Empirische Sonderpädagogik, 2017, Nr. 2, S. 143-164
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Zuverlässigkeit von Verhaltensverlaufsdiagnostik über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen

Gino Casale[1], Michael Grosche[2], Robert J. Volpe[3] & Thomas Hennemann[4]

[1] *Universität Paderborn*
[2] *Bergische Universität Wuppertal*
[3] *Northeastern University Boston*
[4] *Universität zu Köln*

Zusammenfassung

Die vorliegende Studie untersucht die Zuverlässigkeit von Verhaltensverlaufsdiagnostik mittels Direct Behavior Rating (DBR) über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen. Im Rahmen einer vollständig gekreuzten Zwei-Facetten-Generalisierbarkeitsstudie wurde das Verhalten von fünf Viertklässlern mit externalisierendem Problemverhalten (Facette Schüler) von einer Regelschullehrerin, sowie einer Lehrerin für sonderpädagogische Förderung (Facette Rater) zu insgesamt acht Unterrichtssituationen à zehn Minuten (Facette Messzeitpunkt) jeweils mit einer Single-Item-Skala (SIS) und einer Multi-Item-Skala (MIS) beurteilt. Die Ergebnisse weisen auf einen starken Einfluss der Messzeitpunkte auf die Zuverlässigkeit der Messung sowohl bei den SIS als auch bei den MIS hin. Die Facette Rater klärt hingegen nur bei den SIS einen substanziellen Anteil der Gesamtvarianz auf. Die Ergebnisse einer Simulationsstudie über eine unterschiedliche Anzahl an Messzeitpunkten zeigen, dass zuverlässige Ergebnisse zur Interpretation intraindividueller Verhaltensverläufe nur mittels MIS in einem überschaubaren Zeitraum (13 Messzeitpunkte) erzielt werden. Die Ergebnisse ermutigen für den praktischen Einsatz der Methode zur Evaluation des Erfolgs einer Verhaltensfördermaßnahme.

Schlagwörter: Direct Behavior Rating, Generalisierbarkeitstheorie, Verlaufsdiagnostik

Dependability of Direct Behavior Ratings across Rater and Occasion in Students with Externalizing Behavior Problems

Abstract

The purpose of the current study was the analysis of the reliability of Direct Behavior Ratings (DBR) across raters and occasions for students with externalizing problem behaviors. A fully-crossed Generalizability study design was conducted, wherein one female regular classroom teacher as well as one female special education teacher (facet rater) rated the externalizing problem behavior of five male fourth-graders (facet students) in eight instructional phases of independent seatwork (facet occasion) with a single item scale (SIS) and a multi item scale (MIS). The results indicate a strong influence of the time-related factor (occasion) on the reliability of both SIS and MIS. The facet rater contributes a substantial part of the total variance for the SIS. Results

of a decision study across a different number of occasions yield dependable results for interpreting intra-individual behavior progress for the MIS after 13 occasions. The results strengthen the practical use of DBR to collect progress monitoring data in order to evaluate a students' individuals response to an intervention.

Key words: Direct Behavior Rating, Generalizability Theory, progress monitoring

Die Verlaufsdiagnostik des Verhaltens von Schülerinnen und Schülern schafft eine Datengrundlage für pädagogisch-praktische Entscheidungen über den Erfolg oder Misserfolg von Verhaltensfördermaßnahmen. Durch die Anwendung entsprechender diagnostischer Verfahren kann im besten Falle sehr schnell eine Einschätzung darüber erfolgen, ob und in welchem Maße eine Intervention bei dem Kind den gewünschten Fördererfolg erzielt. Damit stellt die Verhaltensverlaufsdiagnostik ein wesentliches Element im Rahmen der Umsetzung mehrstufiger Fördersysteme, wie z. B. Response-to-Intervention (RtI), dar und liefert damit einen maßgeblichen Beitrag zur Umsetzung einer evidenzbasierten und präventiven (sonder-) pädagogischen Praxis (Casale, Hennemann & Grosche, 2015; Huber & Grosche, 2012).

