Empirische Sonderpädagogik, 2020, Nr. 3, S. 207-222
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Können Schülerinnen und Schüler ihr Lernverhalten im Verlauf zuverlässig selbst beurteilen?

Simone Weber, Sina Napiany & Christian Huber

Bergische Universität Wuppertal

Zusammenfassung

Mit der Methode Direct Behavior Rating (DBR) kann Verhalten im Verlauf abgebildet werden (Casale, Hennemann & Grosche, 2015). Bis dato wird DBR hauptsächlich als Fremdbeurteilung durch Lehrkräfte durchgeführt. Eine Selbstbeurteilung durch Schülerinnen und Schüler würde Ressourcen der Lehrkräfte schonen. Zudem erscheint der Zugang aus einer Selbstperspektive je nach zu beurteilendem Verhaltensziel sinnvoller. In der vorliegenden Studie wird daher der Einsatz von DBR als Methode der Selbstbeurteilung untersucht. Im Fokus steht die numerische und strukturelle Übereinstimmung der Selbstbeurteilung durch Schülerinnen und Schüler der vierten Klasse und der Fremdbeurteilung durch Lehrkräfte in Bezug auf die aktive Teilnahme am Unterricht. In einer quantitativen Feldstudie wurden über einen Zeitraum von vier Schulwochen Verlaufsdaten von $N_{Paare} = 18$ auf einer fünfstufigen Ratingskala erhoben. Die Ergebnisse zeigten im Mittel eine höhere Selbstbeurteilung der aktiven Teilnahme am Unterricht durch die Schülerinnen und Schüler als eine Fremdbeurteilung durch die Lehrkräfte. Für die Gesamtstichprobe konnten weder auf numerischer noch auf struktureller Ebene zufriedenstellende Übereinstimmungen zwischen den Beurteilungen nachgewiesen werden. Auf Einzelfallebene wiesen $n_{Paare} = 4$ eine $ICC_{unjustmittel} \geq .5$ und $n_{Paare} = 5$ eine $ICC_{justmittel} \geq .5$ auf. Es werden mögliche Gründe für die niedrigen Beurteilungsübereinstimmungen der Paare diskutiert und Ableitungen für die weitere Forschung getroffen.

Schlüsselwörter: Direct Behavior Rating, Verlaufsdiagnostik, Selbsteinschätzung, Fremdeinschätzung, Beurteilungsübereinstimmung

Are students able to self-assess their learning behavior reliably in terms of progress monitoring?

Abstract

Direct Behavior Rating (DBR) is a behavior assessment method (Casale et al., 2015). So far, DBR is mostly used as a tool of external monitoring. The use of DBR as a method of self-monitoring could be a possibility to look after teachers' resources. Furthermore, the use of DBR as a method of self-monitoring might be more appropriate for some behaviors. Therefore, the present study focuses on DBR as a method of self-monitoring. Particularly, the numeric and structural invariance of self-monitoring and external monitoring data is examined. Fourth-grade students (self-monitoring) as well as teachers (external monitoring) rated students' active engagement.

Within a quantitative field study, N_{Pairs} = 18 pairs of students and teachers collected progress monitoring data over a four-week period on a five-point rating scale. On average, self-monitoring data by students was higher than external monitoring data by teachers. Focusing on N_{Pairs} = 18, there were no significant correlations between the DBR ratings, neither numeric nor structural. Focusing on single cases, n_{Pairs} = 4 presented $ICC_{unjust\_average} \geq .5$ and n_{Pairs} = 5 presented $ICC_{just\_average} \geq .5$. Potential reasons for inadequate interrater agreements as well as conclusions for further research are discussed.

Keywords: Direct Behavior Rating, behavior progress monitoring assessment, self-assessment, external assessment, interrater-reliability

Während der Bereich der Lernverlaufsdiagnostik bereits seit längerem einem erhöhten Forschungsinteresse unterliegt (Deno, Fuchs, Marston & Shin, 2001; Förster, Kuhn & Souvignier, 2017), gewinnt der Bereich der Verhaltensverlaufsdiagnostik ebenfalls zunehmend an Beachtung. Eine hochfrequente Beurteilung des Verhaltens bietet die Möglichkeit, individuelle Verhaltensentwicklungen sichtbar zu machen, Veränderungen sensitiv abzubilden und somit Fördermaßnahmen im Einzelfall zu evaluieren (Casale et al., 2015).

Die systematische direkte Beobachtung gilt testtheoretisch als bislang am besten abgesichertes Verfahren, um Verhalten objektiv, reliabel und valide zu erfassen (Riley-Tillman, Chafouleas, Sassu, Chanese & Glazer, 2008). Der Einsatz einer systematischen direkten Beobachtung ist jedoch nicht ökonomisch. Weder eine Lehrkraft noch eine Schülerin oder ein Schüler hat im Zuge des schulischen Alltags die Möglichkeit, ein bestimmtes Verhalten mittels eines Kategoriensystems engmaschig zu dokumentieren (Casale, Hennemann, Huber & Grosche, 2015). Deshalb wird im Kontext Schule häufig auf Fragebögen zurückgegriffen. Fragebogenverfahren mit Ratingskalen sind in ihrem Einsatz deutlich ökonomischer, da sie eine umfassende und dennoch zeitsparende retrospektive Einschätzung vergangener Verhaltensweisen ermöglichen. Für eine statusdiagnostische Verwendung weisen sie in der Regel eine zufriedenstellende Reliabilität auf. Da in Fragebogenverfahren jedoch häufig das Verhalten von mehreren, vergangenen Wochen eingeschätzt wird, ist ihre Eignung für eine hochfrequente Verhaltensbeurteilung im Zuge einer Verlaufsdiagnostik fraglich (Casale et al., 2015). Eine ökonomischere Möglichkeit zur Beurteilung von Verhaltensverläufen im schulischen Kontext ist die Methode Direct Behavior Rating (DBR).

DBR

Bei DBR handelt es sich um eine Kombination aus einer systematischen, direkten Verhaltensbeobachtung und einer Verhaltensbeurteilung. Die Beurteilungen eines zuvor definierten Verhaltens mittels DBR erfolgen hochfrequent, sodass ausreichend Datenpunkte zur Abbildung eines Verhaltensverlaufs zur Verfügung stehen. In der Schulpraxis wird DBR häufig von einer Lehrkraft durchgeführt, die das Verhalten einer Schülerin oder eines Schülers aus einer Fremdperspektive täglich zu mehreren Zeitpunkten beurteilt. DBR ist flexibel in Bezug auf das zu beobachtende Verhalten, die Dauer, den Zeitraum sowie die Häufigkeit der Messungen und kann somit an zahlreiche Kontexte angepasst werden (Chafouleas, Kilgus & Wallach, 2010; Christ, Riley-Tillman & Chafouleas, 2009). Ein vertiefender Einblick in die Methode DBR findet sich z.B. bei Chafouleas (2011) oder Huber (2016).

In den USA unterliegt die Methode DBR seit einiger Zeit einer systematischen Untersuchung der Testgüte (Chafouleas, Riley-Tillman & Christ, 2009; Volpe & Briesch, 2012). Verschiedene Arbeiten aus dem deutschen Sprachraum ergänzen diese Erkenntnisse (Casale, 2017; Casale, Grosche, Volpe & Hennemann, 2017; Huber & Rietz, 2015a). Der bisherige Forschungsstand bezieht sich primär auf den Einsatz von DBR als Methode der Fremdbeurteilung. Hierbei sind insbesondere die drei externalisierenden Verhaltensweisen aktive Teilnahme am Unterricht, respektvolles Verhalten und störungsfreies Verhalten (engl.: *academically engaged, respectful, non-disruptive)* im Fokus der Forschung. Diese Verhaltensweisen werden als *BIG 3* bezeichnet und gelten als besonders relevant für einen erfolgreichen Schulbesuch (Chafouleas, 2011).

