhammer (Inklusion in Schule und Gesellschaft, / herausgegeben von Erhard Fischer, Ulrich Heimlich, Joachim Kahlert und Reinhard Lelgemann ; Band 5), S. 110–143.

Hesse, Ingrid; Latzko, Brigitte (2017): Diagnostik für Lehrkräfte. 3., vollständig überarbeitete und erweiterte Auflage. Opladen, Toronto, Stuttgart: Verlag Barbara Budrich; UTB GmbH (utb-studi-e-book, 3088). Online verfügbar unter http://www.utb-studi-e-book.de/9783838547510.

Hillenbrand, Clemens (2008): Einführung in die Pädagogik bei Verhaltensstörungen. 4., aktualisierte Aufl. München: Reinhardt (utb, 2103).

Hillenbrand, Clemens (2015): Evidenzbasierte Praxis im Förderschwerpunkt emotional-soziale Entwicklung. In: Roland Stein, Thomas Müller, Erhard Fischer, Ulrich Heimlich, Joachim Kahlert und Reinhard Lelgemann (Hg.): Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung. 1. Auflage. Stuttgart: Verlag W. Kohlhammer (Inklusion in Schule und Gesellschaft, / herausgegeben von Erhard Fischer, Ulrich Heimlich, Joachim Kahlert und Reinhard Lelgemann ; Band 5), S. 170–215.

Huber, Christian; Grosche, Michael (2012): Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik 63 (8), 312-322.

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

Jürgens, Eiko; Lissmann, Urban (2015): Pädagogische Diagnostik. Grundlagen und Methoden der Leistungsbeurteilung in der Schule. Weinheim: Beltz (Reihe Bildungs-Wissen Lehramt, 27).

Klauer, Karl Josef (2011): Lernverlaufsdiagnostik - Konzept, Schwierigkeiten und Möglichkeiten. In: *Empirische Sonderpädagogik*, S. 207–224.

Klauer, Karl Josef (2014): Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In: Marcus Hasselhorn, Wolfgang Schneider und Ulrich Trautwein (Hg.): Lernverlaufsdiagnostik. 1. Aufl. Göttingen, N„Die Abenteuer von Levumi und Malini“iedersachs: Hogrefe Verlag (Tests und Trends, 12), S. 1–18.

Kuhl, J., Gebhardt, M., Bienstein, P., Käppler, C., Quinten, S., Ritterfeld, U., Tröster, H. & Wember, F. (2017). Implementationsforschung als Voraussetzung für eine evidenzbasierte sonderpädagogische Praxis. Sonderpädagogische Förderung, 62(4), 383-393

Kultusministerkonferenz (16.12.2004 (i.D.F. vom 12.06.2014)): Standards für die Lehrerbildung: Bildungswissenschaft.

Kultusministerkonferenz (20.10.2011): Inklusive Bildung von Kindern und Jugendlichen mit Behinderungen in Schulen.

Mahlau, Kathrin; Blumenthal, Yvonne; Diehl, Kirsten; Schöning, Anne; Sikora, Simon; Voß, Stefan; Hartke, Bodo (2014): Das Rügener Inklusionsmodell (RIM) - RTI in der Praxis. In: Marcus Hasselhorn, Wolfgang Schneider und Ulrich Trautwein (Hg.): Lernverlaufsdiagnostik. 1. Aufl. Göttingen, Niedersachs: Hogrefe Verlag (Tests und Trends, 12).

Mayer, Horst O. (2013): Interview und schriftliche Befragung. Grundlagen und Methoden empirischer Sozialforschung. 6., überarb. Aufl. München: Oldenbourg (Sozialwissenschaften 10-2012), zuletzt geprüft am 21.01.2019.

Meyer, Hilbert (2018): Leitfaden Unterrichtsvorbereitung. 9. Auflage. Berlin: Cornelsen.

Mühling, Andreas; Gebhardt, Markus; Diehl, Kirsten (2017): Formative Diagnostik durch die Onlineplattform LEVUMI. In: *Informatik Spektrum* 40 (6), S. 556–561. DOI: 10.1007/s00287-017-1069-7.

Myschker, Norbert (2002): Verhaltensstörungen bei Kindern und Jugendlichen. Erscheinungsformen - Ursachen - hilfreiche Massnahmen. 4., überarb. und aktualisierte Aufl. Stuttgart: Kohlhammer (Kohlhammer Pädagogik).

Prenzel, Manfred (2010): Geheimnisvoller Transfer? In: *Z Erziehungswiss* 13 (1), S. 21–37. DOI: 10.1007/s11618-010-0114-y.

Sauerland, Anna (2018): Konstruktion eine Direct Behavior Ratings. Masterarbeit. Technische Universität, Dortmund. Entwicklung und erforschung inklusiver Bildungsprozesse. Online verfügbar unter https://eldorado.tu-dortmund.de/bitstream/2003/37652/1/Sauerland_DBR.pdf, zuletzt geprüft am 23.01.2019.

Schrader, Friedrich-Wilhelm (2010): Diagnostische Kompetenz von Eltern und Lehrern. In: Detlef H. Rost (Hg.): Handwörterbuch pädagogische Psychologie. 4., überarb. und erw. Aufl. Weinheim: Beltz (Programm PVU, Psychologie-Verlags-Union), S. 102–108.

Schuck, Karl Dieter (2014): Förderdiagnostik. In: Franz B. Wember, Roland Stein und Ulrich Heimlich (Hg.): Handlexikon Lernschwierigkeiten und Verhaltensstörungen: Kohlhammer Verlag, S. 122–125.

Stein, Roland (2017): Grundwissen Verhaltensstörungen. 5., neu überarbeitete Auflage. Baltmannsweiler: Schneider Verlag Hohengehren GmbH.

Voß, Stefan; Gebhardt, Markus (2017): Monitoring der sozial-emotionalen Situation von Grundschülerinnen und Grundschülern - Ist der SDQ ein geeignetes Verfahren? In: *Empirische Sonderpädagogik* (1), S. 19–35.

Voß, Stefan; Marten, Katharina; Diehl, Kirsten; Mahlau, Kathrin; Sikora, Simon; Blumenthal, Yvonne; Hartke, Bodo (2015): Evaluationsergebnisse des Projekts Rügener Inklusionsmodell (RIM) - Präventive und Integrative Schule auf Rügen (PISaR) nach vier Schuljahren. In: *Zeitschrift für Heilpädagogik* 66, S. 133–149.

Willmann, Marc (2018): Vermessung des Verhaltens, Normierung zur Inklusion? In: *ZfG* 11 (1), S. 101–114. DOI: 10.1007/s42278-018-0006-4.

Anhang

Anhang A: „Darstellung Stichprobe“:

$N_{Stichprobe} = 7$	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**L9**
Anzahl beobachteter Schüler	2	2	0	2	2	2	0	2	1
Beteiligte Gespräche	2 (Transkripte 3, 4 jeweils Gespräch 1)	2 (Transkripte 3, 4)	5 (Transkripte 1,2,6,7,11)	2 (Transkripte 1,2)	2 (Transkripte 6, 7)	2 (Transkripte 8,9)	2 (Transkripte 8,9 jeweils Gespräch 1)	2 (Transkripte 9, 10)	1 (Transkript 11)
Regelschullehrkraft (x)/ sonderpädagogische Lehrkraft (o)	X	X	O	X	X	X	O	X	X