Im Rahmen mehrstufiger Fördersysteme wird die Verhaltensverlaufsdiagnostik für eine Auswahl (schätzungsweise ca. 5 – 20%) der Schülerinnen und Schüler einer Klasse umgesetzt (Reschly & Bergstrom, 2008). In der Regel sind dies die Schülerinnen und Schüler, deren Verhalten z. B. durch ein universelles Verhaltensscreening als problematisch nachgewiesen wurde und dementsprechend einen erhöhten Bedarf an zusätzlichen Verhaltensfördermaßnahmen aufweisen (Hawken, Vincent & Schumann, 2008). Die Funktion der Verhaltensverlaufsdiagnostik auf den nächsthöheren Stufen ist es dann, die Verhaltensentwicklung der Schülerinnen und Schüler nach Einführung der Fördermaßnahmen zu überprüfen (Grosche, 2014). Dies sollte möglichst zeitnah erfolgen, damit eine unwirksame Förderung modifiziert werden kann. Damit erfüllt die Verhaltensverlaufsdiagnostik eine Schlüsselfunktion im Rahmen mehrstufiger Fördersysteme, nämlich die Evaluation der individuellen Wirksamkeit (*response*) einer Förderung (*intervention*) im Einzelfall (Casale et al., 2015[a]).

Die Verhaltensverlaufsdiagnostik kann mehrmals täglich angewendet werden und generiert damit eine hohe Anzahl an Messzeitpunkten. Daher sollten Methoden der Verhaltensverlaufsdiagnostik dem Kriterium der *Ökonomie* entsprechen, d. h. sie sollten im Schulalltag häufig und schnell umsetzbar sein (Casale, Hennemann, Huber & Grosche, 2015). Weiterhin werden die Ergebnisse der Verhaltensverlaufsdiagnostik für pädagogisch-praktisch relevante Förderentscheidungen (z. B. Fortführung oder Modifikation einer Fördermaßnahme) genutzt. Daher sollten Methoden der Verhaltensverlaufsdiagnostik in der Lage sein, das Verhalten der Schülerinnen und Schüler in einem vertretbaren Zeitraum *reliabel, valide und möglichst genau* zu erfassen (Casale et al., 2015[b]).

Sowohl international (Christ, Riley-Tillman & Chafouleas, 2009) als auch im deutschsprachigen Raum (Casale et al., 2015; Huber & Rietz, 2015) liegen mittlerweile Überblicksarbeiten vor, die traditionelle Ansätze zur Verhaltensdiagnostik hinsichtlich der Eignung dieser Verfahren für die Verlaufsdiagnostik analysierten. Dabei lassen sich grundsätzlich zwei zentrale Ergebnisse konstatieren: a) sowohl die systematisch-direkte Verhaltensbeobachtung als auch die Verhaltensbeurteilung mit Ratingskalen sind potenziell für die Verhaltensverlaufsdiagnostik geeignet. Allerdings lassen sich b) bei beiden Verfahren Nachteile hinsichtlich der genannten Gütekriterien identifizieren, die den Einsatz zur Verlaufsdiagnostik im o. g. Sinne erheblich einschränken. Systematisch-direkte Verhaltensbeob-

achtungen bspw. sind in der Vorbereitung, Durchführung und Auswertung i. d. R. sehr aufwändig, da sie in vivo während des Auftretens des Verhaltens im Minutentakt erfolgen. Oftmals erfordert dies einen externen und geschulten Beobachter im Klassenraum, der ausschließlich für die Beobachtung zuständig ist. Somit wäre das Kriterium der Ökonomie nicht erfüllt. Die Verhaltensbeurteilungen mit Ratingskalen sind hingegen sehr ökonomisch, da konkrete Verhaltensweisen anhand einer Likert-Skala beurteilt werden. Allerdings ist diese Art der Verhaltensmessung stark fehlerbehaftet, da typische Urteilsfehler (z. B. Strenge-/Mildefehler, Halo-Effekte) mit bis zu 40% an Varianzaufklärung zu Buche schlagen (Hoyt & Kerns, 1999). Damit wäre die Verhaltensbeurteilung mit Ratingskalen wenig zuverlässig, relativ ungenau und ließe nur eingeschränkt valide Interpretationen der Befunde zu (z. B. Iovannone, Greenbaum, Wang, Dunlap & Kincaid, 2014).