In Bezug auf die Interrater-Reliabilität beim Einsatz von DBR liegen einige kontroverse Forschungsbefunde vor. Während in der Studie von Briesch, Chafouleas und Riley-Tillman (2010) eine unzureichende Interrater-Reliabilität nachgewiesen wurde, konnten andere Studien hohe Beurteilungsübereinstimmungen aufzeigen (Chafouleas, Christ, Riley-Tillman, Briesch & Chanese, 2007; Volpe & Briesch, 2012). Aus den bisherigen Befunden wurden von den Autorinnen und Autoren erste Implikationen zur Verbesserung der Interrater-Reliabilität abgeleitet. Es wird empfohlen, DBR durch eine gleichbleibende Person durchzuführen, um bestehende Beurteilungstendenzen im Zeitverlauf konstant zu halten (Briesch et al., 2010; Casale et al., 2015). Weiter ist es nach Briesch, Chafouleas und Riley-Tillman (2016) ratsam, die Länge des Zeitraums, auf welchen sich die Beurteilungen beziehen, konstant zu halten. Um das Auftreten von Beobachtungsverzerrungen möglichst gering zu halten, sollte die Verhaltensbeurteilung unmittelbar nach der Situation des Auftretens stattfinden. Briesch et al. (2010) empfehlen weiter den Einsatz eines vorherigen Trainings zur Erhöhung der Interrater-Reliabilität. Bis dato liegen noch keine eindeutigen Befunde in Bezug auf die für eine ausreichende Interrater-Reliabilität benötigte Anzahl von Beurteilungen mittels DBR vor. Die empfohlenen Angaben schwanken zwischen fünf und 20 Messzeitpunkten, wobei die zugrundeliegenden Studien Unterschiede in Bezug auf die Verhaltensweise und den Beobachtungszeitraum aufweisen (Huber & Rietz, 2015a). Ferner handelt es sich bei den bisherigen Forschungsarbeiten primär um Studien, die in einem Laborsetting durchgeführt wurden. Es ist daher anzunehmen, dass für die schulische Praxis weitaus mehr direkte Verhaltensbeurteilungen nötig sind, um eine zufriedenstellende Interrater-Reliabilität zu erzielen.

Einige Autorinnen und Autoren plädieren für eine globale Formulierung der Verhaltensweisen (Christ, Riley-Tillman, Chafouleas & Jaffery, 2011). Andere empirische Befunde weisen hingegen auf eine erhöhte Interrater-Reliabilität bei einer konkreten Formulierung des Verhaltens hin (Chafouleas et al., 2007; Volpe & Briesch, 2012). Des Weiteren wurde der Einfluss einer positiven oder negativen Formulierung der Verhaltensweise auf die Interrater-Reliabilität untersucht. Erste Befunde deuten auf eine höhere Interrater-Reliabilität bei einer positiven anstelle einer negativen Formulierung der Verhaltensweise hin (Christ et al., 2011). Eine positive Formulierung ist zudem aus pädagogischen Gründen einer negativen Formulierung vorzuziehen (Riley-Tillman, Christ, Chafouleas, Boice Mallach & Briesch, 2011).

Die Methode DBR unterliegt weiteren Limitationen. Die beurteilende Person nimmt das Verhalten über die Schulstunde hinweg eher implizit wahr und beurteilt dieses zeitlich verzögert im Anschluss. Diese eher implizite Registrierung erhöht die Wahrscheinlichkeit des Auftretens von Beobachtungsfehlern in Bezug auf die Wahrnehmung, Erinnerung und Interpretation des Verhaltens. Sofern DBR von mehreren

Personen durchgeführt wird, könnten sich diese subjektiven Beobachtungsverzerrungen addieren (Chafouleas et al., 2007). Mit Blick auf die bislang vorliegende DBR-Forschung wurde von zahlreichen Arbeitsgruppen betont, dass das Zielverhalten eindeutig beobachtbar sein muss (z.B. Casale et al., 2017). Eine eindeutige Beobachtbarkeit erscheint insbesondere bei externalisierenden Verhaltensweisen möglich. Internalisierende Verhaltensweisen, wie Unsicherheit und Angst, sind hingegen aufgrund ihrer Ausprägung, die sich hauptsächlich auf das Selbst und das Innere eines Menschen richtet, von außen nur bedingt beobachtbar. Dies führt dazu, dass externe Personen das Auftreten internalisierender Verhaltenssymptome häufig anders beurteilen als die betroffenen Personen selbst (Klasen et al., 2016).

DBR als Methode der Selbstbeurteilung

In Bezug auf den Einsatz von DBR deutet dies darauf hin, dass eine Fremdbeurteilung von internalisierenden Verhaltensweisen durch eine Lehrkraft wahrscheinlich nur bedingt aussagekräftig wäre. Hinzu kommt, dass eine Fremdbeurteilung, insbesondere wenn das Verhalten von mehreren Schülerinnen und Schülern beurteilt werden soll stets Ressourcen der Lehrkraft erfordert. Eine Lösung dieser Problematik könnte in einer Beurteilung durch die Schülerinnen und Schüler selbst liegen.

Es ist jedoch zu bedenken, dass Selbsteinschätzungen häufig selbstwertdienlichen und sozial erwünschten Verzerrungen unterliegen (Bortz & Döring, 2016). Während Kinder insbesondere im Elementar- sowie im frühen Primarbereich die eigenen Kompetenzen häufig noch überschätzen, entwickeln sie im Laufe ihrer Grundschulzeit eine zunehmend realistischere Selbsteinschätzung (Helmke, 1998; Möller & Trautwein, 2009; Nicolls, 1978). Durch den Schuleintritt trägt der alltägliche Kontakt zu Gleichaltrigen und erwachsenen Bezugspersonen dazu bei, die Entwicklung realistischer Selbsteinschätzungen von Kindern zu unterstützen. Durch Interaktionen mit Peers und Lehrkräften erhalten Kinder Rückmeldungen über sich selbst und ihr Verhalten. Dies sensibilisiert das Bewusstsein des Kindes über die eigenen Kompetenzen und unterstützt die Entwicklung des Selbstkonzepts (Hellmich & Günther, 2011). Trotz der Entwicklung einer zunehmend realistischeren Selbsteinschätzung im Verlauf der Grundschulzeit ist zu berücksichtigen, dass sich auch im Erwachsenenalter noch höhere Einschätzungen der eigenen Fähigkeiten und Kompetenzen aus einer Selbstperspektive im Vergleich zu Einschätzungen aus einer Fremdperspektive finden lassen (Gold & Kuhn, 2017; Harris & Schaubroeck, 1988; Von Stumm, 2014).

Während für den Einsatz von Verhaltensverlaufsdiagnostiken aus Fremdperspektive (DBR_{fremd}) bereits vergleichsweise viele empirische Befunde vorliegen, ist der Bereich der Verhaltensverlaufsdiagnostik aus der Selbstperspektive (DBR_{selbst}) noch nahezu unerforscht. Von der Embse, Scott und Kilgus (2015) untersuchten den Einsatz von DBR_{selbst} zur Messung von Leistungsangst im universitären Kontext. Die Ergebnisse sprechen grundlegend für den Einsatz von DBR_{selbst} zur Messung von Verhalten im Verlauf. Die Studie ist jedoch nur bedingt auf den vorliegenden Forschungskontext übertragbar. Zum einen wurde die Studie mit Studierenden und nicht mit Kindern durchgeführt. Zum anderen führten die Studierenden insgesamt nur sieben Verlaufsbeurteilungen mittels DBR_{selbst} durch. Eine hochfrequente Beurteilung mittels DBR sieht jedoch deutlich mehr Beurteilungen im Verlauf vor, wodurch der alleinige Einsatz eines statusdiagnostischen Inventars als externes Kontrollkriterium, wie in der Studie von Von der Embse et al. (2015), nur bedingt aussagekräftig erscheint. Weiter würde der Einsatz einer statusdiagnostischen Selbsteinschätzung die Frage nach der

Selbsteinschätzungskompetenz der Schülerinnen und Schüler im Primarbereich nur unzureichend beantworten. Es erscheint daher naheliegender, DBR_{selbst}- und DBR_{fremd}-Daten hinsichtlich ihrer Übereinstimmung zu betrachten.

Mit Blick auf die Interrater-Reliabilität könnten DBR_{selbst}- und DBR_{fremd}-Messungen sowohl strukturell als auch numerisch übereinstimmen (Huber & Rietz, 2015b; Wirtz & Caspar, 2002). Eine hohe numerische Invarianz der Messungen bedeutet, dass die Einzelwerte mehrerer Messreihen absolut übereinstimmen. Von diesem Fall ist die strukturelle Invarianz der Messreihen zu unterscheiden. Bei einer hohen strukturellen Invarianz sind Messreihen in ihren Einzelwerten nicht identisch, aber in ihrem Verlauf ähnlich. Huber und Rietz (2015b) weisen darauf hin, dass hohe numerische Übereinstimmungen von unterschiedlichen Personen bei hochfrequenten Verhaltensbeurteilungen weder zu erwarten noch zwingend notwendig sind. Da das Ziel einer Verlaufsdiagnostik in der Evaluation und Optimierung von Fördermaßnahmen liegt, ist eine hohe strukturelle Übereinstimmung unterschiedlicher beurteilender Personen, bei Berücksichtigung einer individuellen Bezugsnorm, ausreichend. Eine hohe numerische Übereinstimmung ist hingegen für statusdiagnostische Entscheidungen und unter Berücksichtigung einer kriterialen Bezugsnorm von höherer Bedeutung.

Fragestellung und Hypothesen

Das Ziel der vorliegenden Arbeit ist es, die Beurteilungsübereinstimmung zwischen Schülerinnen und Schülern (DBR_{selbst}) und deren Lehrkräften (DBR_{fremd}) zu überprüfen. Da für die Verhaltensweise der aktiven Teilnahme am Unterricht bereits empirische Befunde vorliegen (z.B. Kilgus, Chafouleas, Riley-Tillman & Welsh, 2012), sollen die Beurteilungsübereinstimmungen anhand dieser Verhaltensweise überprüft werden. Zunächst wird hierzu das Antwortverhalten der Schülerinnen und Schüler und der Lehrkräfte betrachtet. Da bisher noch keine empirischen Befunde zur Reliabilität von DBR_{selbst}-Beurteilungen vorliegen, werden in der folgenden Studie folgende Forschungsfragen untersucht:

Fragestellung 1: Wie hoch fällt die absolute Übereinstimmung zwischen den Fremdbeurteilungen und den Selbstbeurteilungen aus?

Fragestellung 2: Wie hoch fällt die strukturelle Übereinstimmung zwischen den Fremdbeurteilungen und den Selbstbeurteilungen aus?

Methode

Stichprobe

An der Studie nahmen insgesamt elf Grundschulklassen aus zehn unterschiedlichen Regelschulen teil. Es handelte sich um vierte Klassen aus Nordrhein-Westfalen. Die Stichprobe bestand aus $N_{SuS} = 22$ Schülerinnen und Schülern. Für diese Schülerinnen und Schüler lagen neben Selbstbeurteilungen verdeckte Fremdbeurteilungen der jeweiligen Klassenlehrkraft vor. Die Schülerinnen und Schüler waren zwischen acht und zehn Jahren alt ($M = 9.5$; $SD = 0.59$). Der Anteil der Mädchen lag bei 50 Prozent. Zwölf Lehrkräfte führten DBR_{fremd} für zwei ausgewählte Schulkinder aus jeder Klasse durch. Elf der Lehrkräfte waren weiblich. In einer dieser elf Klassen bestand eine geteilte Klassenleitung und beide Lehrkräfte nahmen an dem Projekt teil. Somit entstanden insgesamt $N_{Paare} = 24$ Beurteilungspaare (ein Beurteilungspaar bestand aus einem Schulkind und einer Lehrkraft). Die Zahl der Berufsjahre der Lehrkräfte belief sich im Mittel auf $M = 13.5$ ($SD = 9.17$).

Studiendesign

Die Feldstudie war in einem Längsschnittdesign konzipiert. Über einen Zeitraum von vier aufeinanderfolgenden Schulwochen wurden viermal täglich Verlaufsdaten aus Sicht von Schülerinnen und Schülern sowie von Lehrkräften mittels DBR in Bezug auf den gleichen Zeitraum erhoben. Insgesamt waren in diesem vierwöchigen Zeitrahmen theoretisch bis zu 80 Messwiederholungen möglich.

Instrumente

DBR

Für die Einschätzung der aktiven Teilnahme am Unterricht im Verlauf wurde DBR genutzt. Die DBR-Daten wurden appbasiert auf Tablet-PCs erhoben, die teilnehmenden Schulklassen für den Erhebungszeitraum zur Verfügung gestellt wurden. Die notwendige App wurden eigens für diese Studie entwickelt. Die Schülerinnen und Schüler gingen einzeln an die Tablet-PCs und wählten in einem ersten Schritt ihren Namen aus. Daraufhin wurden sie an das Ziel „Ich arbeite im Unterricht mit" erinnert. Die Schülerinnen und Schüler wurden gefragt, wie sie in der letzten Stunde mitgearbeitet haben. Ihre Beurteilung gaben sie auf einer fünfstufigen Skala ab. In einer Vorstudie wurden die Handhabbarkeit sowie das Verständnis der DBR-Skala mit zwei vierten Grundschulkassen überprüft. Da ein Großteil der Schülerinnen und Schüler eine fünfstufige Skala als verständlicher und leichter handhabbar einschätzte als eine sechsstufige, wurde in der Hauptstudie die fünfstufige Skala genutzt. Die einzelnen Abstufungen wurden symbolisch in Form von Smileys dargestellt. Ihre abgegebene Beurteilung bestätigten die Schülerinnen und Schüler und setzten sich zurück an ihren Platz, woraufhin das nächste Schulkind die Selbstbeurteilung durchführte. Die Lehrkraft führte die Fremdbeurteilung für zwei ausgewählte Schülerinnen und Schüler ebenfalls am Tablet-PC durch. Der Aufbau war äquivalent zu dem der Schülerinnen und Schüler. Die Verhaltensweise der aktiven Teilnahme am Unterricht wurde durch den Einsatz einer *Single-Item-Scale* (Casale et al., 2017) beurteilt und es wurde folgende Operationalisierung vorgenommen: „Ich arbeite im Unterricht mit. Das bedeutet: Ich passe gut auf; Ich weiß, was wir gerade machen; Ich erledige meine Aufgaben; Ich beteilige mich."

Für DBR_{fremd} liegen moderate bis gute Befunde für die Kriteriums-Validität und die Interrater-Reliabilität vor (Huber & Rietz, 2015a).