Direct Behavior Rating als Methode der Verhaltensverlaufsdiagnostik

Diesen Aspekten zufolge wird in den o. g. Überblicksarbeiten geschlussfolgert, dass eine Kombination aus beiden diagnostischen Zugängen – systematisch-direkte Verhaltensbeobachtung sowie Verhaltensbeurteilung mit Ratingskalen – das Potential zur Verhaltensverlaufsdiagnostik haben könnte (v. a. Christ et al., 2009). Das Direct Behavior Rating (DBR) stellt eine Methode dar, die beide Ansätze wie folgt miteinander verbindet. Ein konkret operationalisierter Verhaltensausschnitt einer Schülerin/ eines Schülers (z. B. konzentriertes Arbeiten) wird in einer Situation, in der dieses Verhalten relevant ist (z. B. Stillarbeitsphasen), beobachtet und unmittelbar im Anschluss an diese Situation anhand einer Ratingskala beurteilt. Dies kann im besten Falle mehrfach täglich im Schulalltag erfolgen. An dieser Stelle ist wichtig anzumerken, dass sowohl die Auswahl der zu beurteilenden Verhaltensweisen als auch die Situationen abhängig von den individuellen Ausgangslagen der Schülerinnen und Schüler stark variieren können (Christ et al., 2009). Dadurch ist die Methode sehr flexibel.

Grundsätzlich lassen sich zwei Formen von DBR unterscheiden: Single-Item-Skalen (SIS) und Multi-Item-Skalen (MIS) (Christ et al., 2009). Single-Item-Skalen zeichnen sich dadurch aus, dass mit nur einem einzigen Item eine übergeordnete Verhaltensdimension (z. B. Arbeitsverhalten, Störverhalten) erfasst wird. Damit stellt die SIS eine ökonomische Methode dar, um einen sehr globalen Verhaltensausschnitt zu erfassen. Dies ist in der Praxis besonders dann nützlich, um bei einer Schülerin/ einem Schüler ein eher breit gefächertes Verhaltensproblem zu beurteilen (Chafouleas, Kilgus & Hernandez, 2009; Miller et al., 2015). Die SIS eignet sich jedoch nur begrenzt, sehr konkrete und spezifische Verhaltensweisen (z. B. Sich melden, Konzentriert arbeiten) zu erfassen. Diese spezifischen Verhaltensmerkmale sind jedoch besonders dann hilfreich, wenn es um die konkrete Information über den individuellen Fördererfolg bei Schülerinnen und Schülern geht. Für diese Zwecke eignet sich die MIS besser, da diese in der Regel drei bis fünf spezifische Verhaltensweisen (z. B. Beginnt selbstständig mit der Aufgabenbearbeitung, Beendet Aufgaben in der vorgegeben Zeit) zur Erfassung einer übergeordneten Verhaltensdimension (z. B. lernförderliches Verhalten) operationalisiert. Je nach Bedarfslage können die Ergebnisse der einzelnen Items individuell analysiert oder zu einem Summenscore aufaddiert werden (Volpe & Briesch, 2012).

Konzeptionell stellt DBR demnach eine ökonomische und flexible Methode dar, da sie erstens keine teuren Materialien benötigt, zweitens in wenigen Minuten zu bearbeiten ist und drittens eine große Bandbreite an Verhaltensweisen in einer Vielzahl von Situation erfassen kann (Christ et al., 2009). Empirisch zu prüfen bleiben allerdings die psychometrischen Eigenschaften der Methode, also die Zuverlässigkeit über verschiedene Bedingungen, die Genauig-

keit der Messung sowie die Validität der erzielten Ergebnisse (Chafouleas, 2011).

Psychometrische Anforderungen an DBR

Im Kontext der Verhaltensverlaufsdiagnostik wird unter psychometrischer Qualität verstanden, dass ein Verfahren reliabel, valide und möglichst genau den individuellen Verhaltensverlauf einer Schülerin/ eines Schülers erfassen kann (Christ, Nelson, Van Norman, Chafouleas & Riley-Tillman, 2014). Reliabilität bezieht sich auf die Ausprägung des zufälligen, also nicht erklärbaren Fehlers während einer Messung (Salvia, Ysseldyke & Witmer, 2012, S. 54). Je geringer das Ausmaß des zufälligen und damit nicht erklärbaren Messfehlers, desto höher ist die Reliabilität eines Verfahrens. Unter Validität versteht man im Allgemeinen die Plausibilität der Interpretation eines Testwertes auf Basis empirischer und theoretischer Evidenz (Kane, 2013; Salvia et al., 2012). Bei der Verhaltensverlaufsdiagnostik bedeutet dies, dass die Testwerte genutzt werden können, um die individuelle Verhaltensentwicklung einer Schülerin/ eines Schülers zu erfassen und daraus Förderentscheidungen abzuleiten (Fan & Hansmann, 2015; Grosche, 2014). Die Messgenauigkeit schließlich ist definiert als das Ausmaß der Sensitivität eines Messinstruments in Bezug auf die objektivierbare Ausprägung des Verhaltens (z. B. Dauer, Häufigkeit; Cone, 1998). Im Kontext der Verhaltensverlaufsdiagnostik bedeutet dies, dass DBR in der Lage ist, möglichst genau die tatsächliche Auftretensdauer oder – häufigkeit der Zielverhaltensweisen zu erfassen.