Ablauf

Die Hauptstudie begann mit einer Übungswoche, in welcher die teilnehmenden Personen die eingesetzten Verfahren kennenlernten und erprobten. Eine vorgefertigte Operationalisierung des Verhaltensziels wurde besprochen und verblieb während der gesamten Zeit, in Form eines Plakates, im Klassenzimmer. Durch die Operationalisierung sollte sichergestellt werden, dass die Lehrkraft und die Schülerinnen und Schüler ihre Verhaltensbeurteilungen basierend auf denselben Merkmalsausprägungen durchführten. In den darauffolgenden vier Wochen beurteilten die Schülerinnen und Schüler der teilnehmenden Klassen viermal täglich ihre aktive Teilnahme am Unterricht mittels DBR_{selbst}. Eine Festlegung auf bestimmte Schulfächer fand nicht statt. Des Weiteren wurde den Lehrkräften nicht vorgegeben, zu welchen Zeiten DBR stattfinden sollte. Die Fremdbeurteilung durch die Lehrkraft wurde zeitlich parallel zur Selbstbeurteilung der Schülerinnen und Schüler vorgenommen und bezog sich somit immer auf denselben Beurteilungszeitraum.

Datenaufbereitung & Datenauswertung

Zur Prüfung der Beurteilungsübereinstimmung wurden nur diejenigen Messzeitpunkte genutzt, für welche eine Selbstbeurteilung durch die Schülerin oder den Schüler sowie eine Fremdbeurteilung durch die Lehrkraft vorlag. Zudem wurden DBR_{selbst}-und DBR_{fremd}-Daten, bei welchen die Selbst- und die Fremdbeurteilung 20 Minuten oder länger auseinanderlagen, aussortiert. Es bestand die Gefahr, dass sich diese Messzeitpunkte in diesen Fällen nicht mehr auf den gleichen Kontext bezogen. Sechs Datensätze mussten aufgrund längerer Fehlzeiten der teilnehmenden Personen oder technischen Schwierigkeiten mit den Tablet-PCs vollständig aus der Analyse ausgeschlossen werden. Somit lagen zur Prüfung der Übereinstimmung von Selbstbeurteilung und Fremdbeurteilung nach Ausschluss Daten von N_{Paare} = 18 Beurteilungspaaren vor.

Für die Betrachtung der numerischen Übereinstimmung der DBR_{selbst}- und DBR_{fremd}-Beurteilungen wurden die Häufigkeiten der identischen Beurteilungen berechnet.

Für die Betrachtung der strukturellen Übereinstimmung wurden die einzelnen Messwertreihen der DBR_{selbst}- und DBR_{fremd}-Beurteilungen daraufhin überprüft, ob eine Verbesserung, eine Verschlechterung oder keine Veränderung zum vorherigen Messzeitpunkt wahrgenommen wurde. In diesem ersten Schritt soll zunächst überprüft werden, ob Lehrkräfte und Schülerinnen und Schüler Veränderungen und deren Tendenzen überhaupt ähnlich wahrnehmen, weshalb das Ausmaß der wahrgenommenen Veränderung nicht beachtet wurde.

Darüber hinaus wurden Übereinstimmungen durch Intraklassenkorrelationen (ICC) berechnet. Die ICC ermöglicht die Überprüfung der Beurteilungsübereinstimmung mehrerer Personen und kann laut Gross Portney und Watkins (2014) bereits für ordinalskalierte Daten genutzt werden. Die ICC als Maß der Varianzaufklärung nimmt einen Wert zwischen 0 und 1 an. Negative ICC-Werte sind als 0 zu werten (Wirtz & Caspar, 2002). Die Überprüfung der Beurteilungsübereinstimmung mittels ICC ermöglicht sowohl die Überprüfung der numerischen (unjustiert, ICC_{unjust}) als auch der strukturellen Beurteilungsübereinstimmung (justiert, ICC_{just}).

Aus der Berechnung der ICC resultieren sowohl einzelne Maße (Einzelschätzer) als auch durchschnittliche Maße (Mittelschätzer). Während sich der Einzelschätzer auf das einzelne Rating bezieht, fokussiert der Mittelschätzer den Mittelwert mehrerer Ratings (Shrout & Fleiss, 1979). In Tabelle 4 und 5 sind sowohl der Einzel- als auch der Mittelschätzer für alle Beurteilungspaare aufgeführt. Im Fließtext wird ausschließlich der Mittelschätzer berichtet. Für die Berechnung der ICC in der vorliegenden Studie wurde ein zweifaktorielles, zufälliges Modell genutzt.

Eine ICC ≥ .7 wird in der Literatur als Maß für eine hohe Übereinstimmung beschrieben (Bortz & Döring, 2016). Wirtz und Caspar (2002) sowie Shoukri, Asyali und Donner (2004) vertreten die Ansicht, dass es sich bei der Angabe ICC ≥ .7 lediglich um einen ungefähren Richtwert handelt und die ICC immer unter Berücksichtigung der jeweiligen Bedingungen der Datenerhebung interpretiert werden muss. Es finden sich jedoch keine Empfehlungen über das Ausmaß der Berücksichtigung. Begründet durch die unterschiedlichen Beurteilungsperspektiven von Lehrkräften (DBR_{fremd}) und Kindern (DBR_{selbst}) sowie die Beurteilung während des schulischen Alltags wurde für diese Studie eine ICC ≥ .5 als zufriedenstellendes Maß der Beurteilungsüberstimmung definiert. Nach unserem aktuellen Wissensstand liegt für eine Feldstudie im schulischen Kontext kein definiertes ICC-Maß vor.

Ergebnisse

Die Selbstbeurteilung der Schülerinnen und Schüler auf der fünfstufigen Skala lag im Mittel bei $M_{DBRselbst} = 4.37$ ($SD_{DBRselbst} = 0.19$) und die Fremdbeurteilung der Lehrkräfte bei $M_{DBRfremd} = 3.81$ ($SD_{DBRfremd} = 0.38$). Die Werte für die einzelnen Paare sind in Tabelle 1getrennt für Schülerinnen und Schüler (DBR_{selbst}) sowie Lehrkräfte (DBR_{fremd}) aufgeführt.

Tabelle 2 gibt Auskunft über die Nutzung der DBR-Skala. Die Schülerinnen und Schüler wählten in mehr als der Hälfte ihrer Beurteilungen die Antwortstufe 5, wohingegen die Antwortstufen 1 und 2 nur selten genutzt wurden. Die Lehrkräfte nutzten die Antwortstufe 5 in etwa halb so häufig wie die Schülerinnen und Schüler, wobei sie die Antwortstufe 4 häufiger nutzten. Auch bei den Lehrkräften zeigte sich eine seltene Nutzung der unteren Antwortstufen.

Die theoretisch mögliche Anzahl von 80 Messwiederholungen wurde von einem Beurteilungspaar erreicht. Für die $N_{Paare} = 18$ lagen im Mittel $M_{MZP} = 46.33$ ($SD_{MZP} = 17.77$) gültige DBR-Beurteilungen vor. Tabelle 4 gibt Auskunft über die Anzahl der gültigen DBR-Beurteilungen für jedes Paar.