Wie dargestellt, stellt DBR eine sehr flexible diagnostische Methode dar. So können z. B. verschiedene Verhaltensweisen in verschiedenen Situationen von verschiedenen Personen je nach individuellem Bedarf erfasst werden. Diese starke Flexibilität geht jedoch gleichzeitig mit großen testtheoretischen Herausforderungen bei der Überprüfung der psychometrischen Qualität einher, da die Methode unter verschiedenen Messbedingungen (z. B. unterschiedliche Verhaltensweisen, unterschiedliche Situationen, unterschiedliche Rater) angewendet werden kann. In diesem Kontext spielt vor allem die Zuverlässigkeit über verschiedene Messbedingungen, also die Reliabilität, eine wichtige Rolle bei der Überprüfung der psychometrischen Qualität von Verhaltensverlaufsdiagnostik (Fan & Hansmann, 2015). Im vorliegenden Beitrag möchten wir daher den Aspekt der Reliabilität genauer betrachten und a) die Generalisierbarkeitstheorie (GT) als Ansatz zur Überprüfung der Reliabilität von verlaufsdiagnostischen Verfahren vorstellen, b) den aktuellen Forschungsstand zur Reliabilität von DBR skizzieren und c) die Anwendung der GT zur Reliabilitätsprüfung von DBR anhand einer eigenen Studie exemplifizieren.

Die Generalisierbarkeitstheorie als Ansatz zur Reliabilitätsprüfung von DBR

Wie bereits erwähnt, hängt die Reliabilität mit der Stärke des zufälligen und damit nicht erklärbaren Fehlers während einer Messung zusammen. Diese Definition impliziert, dass es auch einen nicht zufälligen und damit bekannten Messfehler gibt, der auch als erklärbare Merkmalsvarianz bezeichnet werden kann. Im Kontext der Verhaltensdiagnostik stellen z. B. die beurteilenden Personen (Inter-Rater-Reliabilität), die genutzten Items (interne Konsistenz) oder die Messzeitpunkte (Test-Retest-Reliabilität) solche erklärbaren Varianzquellen dar (Hintze, 2005).

Im Paradigma der Klassischen Testtheorie (KTT) steht bei der Reliabilitätsprüfung in der Regel immer nur einer dieser Faktoren im Vordergrund, d. h. ein empirisch beobachteter Wert setzt sich aus dem wahren Wert und dem Messfehler, der global mit einem Faktor erklärt wird, zusammen (Shavelson & Webb, 1991). Interessiert bei der Evaluation eines Verhaltensratings bspw., wie zuverlässig die entwickelten Items das Ver-

haltenskonstrukt erfassen, wird der Anteil an Gesamtvarianz, der durch die Items erklärt werden kann, statistisch über die Berechnung von Maßen zur internen Konsistenz (z. B. Cronbach's Alpha) geprüft. Ein Nachteil eines solchen Vorgehens ist unter anderem, dass alle anderen Faktoren, die während der Messung auf den Messfehler wirken, außer Acht gelassen werden. So könnte es z. B. sein, dass die Items durchaus zuverlässig eine Verhaltensdimension erfassen, die Bearbeitung der Items von verschiedenen Ratern mit unterschiedlichen Verhaltensnormen jedoch sehr unterschiedlich erfolgt. Als Ergebnis würde dann ein geringer Reliabilitätsindex der internen Konsistenz entstehen, da durch das statistische Modell nur ein kleiner Anteil der Fehlervarianz aufzuklären wäre. Empirisch erklärbar wäre dieser Befund aber tatsächlich nicht über eine mangelnde Zuverlässigkeit der Items, sondern über Unterschiede zwischen den beurteilenden Personen.

Ausgehend von dieser Schwäche der KTT wurde die GT entwickelt (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Sie stellt eine Erweiterung des Paradigmas in der KTT dar und zwar dergestalt, dass der Messfehler nicht wie in der KTT als global und unbekannt, sondern als spezifisch und bekannt vorausgesetzt wird. Daher ist es im Rahmen der GT möglich den Einfluss mehrerer Faktoren (z. B. Items und Rater) auf den Messfehler sowie deren Interaktionen simultan zu schätzen (Brennan, 2001). Zum Verständnis des Ansatzes ist es wichtig, zwischen *konzeptionellen Grundlagen* und der *methodischen Vorgehensweise* der GT zu unterscheiden.

Konzeptionelle Grundlagen der GT

Die GT stellt zwar eine Erweiterung der KTT dar, basiert aber auf spezifischen konzeptionellen Grundannahmen, die sich auch in einer eigenen Terminologie niederschlagen. Die wohl wichtigste Grundannahme stellt jene dar, die besagt, dass sich der Messfehler bei empirischen Beobachtungen nicht global, sondern spezifisch durch mehrere Faktoren simultan zusammensetzt. Diese Faktoren werden in der GT auch als *Facetten* bezeichnet, die als "set of similar conditions of measurement" (Brennan, 2001, S. 5) definiert werden. Theoretisch gibt es eine unendlich große Anzahl an diesen vergleichbaren Bedingungen, die innerhalb einer Facette modelliert werden können. Daher spricht man in der GT vom *Universum aller zulässigen Bedingungen*, d. h. die Bedingungen innerhalb einer Facette stellen eine repräsentative Auswahl aus diesem Universum dar. In der Terminologie der KTT würden die Facettenstufen als *Stichprobe* und das Universum aller zulässigen Bedingungen als *Population* bezeichnet. Die Bedingungen, aus denen sich die Facetten zusammensetzen, werden hinsichtlich der Fragestellung und des Forschungsinteresses ausgewählt. Sie stellen i. d. R. eine Zufallsauswahl aus dem Universum aller zulässigen Bedingungen dar (Brennan, 2001). Ausgehend von diesem Grundverständnis wird in der GT geprüft, wie stark der simultane Einfluss verschiedener Facetten, und deren jeweiligen Bedingungen auf den Messfehler eines empirisch beobachteten Wertes ist.

Ein wichtiger Aspekt bei der Auswahl der Facetten stellt deren Klassifikation dar. Es wird zwischen *zufälligen* und *festen* Facetten unterschieden. Von zufälligen Facetten spricht man, wenn man von einer grundsätzlichen Austauschbarkeit der Facettenbedingungen in weiteren Replikationsstudien ausgeht (z. B. andere Rater oder Items), d. h. die gewählten Bedingungen stellen nur eine Auswahl einer Population dar, über die Generalisierungen vorgenommen werden sollen (Briesch, Swaminathan, Welsh & Chafouleas, 2014). Im Gegensatz dazu wird eine Facette als fest klassifiziert, wenn die gewählten Facettenbedingungen in Replikationsstudien nicht austauschbar sind und Generalisierungen ausschließlich über die gewählten Bedingungen erfolgen sollen (z. B. nur eine spezifische Testaufgabe; Briesch et al., 2014).. Da ein zentrales Anliegen der GT allerdings die Generalisier-

barkeit über alle zulässigen Bedingungen eines Universums ist, werden in schulischen Settings Facetten zumeist nur dann als fest klassifiziert, wenn es sich um etablierte und standardisierte Testaufgaben handelt, die nicht austauschbar sind (Briesch et al., 2014). Allen anderen potenziellen Facetten (Rater, Items, Messzeitpunkte usw.) wird grundsätzlich eine Austauschbarkeit unterstellt.

Methodische Vorgehensweise der GT

Das methodische Vorgehen innerhalb der GT gliedert sich grob in zwei Schritte. Der erste Schritt erfolgt in Form einer sogenannten Generalisierbarkeitsstudie (G-Studie). G-Studien schätzen die Varianz der einzelnen Facetten und deren Interaktionen untereinander, um festzustellen, in welchem Ausmaß sie zum Messfehler beitragen. Dies geschieht mittels Varianzanalysen, wobei die Facetten als Faktoren und die Bedingungen der Facetten als Faktorstufen behandelt werden (Brennan, 2001). In diesem Kontext muss auf eine weitere Liberalisierung der GT im Gegensatz zur KTT eingegangen werden. Die statistische Modellierung der Faktoren in den Varianzanalysen als zufällige oder feste Faktoren orientiert sich in der GT an der theoretisch angenommenen Klassifikation der Facette als zufällig oder fest (Brennan, 2001). Dies bedeutet, dass die Varianzanalysen in der GT dahingehend von Einschränkungen befreit werden, dass auch Facetten, die im Verständnis der KTT niemals als zufällige Faktoren modelliert würden (z. B. aufgrund einer nicht zufälligen Auswahl der Faktorenstufen) in der GT durchaus als zufällig angenommen werden dürfen (Brennan, 2001). Aus diesem Grund wird die GT auch als Liberalisierung der KTT bezeichnet (Cronbach, Rajaratnam & Gleser, 1963).