Tabelle 1: Deskriptive Darstellung der einzelnen Paare getrennt nach Schülerinnen und Schüler (DBRselbst) und Lehrkräften (DBR_{fremd})

	DBR_{selbst} (SD)	DBR_{fremd} (SD)	DBR_{selbst} - DBR_{fremd}	d_{Cohen}
Paar 01	4.24 (1.11)	3.47 (1.07)	0.76	0.71
Paar 02	4.07 (0.73)	3.79 (0.63)	0.28	0.41
Paar 03	4.37 (1.06)	4.05 (0.88)	0.32	0.33
Paar 04	4.16 (1.14)	3.95 (0.89)	0.21	0.21
Paar 05	4.18 (1.32)	3.56 (1.10)	0.62	0.51
Paar 06	4.25 (0.88)	3.87 (0.92)	0.38	0.42
Paar 07	4.17 (1.34)	3.57 (0.82)	0.61	0.54
Paar 08	4.28 (0.73)	3.40 (0.89)	0.88	1.08
Paar 09	4.41 (0.58)	3.70 (0.89)	0.70	0.95
Paar 10	4.73 (0.60)	3.69 (0.88)	1.04	1.38
Paar 11	4.63 (0.82)	3.91(1.04)	0.72	0.77
Paar 12	4.26 (0.70)	4.41 (0.62)	-0.15	0.23
Paar 13	4.71 (0.68)	3.43 (0.79)	1.29	1.77
Paar 14	4.56 (0.74)	3.69 (0.72)	0.87	1.19
Paar 15	4.56 (0.50)	4.80 (0.46)	-0.24	0.5
Paar 16	4.51 (0.62)	3.51 (0.75)	1.00	1.45
Paar 17	4.35 (1.00)	3.43 (0.77)	0.91	1.03
Paar 18	4.26 (0.62)	4.44 (0.61)	-0.18	0.29

Tabelle 2: Deskriptive Darstellung des Antwortverhaltens im DBR getrennt nach Schülerinnen und Schülern (DBR_{selbst}) sowie Lehrkräften (DBR_{fremd}) in absoluten Häufigkeiten und Prozent

	Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5	Gesamt
DBR_{selbst}	17 (2%)	12 (1.4%)	58 (7%)	268 (32.1%)	479 (57.4%)	834 (100%)
DBR_{fremd}	11 (1.3%)	54 (6.5%)	191 (22.9%)	362 (43.3%)	216 (25.9%)	834 (100%)

Anmerkungen. Für die $N_{Paare} = 18$ liegen unterschiedlich viele Messzeitpunkte vor (siehe Tab. 4). Die vorliegende Darstellung des Antwortverhaltens ist insofern verzerrt, als dass alle vorhandenen Messzeitpunkte aller Paare einbezogen wurden.

Mit Fragestellung 1 wurde untersucht, wie die absolute Übereinstimmung zwischen den Selbstbeurteilungen und den Fremdbeurteilungen ausfiel. Auf deskriptiver Ebene zeigte sich, dass die $N_{Paare} = 18$ in knapp einem Drittel der Fälle ($N_{MZP} = 275$) in ihren absoluten Werten übereinstimmten. Die höchste prozentuale Übereinstimmung zwischen den Fremd- und Selbstbeurteilungen lag in der obersten Stufe vor und nahm mit niedrig werdender Stufe ab. Tabelle 3 gibt Auskunft über die Verteilung der numerischen Übereinstimmungen auf den Stufen 1 bis 5 in Bezug auf die Gesamtstichprobe.

In Tabelle 4 sind die numerischen Übereinstimmungen der einzelnen Paare in absoluten Häufigkeiten und Prozent angegeben. Zur Überprüfung der numerischen Übereinstimmungen wurden ICC_{unjust} berechnet.

Auf deskriptiver Ebene zeigte sich, dass zwei Paare in mehr als der Hälfte ihrer Beurteilungen absolut übereinstimmten. Sieben weitere Paare stimmten in mehr als einem Drittel ihrer Beurteilungen absolut überein.

Vier Paare wiesen im Mittelschätzer $ICC_{unjustmittel} \geq .5$ auf. Die verbleibenden

Tabelle 3: Deskriptive Darstellung der Verteilung der Beurteilungsübereinstimmungen im DBR der vorliegenden $N_{Paare} = 18$ in absoluten Häufigkeiten und Prozent

Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5	Übereinstimmungen gesamt	Beurteilungen gesamt
2 (0.2%)	6 (0.7%)	32 (3.8%)	111 (13.3%)	124 (14.9%)	275 (33%)	834

Tabelle 4: Darstellung der Verteilung der numerischen Beurteilungsübereinstimmungen im DBR in absoluten Häufigkeiten und Prozent sowie ICC_{unjust} im Einzel- und Mittelschätzer aller einzelnen Beobachtungspaare absteigend nach $ICC_{unjustmittel}$ sortiert

	N_{MZP}[a]	numerische Übereinstimmungen	$ICC_{unjustmittel}$	$ICC_{unjusteinzel}$
Paar 01	38	12 (32%)	.682	.518
Paar 02	43	24 (56%)	.664	.497
Paar 03	63	25 (40%)	.538	.368
Paar 04	19	8 (42%)	.538	.368
Paar 05	39	16 (41%)	.478	.314
Paar 06	55	15 (27%)	.449	.289
Paar 07	23	5 (22%)	.432	.276
Paar 08	43	12 (28%)	.345	.208
Paar 09	44	15 (34%)	.305	.180
Paar 10	74	16 (22%)	.290	.170
Paar 11	32	11 (34%)	.285	.166
Paar 12	27	11 (41%)	.178	.098
Paar 13	70	6 (9%)	.149	.080
Paar 14	54	12 (22%)	.134	.072
Paar 15	80	44 (55%)	.094	.049
Paar 16	45	12 (27%)	.057	.029
Paar 17	23	5 (22%)	-.036	-.018
Paar 18	62	26 (42%)	-.084	-.040

Anmerkungen: [a] Es wurden nur diejenigen Messzeitpunkte ausgewertet, für die sowohl eine Beurteilung der Lehrkraft als auch Beurteilung des Schulkindes vorlag.

vierzehn Paare wiesen im Mittelschätzer $ICC_{unjustmittel} \leq .5$ auf.

Anhand Fragestellung 2 wurde die strukturelle Übereinstimmung zwischen den Fremdbeurteilungen und Selbstbeurteilungen untersucht. In Tabelle 5 sind die strukturellen Übereinstimmungen der einzelnen Paare in absoluten Häufigkeiten und Prozent angegeben. Zur Überprüfung der strukturellen Übereinstimmungen der Beurteilungen wurden ICC_{just} berechnet. Diese sind ebenfalls in Tabelle 5 aufgeführt.

Auf deskriptiver Ebene zeigte sich, dass zwei Paare in mehr als der Hälfte ihrer Beurteilungen strukturell übereinstimmten. Zwölf weitere Paare stimmten in mehr als einem Drittel ihrer Beurteilungen strukturell überein. Zur exemplarischen Ansicht ist in Abbildung 1 die Beurteilungsübereinstimmung von Paar 07 grafisch dargestellt. Auch wenn die Kurvenverläufe numerisch nicht übereinstimmen, so ist ein ähnlicher Verlauf hinsichtlich der Wahrnehmung einer Verbesserung, Verschlechterung oder Stagnation zu erkennen.

Es zeigte sich, dass fünf Paare im Mittelschätzer $ICC_{justmittel} \geq .5$ und 13 Paare im Mittelschätzer $ICC_{justmittel} \leq .5$ aufwiesen.