Im zweiten Schritt werden dann die Ergebnisse der G-Studie als Ausgangspunkt für so genannte Entscheidungsstudien (D-Studie; *decision study*) genutzt. In der D-Studie wird simuliert, wie sich die Varianzaufklärung verändern würde, wenn man die Anzahl der zulässigen Bedingungen innerhalb bestimmter Facetten variiert. Zusätzlich werden zwei Gütemaße zur Beurteilung der Reliabilität berechnet. Der Generalisierbarkeitskoeffizient (εp^2) gibt an, wie viel Prozent der relativen Fehlervarianz durch das gewählte Messmodell erklärt werden kann. Er kann damit als Grundlage für pädagogisch-praktische Entscheidungen, die sich auf relative Vergleiche beziehen, genutzt werden (z. B. die Identifikation der Schülerinnen und Schüler mit dem größten Problemverhalten in einer Klasse). εp^2 eignet sich dementsprechend eher als Kriterium für die Reliabilität eines Screeningverfahrens. Wichtiger für die Verlaufsdiagnostik ist der Zuverlässigkeitsindex Φ, der angibt, wie viel Prozent der absoluten Fehlervarianz durch das gewählte Modell erklärt werden kann. Damit kann er als Grundlage für pädagogisch-praktische Entscheidungen dienen, die sich auf intra-individuelle Vergleiche beziehen (z. B. den Verhaltensverlauf einer Schülern/ eines Schülers gemessen über mehrere aufeinander folgende Messzeitpunkte). Damit eignet sich Φ als Kriterium für die Beurteilung der Reliabilität eines Verfahrens für verlaufsdiagnostische Zwecke (Hintze, Owen, Shapiro & Daly, 2000). Beide Koeffizienten können Werte zwischen 0 (0% Varianzaufklärung durch das Modell) und 1 (100% Varianzaufklärung durch das Modell) annehmen (Brennan, 2001). Für verlaufsdiagnostische Zwecke wird häufig ein Wert von .70 als Minimalstandard definiert (Salvia et al., 2012).

Forschungsstand zur Reliabilität von DBR

Aufgrund der genannten Vorteile (Möglichkeit der simultanen Schätzung des Varianzanteils verschiedener Facetten, Simulation der Facettenbedingungen zur Verbesserung der Reliabilität, Möglichkeit der Berechnung eines Reliabilitätskoeffizienten als Kriterium für intra-individuelle Entwicklungen) wurde die GT seit 2007 in insgesamt elf Stu-

dien zur Reliabilitätsprüfung von DBR eingesetzt. Am häufigsten wurde überprüft: a) welchen Einfluss die Rater und die Messzeitpunkte auf die Beurteilung des Verhaltens von Schülerinnen und Schülern haben und b) nach wie vielen DBR-Messungen reliable Ergebnisse erzielt werden können (Huber & Rietz, 2015b).

Einfluss der Rater

Die Varianz, die durch die Facette der Rater erklärbar ist, basiert auf Unterschieden in den Werten der einbezogenen Rater. Ein großer Anteil an Ratervarianz würde dementsprechend bedeuten, dass die verschiedenen Rater das Verhalten der Schülerinnen und Schüler sehr unterschiedlich beurteilen. Dies wäre für die praktische Umsetzung von Verhaltensverlaufsdiagnostik in der Schule erst einmal unproblematisch, da es unrealistisch ist, dass verschiedene Lehrkräfte das Verhalten ihrer Schülerinnen und Schüler absolut identisch beurteilen (numerisches Invarianzkonzept; Huber & Rietz, 2015a). Wichtiger ist in diesem Zusammenhang, dass die gleichen Trends (z. B. die Reduktion des Störverhaltens im Unterricht nach der Implementation einer Förderung) bei einer Schülerin/ einem Schüler erkannt und beurteilt werden (strukturelles Invarianzkonzept; Huber & Rietz, 2015a). Daher stellt eine geringe bis moderate Varianzaufklärung durch die Facette Rater ein akzeptables Resultat dar.