Diskussion

Die vorliegende Studie greift die bislang offene Frage auf, ob eine Beurteilung des eigenen Verhaltens im Verlauf durch Schülerinnen und Schüler selbst durchgeführt werden kann. Dies würde zum einen Ressourcen der Lehrkraft schonen und es wäre möglich, auch solche Verhaltensweisen im Verlauf zu erfassen, die aus einer Fremdperspektive nur schwer zugänglich sind. In der vorliegenden Studie wurde daher untersucht, inwieweit die für die Fremdbeurtei-

Tabelle 5: Darstellung der Verteilung der strukturellen Beurteilungsübereinstimmungen im DBR in absoluten Häufigkeiten und Prozent sowie ICC_{just} im Einzel- und Mittelschätzer aller einzelnen Beobachtungspaare absteigend nach $ICC_{justmittel}$ sortiert

	N_{MZP}[a]	strukturelle Übereinstimmungen	$ICC_{justmittel}$	$ICC_{justeinzel}$
Paar 01	38	18 (49%)	.778	.636
Paar 02	43	21 (50%)	.695	.533
Paar 03	63	22 (35%)	.554	.383
Paar 04	19	6 (33%)	.532	.363
Paar 05	39	15 (39%)	.516	.348
Paar 06	43	19 (45%)	.491	.325
Paar 07	74	32 (44%)	.484	.320
Paar 08	55	26 (48%)	.475	.311
Paar 09	23	14 (64%)	.470	.307
Paar 10	44	17 (40%)	.405	.254
Paar 11	32	10 (32%)	.346	.209
Paar 12	70	29 (42%)	.334	.201
Paar 13	54	23 (43%)	.216	.121
Paar 14	27	10 (38%)	.177	.097
Paar 15	45	26 (59%)	.111	.059
Paar 16	80	33 (42%)	.104	.055
Paar 17	23	6 (27%)	-.053	-.026
Paar 18	62	18 (30%)	-.086	-.041

Anmerkungen: [a] Es wurden nur diejenigen Messzeitpunkte ausgewertet, für die sowohl eine Beurteilung der Lehrkraft als auch Beurteilung des Schulkindes vorlag.

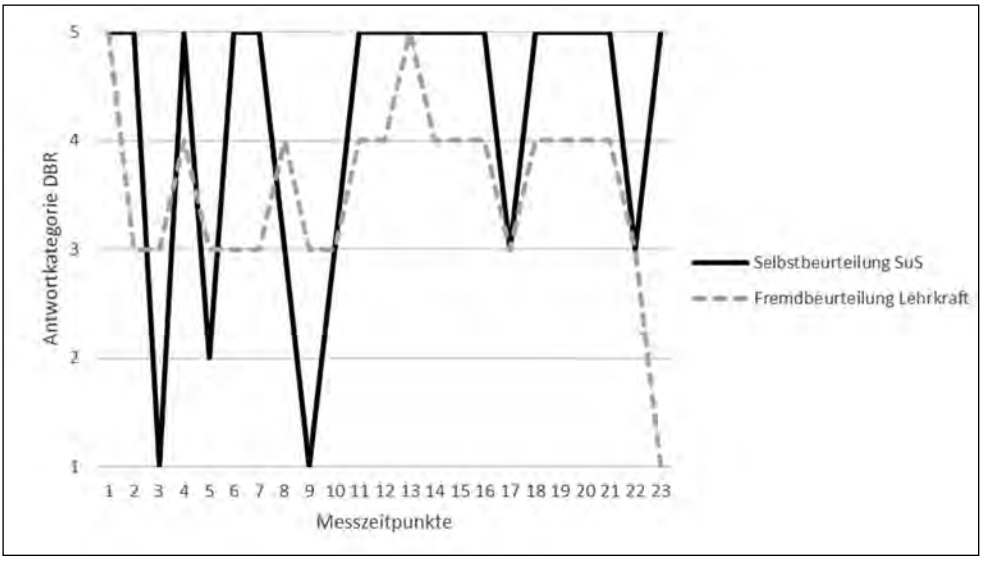

Abbildung 1. Beurteilungsübereinstimmung von Paar 07.

lung gut untersuchte Methode DBR auch zur Selbstbeurteilung im Primarbereich eingesetzt werden könnte. Da das Verhalten sowohl für die Kinder als auch für ihre Lehrkräfte eindeutig beurteilbar sein sollte, wurde als Zielverhalten die aktive Teilnahme am Unterricht gewählt. Die Ergebnisse zeigten, dass sich sowohl die numerische als auch die strukturelle Übereinstimmung der DBR_{fremd}- und DBR_{selbst}-Daten nur für einen Teil der Gesamtstichprobe nachweisen ließ. Für die ausbleibenden Effekte können unterschiedliche Gründe angenommen werden.

Ein Grund könnte in den bestehenden Deckeneffekten der Selbsteinschätzungen der Schülerinnen und Schüler liegen. Diesbezüglich können vier mögliche Ursachen angenommen werden:

1. Die Schülerinnen und Schüler überschätzten sich systematisch.
2. Die Selbsteinschätzung wurde durch die soziale Erwünschtheit systematisch verzerrt.
3. Das Skalenformat war nicht passend gewählt.
4. Die Teilnahme der Schülerinnen und Schüler am Unterricht war tatsächlich hoch.

Das Auftreten systematischer Überschätzungen der Grundschülerinnen und Grundschüler (Annahme 1) ist für statusdiagnostische Beurteilungen, auch für die Selbstbeurteilung schulischer Fähigkeiten, empirisch nachgewiesen (Filipp, 2006). Aus diesem Grunde wurden Schülerinnen und Schüler der vierten Jahrgangsstufe ausgewählt, da eine zunehmend realistischere Selbsteinschätzung zum Ende der Grundschulzeit angenommen wird (Weichbold, 2009). Gleichwohl lassen sich auch im Erwachsenenalter noch höhere Selbsteinschätzungen der eigenen Fähigkeiten im Vergleich zu Fremdeinschätzungen nachweisen (Gold & Kuhn, 2017; Harris & Schaubroeck, 1988; Von Stumm, 2014). Dieser Befund könnte eine mögliche Erklärung für die unzureichende numerische Übereinstimmung zwischen DBR_{selbst} und DBR_{fremd} sein. Die unzureichende strukturelle Übereinstimmung vieler Beurteilungspaare in den Verlaufsdaten kann dadurch jedoch nicht erklärt werden.

Die Deckeneffekte könnten ferner durch soziale Erwünschtheitsprozesse ausgelöst worden sein (Annahme 2). So könnten die Beurteilungen der Schülerinnen und Schüler durch die in Klassenräumen vorherr-

schenden Erwartungen nach hoher Teilnahme am Unterricht verzerrt worden sein. Es besteht für Schulkinder immer auch ein latenter Druck zur aktiven Teilnahme, was sich auch in den insgesamt höheren Selbstbeurteilungen der Schulkinder widerspiegeln könnte. Damit stellt sich auch die Frage, inwieweit die Teilnahme am Unterricht überhaupt ein geeignetes Kriterium zur Überprüfung der Beurteilungsübereinstimmung ist.

Mit Blick auf das Skalenformat (Annahme 3) empfehlen Briesch, Kilgus, Chafouleas, Riley-Tillman und Christ (2012) für Fremdbeurteilungen eine mindestens sechsstufe Skala, um Veränderungen differenziert abbilden zu können. Chafouleas (2011) rät sogar zum Einsatz einer zehnstufigen Skala. Die genutzte fünfstufige Skala war möglichweise zu undifferenziert, um Verhalten im Verlauf abbilden zu können. Für weitere Studien sollte daher eine differenziertere Skala mit mehr Abstufungen gewählt werden.