Bisherige Studien zur Reliabilität von DBR konnten zeigen, dass die Varianzaufklärung durch die Facette Rater je nach Personengruppe zwischen 0 und 41% schwankt. Chafouleas et al. (2010) verglichen bspw. die Ergebnisse aus DBR-SIS von sechs aufeinanderfolgenden Schultagen durch zwei Mittelstufenlehrkräfte und zwei geschulte Beobachter. Die Ratervarianz betrug 1 bis 8% in den Lehrkraftratings, allerdings nur 0 bis 1% in den Ratings der Beobachter. Die Interaktionen mit der Facette Rater konnten keine bedeutsame Varianz aufklären. In einer weiteren Studie wurde das Verhalten von Schülerinnen und Schüler der Vorschule mit DBR-SIS an 13 aufeinanderfolgenden Schultagen durch die Lehrkräfte beurteilt (Chafouleas, Christ, Riley-Tillman, Briesch & Chanese, 2007). Ein bedeutsamer Anteil an Varianz (20 bis 41%) konnte auf die Rater zurückgeführt werden. Auch die Rater-Interaktionen erklärten einen moderaten Anteil (4 bis 8%) an der Gesamtvarianz. Weiterhin wurde in einer Studie von Briesch et al. (2010) das lernförderliche Verhalten von zwölf Kindergarten-Kindern durch zwei Erzieherinnen und zwei geschulte Beobachter mit DBR-SIS und systematisch-direkten Verhaltensbeobachtungen erfasst. Die Befunde wiesen auch hier auf einen bedeutsamen Anteil an Ratervarianz für die Erzieherinnen (7.5%) und deren Interaktionen hin, während keine Ratervarianz auf die Ergebnisse der geschulten Beobachter zurückgeführt werden konnte. In einer Studie (Volpe & Briesch, 2012) wurden SIS und MIS direkt miteinander verglichen. Hier wurde bei den SIS 0% an Varianz und bei den MIS 3 bis 5% an Varianz durch die Rater aufgeklärt.

Einfluss der Messzeitpunkte

Der Anteil der Varianz, der durch die Messzeitpunkte erklärbar ist, basiert auf den Werten, die an den verschiedenen Messzeitpunkten erhoben wurden. (z. B. Schultage). Das bedeutet, dass hohe Varianzaufklärungen durch diese Facette auf stark unterschiedliche Verhaltensbeurteilungen zu den verschiedenen Messungen hinweisen. Für die praktische Umsetzung von Verhaltensverlaufsdiagnostik wäre bei standardisierten Messzeitpunkten (z. B. wenn das Verhalten immer in Stillarbeitsphasen erfasst wird) eine möglichst geringe Varianzaufklärung wünschenswert, da dies eine höhere Stabilität der Daten und damit eine bessere Interpretierbarkeit impliziert (Dever, Dowdy, Raines & Carnazzo, 2015).

In den bisherigen Studien zur Reliabilität von DBR konnte zwischen 0 und 20% an Gesamtvarianz durch die Messzeitpunk-

te erklärt werden. Insgesamt lässt sich auch in Bezug auf den Einfluss der Messzeitpunkte festhalten, dass die Stärke der Varianzaufklärung davon abhängt, welche Personengruppe die DBR umgesetzt hat. So konnte in der Studie von Chafouleas et al. (2010) gezeigt werden, dass die Tage, an denen die Beurteilungen erfolgten, bei Vorschul- oder Kindergartenkindern keine bedeutsame Varianz aufklärten, bei Schülerinnen und Schüler in der Mittelstufe hingegen schon. Hier konnte zwischen 16 und 20% der Varianz durch die Tage, an denen die DBR umgesetzt wurden, erklärt werden. Der direkte Vergleich zwischen SIS (1 bis 3%) und MIS (0 bis 1%) ergab keine nennenswerten Unterschiede (Volpe & Briesch, 2012).