Neben den Schülerinnen und Schülern selbst schätzten auch die Lehrkräfte die aktive Teilnahme am Unterricht aus einer Fremdperspektive im Mittel als hoch ein. Huber und Rietz (2015a) führen in ihrem Review auf, dass für die Beurteilung der aktiven Teilnahme am Unterricht mittels DBR_{fremd} bereits zufriedenstellende empirische Befunde vorliegen. Daher ist anzunehmen, dass es sich um keine systematische Überschätzung der teilnehmenden Lehrkräfte handelt, sondern die Unterrichtsteilnahme der Schülerinnen und Schüler tatsächlich hoch war (Annahme 4).

Ein zweiter Grund für die ausbleibenden numerischen und strukturellen Beurteilungsübereinstimmungen könnte darin liegen, dass die Schwierigkeit, das Verhalten aus Selbst- und Fremdperspektive einzuschätzen; über die einzelnen Stufen der Skala variierte. So stimmten die Paare insgesamt in knapp einem Drittel ihrer Fremd- und Selbstbeurteilungen in ihren absoluten Werten überein. Hierbei ist ein Großteil der Übereinstimmungen in den Antwortkategorien 4 und 5 zu finden. Dies könnte darauf hindeuten, dass es für Lehrkräfte und Schülerinnen und Schüler der vorliegenden Studie leichter war, eine hohe aktive Teilnahme am Unterricht zu beurteilen als eine niedrige oder eine mittlere. Chafouleas, Jaffery, Riley-Tillman, Christ und Sen (2013) untersuchten den Einfluss unterschiedlicher Intensitäten (gering, mittel, hoch) von verschiedenen Verhaltensweisen auf die Beurteilungsübereinstimmung. Da für einige Verhaltensweisen bedeutsame Unterschiede in der Beurteilungsübereinstimmung in unterschiedlichen Intensitäten nachgewiesen werden konnte, sollten mögliche Herausforderungen bei der Beurteilung von Verhalten in niedrigen oder mittleren Intensitäten in Zukunft bedacht werden.

Der dritte Grund für die ausbleibenden Effekte könnte in der Wahl der Analysemethode liegen. Die Betrachtung der Einzelfälle zeigte, dass 14 der 18 Paare zu mehr als einem Drittel ihrer Beurteilungen strukturell übereinstimmten (Tabelle 5). Dies liefert einen ersten Hinweis auf eine ähnliche Wahrnehmung von strukturellen Veränderungen der aktiven Teilnahme am Unterricht im Verlauf aus Fremd- und Selbstperspektive. Es stellt sich die Frage, ob die ICC das richtige Maß für die Bewertung der Beurteilungsübereinstimmung zwischen DBR_{selbst}- und DBR_{fremd}-Daten darstellt. Aufgrund der Beurteilung aus Selbst- und Fremdperspektive konnte ein unterschiedlicher Fokus der teilnehmenden Personen, trotz gemeinsamer Operationalisierung im Vorfeld, nicht ausgeschlossen werden. Möglicherweise wirkten zudem im Klassenraum entstandene Störeinflüsse auf die Beurteilungen ein. Die Festlegung einer ICC ≥ .5 als Maß für eine zufriedenstellende Beurteilungsübereinstimmung wurde aufgrund ebendieser möglichen Störeinflüsse sowie den unterschiedlichen Beurteilungsperspektiven von Lehrkräften (DBR_{fremd}) und Kindern (DBR_{selbst}) gewählt. Diese Festlegung ist diskutabel und bedarf einer empirischen Absicherung.

Ein vierter Grund für die hier skizzierten Befunde könnte auch an einer fehlenden

Schulung der Studienteilnehmerinnen und -teilnehmer liegen. So lässt sich diskutieren, dass innerhalb der Studie aus ökonomischen Gründen keine ausführliche Schulung der teilnehmenden Personen im Vorfeld stattfand. In einigen Studien konnte eine Erhöhung der Interrater-Reliabilität durch ein vorheriges Training der teilnehmenden, beurteilenden Personen nachgewiesen werden (Chafouleas, Kilgus, Riley-Tillman, Jaffery & Harrison, 2012; Schlientz, Riley-Tillman, Briesch, Walcott & Chafouleas, 2009). In einer Folgestudie sollten die teilnehmenden Personen im Vorfeld geschult werden. In diesem Zuge könnte ebenfalls das Auftreten von verschiedenen Intensitäten einer Verhaltensweise thematisiert und deren Beurteilung konkret geübt werden. Bisherige Studien führten Trainings für den Einsatz von DBR aus einer Fremdperspektive anhand von Videosequenzen durch (z.B. Chafouleas et al., 2012). Ob dieses Format ebenfalls für das Training von Schülerinnen und Schülern in Bezug auf deren Selbstbeurteilung passend ist, wurde bis dato noch nicht überprüft.

Eine Betrachtung der Anzahl der Messzeitpunkte der einzelnen Paare (Tabelle 4) zeigte, dass lediglich ein Paar die theoretisch mögliche Anzahl von 80 Messwiederholungen erreichte. Die Gründe für die Unvollständigkeit der möglichen Messzeitpunkte lagen zum Teil in anderweitigen Schul- und Unterrichtsaktivitäten begründet. Des Weiteren konnte aufgrund von Fehlzeiten der Lehrkräfte oder der Schülerinnen und Schüler nicht immer eine Beurteilung durchgeführt werden. Die vorliegenden Ergebnisse geben einen ersten Hinweis darauf, dass eine Beurteilung mit vier Messzeitpunkten täglich womöglich nicht notwendig ist, um eine zufriedene Interrater-Reliabilität zu erzielen. Keines der fünf Paare mit einer $ICC_{justmittel} \geq .5$ erreichte die maximale Anzahl von 80 Messwiederholungen. Vielmehr repräsentierte Paar 04 ($ICC_{justmittel} = .532$) mit 19 Messzeitpunkten sogar das Minimum der gemeinsamen Messzeitpunkte der vorliegenden Stichprobe.

Es bleibt die Frage offen, ob die einzelnen Beurteilungspaare mit hohen Beurteilungsübereinstimmungen gegenüber den anderen Paaren spezifische Merkmale aufweisen. Aus den vorliegenden Daten lassen sich diesbezüglich keine Schlüsse ableiten. Es wäre möglich, dass spezifische Merkmale der Lehrkräfte (z.B. Berufserfahrung, Anzahl der gemeinsamen Stunden mit dem jeweiligen Schulkind), der Schülerinnen und Schüler (z.B. Selbstkonzept, Merkfähigkeit) oder des Unterrichts (z.B. Classroom Management, Schulfach, Zeitpunkt der Beurteilung) die Übereinstimmungen der Selbst- und Fremdbeurteilungen beeinflussen. In weiteren Studien sollten mögliche weitere Einflussfaktoren daher miterfasst werden.

Aufgrund der geringen Stichprobe können keine generalisierbaren Aussagen zur Testgüte von Selbstratings getroffen werden. Die vorliegende Studie liefert lediglich erste Hinweise darauf, dass der Einsatz von DBR_{selbst} durch Schülerinnen und Schüler zum Ende des Primarbereichs grundlegend möglich ist.

Offen bleibt zudem die Frage, ob eine kontinuierliche Beobachtung und Beurteilung des eigenen Verhaltens im Verlauf der Zeit, wie sie im DBR stattfindet, zu einer besseren Selbstwahrnehmung führt. In diesem Fall wäre DBR_{selbst} nicht nur als ein Instrument, sondern auch als eine Intervention zu betrachten. Sollte sich dies empirisch bestätigen lassen, so wäre hierin ein weiterer großer Vorteil von Verlaufsdiagnostik aus Selbstperspektive zu sehen.

Der Einsatz von DBR_{selbst} könnte eine ressourcenschonende Möglichkeit bieten, um Verhalten im Verlauf aus Perspektive der Schülerinnen und Schüler zu erfassen und die Entwicklung abzubilden. Somit könnte das pädagogische Handeln passgenauer auf die Bedürfnisse der Schülerinnen und Schüler abgestimmt werden und auf bisher schwer zu beobachtende Aspekte erweitert werden.

Literaturverzeichnis

Bortz, J., & Döring, N. (2016). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (vollst. überarb., akt. u. erw. Aufl. 2016). Heidelberg: Springer.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and Direct Behavior Rating. *School Psychology Review, 39*, 408–421.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. and Contributors. (2016). *Direct Behavior Rating: Linking assessment, communication, and intervention*. New York: The Guilford Press.

Briesch, A. M., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2012). The influence of alternative scale formats on the generalizability of data obtained from Direct Behavior Rating single-item scales (DBR-SIS). *Assessment for Effective Intervention, 38*, 127–133. https://doi.org/10.1177/1534508412441966

Casale, G. (2017). „Nützt es was oder nützt es nichts?": Direct Behavior Rating (DBR) als diagnostische Methode zur zeitnahen Überprüfung des Fördererfolgs bei unterrichtlichem Schülerinnen- und Schülerverhalten. *Potsdamer Zentrum Für Empirische Inklusionsforschung (ZEIF),* (1).

Casale, G., Grosche, M., Volpe, R. J., & Hennemann, T. (2017). Zuverlässigkeit von Verhaltensverlaufsdiagnostik über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen. *Empirische Sonderpädagogik, 9*, 143–164.

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

Casale, G., Hennemann, T., Huber, C., & Grosche, M. (2015). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt emotionale und soziale Entwicklung. *Heilpädagogische Forschung, 41*, 37–45.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education and Treatment of Children, 3*, 575–591.

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

Chafouleas, S. M., Jaffery, R., Riley-Tillman, T. C., Christ, T. J., & Sen, R. (2013). The impact of target, wording, and duration on rating accuracy for Direct Behavior Rating. *Assessment for Effective Intervention, 39*, 39–53. https://doi.org/10.1177/1534508413489335

Chafouleas, S. M., Kilgus, S. P., Riley-Tillman, T. C., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of Direct Behavior Ratings. *Journal of School Psychology, 50*, 317–334. https://doi.org/10.1016/j.jsp.2011.11.007

Chafouleas, S. M., Kilgus, S. P., & Wallach, N. (2010). Ethical dilemmas in school-based behavioral screening. *Assessment for Effective Intervention, 35*, 245–252. https://doi.org/10.1177/153450841037 9002

Chafouleas, S. M., Riley Tillman, T. C., & Christ, T. J. (2009). Direct Behavior Rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention*, *34*, 195–200. https://doi.org/10.1177/1534508409340391

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*, 201–213. https://doi.org/10.1177/1534508409340390

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Jaffery, R. (2011). Direct Behavior Rating: An evaluation of alternate defini-

tions to assess classroom behaviors. *School Psychology Review, 40*, 181–199.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurements to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507–524.

Filipp, S.-H. (2006). Kommentar zum Schwerpunktthema: Entwicklung von Fähigkeitsselbstkonzepten. (1/2), 65–72. https://doi.org/10.1024/1010-0652.20. 12. 65

Förster, N., Kuhn, J.-T., & Souvignier, E. (2017). Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, *9*, 116–122.

Gold, B., & Kuhn, J.-T. (2017). A longitudinal study on the stability of self-estimated intelligence and its relationship to personality traits. *Personality and Individual Differences*, *106*, 292–297. https://doi.org/10.1016/j.paid.2016.10.052

Gross Portney, L., & Watkins, M. P. (2014). *Foundations of clinical research*. *Application to practice*. Pearson New International Edition. Edinburgh: Pearson Education Limited.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43–62.

Hellmich, F., & Günther, F. (2011). *Entwicklung des Selbstkonzepts im Grundschulalter*. In F. Hellmich (Ed.), *Selbstkonzepte im Grundschulalter: Modelle, empirische Ergebnisse, pädagogische Konsequenzen* (S. 17–46). Stuttgart: W. Kohlhammer.

Helmke, A. (1998). *Vom Optimisten zum Realisten? Zur Entwicklung des Fähigkeitsselbstkonzeptes vom Kindergarten bis zur 6. Klassenstufe*. In F. E. Weinert (Ed.), *Entwicklung im Kindesalter* (S. 115–132). Weinheim: Psychologie Verlags Union.

Huber, C. (2016). Verhaltensverlaufsdiagnostik. In K. Seifried, S. Drewes, & M. Hasselhorn (Eds.), *Handbuch Schulpsychologie: Psychologie für die Schule* (2nd ed.). Stuttgart: Verlag W. Kohlhammer.

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

Huber, C., & Rietz, C. (2015b). Behavior assessment using Direct Behavior Rating (DBR) – a study on the criterion validity of DBR single-item-scales. *Insights into Learning Disabilities, 12*, 73–90.

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

Klasen, F.; Petermann, F.; Meyrose, A. Barkmann, C.; Otto, C.; Haller, A. et al. (2016). Verlauf psychischer Auffälligkeiten von Kindern und Jugendlichen. *Kindheit und Entwicklung, 25*, 10–20. https://doi.org/10.1026/0942-5403/a000184

Möller, J., & Trautwein, U. (2009). Selbstkonzept. In E. Wild (Ed.), *Pädagogische Psychologie* (1. Aufl., S. 177–200). Berlin: Springer.

Nicolls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development*, *49*, 800–814.

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A., & Glazer, A. D. (2008). Examining the agreement of Direct Behavior Ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions, 10*, 136–143.

Riley-Tillman, T. C., Christ, T. J., Chafouleas, S. M., Boice Mallach, C. H., & Briesch, A. (2011). The impact of observation duration on the accuracy of data obtained from Direct Behavior Rating (DBR). *Journal of Positive Behavior Interventions, 13*, 119–128. https://doi.org/10.1177/1098300710361954

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

racy of Direct Behavior Ratings (DBR). *School Psychology Quarterly, 24*, 73–83. https://doi.org/10.1037/a0016255

Shoukri, M., Asyali, M., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*, 251-271. https://doi.org/10.1191/0962280204sm365ra

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item Direct Behavior Rating scales for engagement and disruptive behavior. *School Psychology Review, 41*, 246–261.

Von der Embse, N. P., Scott, E.-C., & Kilgus, S. P. (2015). Sensitivity to change and concurrent validity of Direct Behavior Ratings for academic anxiety. *School Psychology Quarterly, 30*, 244–259. https://doi.org/10.1037/spq0000083

Von Stumm, S. (2014). Intelligence, gender, and assessment method affect the accuracy of self-estimated intelligence. *British Journal of Psychology (London, England: 1953)*, *105*, 243–253. https://doi.org/10.1111/bjop.12031

Weichbold, M. H. (2009). Umfrageforschung: Herausforderungen und Grenzen. *Österreichische Zeitschrift für Soziologie. Sonderheft, Teil 9*. Wiesbaden: VS, Verl. für Sozialwiss.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe Verlag GmbH & Co KG

Simone Weber
Bergische Universität Wuppertal
School of Education
Institut für Bildungsforschung
Gaußstraße 20
42119 Wuppertal
E-Mail: siweber@uni-wuppertal.de

Erstmalig eingereicht: 01.04.2019
Überarbeitung eingereicht: 14.10.2019
Angenommen: 08.06.2020