Anzahl an notwendigen Messungen für reliable Befunde

Die Ergebnisse aus bisherigen Studie weisen darauf hin, dass ein Generalisierbarkeitskoeffizient von $\varepsilon p^2 \geq .70$ nach einem bis sechs Messzeitpunkten erzielt werden kann (Briesch et al., 2010; Chafouleas et al., 2007). Vergleichbare Werte konnten für den Zuverlässigkeitskoeffizienten Φ nach zwei bis zwölf Messungen erzielt werden (ebd.). Grundsätzlich wurden für die Beurteilung störenden Verhaltens (z. B. Benutzen von Schimpfwörtern, In Streitereien verwickelt sein) mehr Messungen benötigt als für die Beurteilung lernförderlichen Verhaltens (Volpe & Briesch, 2012). Dies wird unter anderem mit der schwieriger objektivierbaren Interpretation dieser Verhaltensweisen sowie einer generell geringen Auftretenshäufigkeit in schulischen Settings erklärt (z. B. Chafouleas, 2011; Volpe & Briesch, 2012). In Bezug auf die Schultage, die für reliable Ergebnisse notwendig sind, waren bei der Beurteilung durch Vorschullehrkräfte sieben Tage notwendig um den Wert von .70 zu übersteigen (Chafouleas et al., 2007). Bei Beurteilungen durch Lehrkräfte der Mittelstufe mittels SIS konnte der Wert auch nach 20 Tagen nicht erreicht werden (Chafouleas et al., 2010). Dieser Befund wurde in der Studie von Briesch et al. (2010) ebenfalls erzielt. Interessant ist das Resultat aus einer Studie von Chafouleas et al. (2010), in der gezeigt wurde, dass es der Klassenlehrkraft bereits nach zehn Schultagen gelingen kann, reliable Ergebnisse zu erzielen, wohingegen eine andere Lehrkraft aus dem Kollegium, die nicht täglich in der Klasse unterrichtete, die doppelte Anzahl an Tagen benötigte. Insgesamt lässt sich demnach festhalten, dass Lehrkräfte zwischen sieben und 20 Tagen benötigen, um reliable Ergebnisse aus der Verhaltensverlaufsdiagnostik zu generieren.

Wenngleich somit einige Studien zur Reliabilität von DBR vorliegen, ist der bisherige Erkenntnisstand zum einen inkonsistent und zum anderen noch lückenhaft: So resultiert der Großteil der Befunde aus Studien, in denen das Schülerverhalten mittels Videosequenzen beurteilt wurde und häufig geschah dies durch Studierende. Die oben dargestellten Studien stellen diesbezüglich einige wenige Ausnahmen dar. Dementsprechend gibt es nur unzureichend Studien, die den Einfluss dieser Faktoren auf Lehrerratings im realen insbesondere inklusiven Setting untersucht haben. Dieses Setting soll allerdings insbesondere von der Umsetzung der Verlaufsdiagnostik profitieren (Huber & Grosche, 2012). Zweitens liegen bislang keine Studien vor, in denen explizit das Verhalten von Schülerinnen und Schülern mit Problemverhalten – also einer der zentralen Zielgruppen der Verlaufsdiagnostik im RtI-Modell – fokussiert wurde.

Fragestellungen

Es ist allerdings anzunehmen, dass sowohl der Einbezug von Lehrkräften (z. B. aufgrund unterschiedlicher Expertise im Vergleich zu Studierenden; Bromme, 2014) im realen Setting (z. B. aufgrund anderer auditiver und visueller Anforderungen in der Beobachtungssituation im Vergleich zu Videostudien; Casabianca et al., 2013) als auch die Fokussierung auf Schülerinnen und Schüler mit Verhaltensproblemen (z. B. auf-

grund einer höheren Sprunghaftigkeit im Verhalten im Vergleich zu „unauffälligen" Schülerinnen und Schüler; Hayling, Cook, Gresham, State & Kern, 2008) zu anderen Ergebnissen in Bezug auf die o. g. Einflussfaktoren führen könnte. Für die Anwendung von DBR zur Verlaufsdiagnostik in der Praxis ist es allerdings wichtig, die Bedingungen zu kennen, unter denen sich das Verhalten möglichst zuverlässig erfassen lässt. Aus diesen Gründen untersucht die vorliegende Studie zwei zentrale Fragestellungen:

1. Welchen Einfluss haben die Lehrkräfte und die Messzeitpunkte auf die Zuverlässigkeit von DBR bei Schülerinnen und Schüler mit externalisierenden Verhaltensproblemen im inklusiven Setting?
2. Wie viele Messzeitpunkte sind notwendig, um mittels DBR zuverlässig interpretierbare Befunde ($\Phi \geq .70$) zu erzielen?

Methode

Stichprobe

Die vierte Klasse einer inklusiven Grundschule im nahen Umland einer Großstadt in Nordrhein-Westfalen nahm an der Studie teil. Die Auswahl der Grundschule erfolgte nach freiwilliger Meldung auf einen offiziellen Aufruf zur Teilnahme an der Studie. Die Klasse wurde von einer Regelschullehrerin (45 Jahre) und einer Lehrerin für sonderpädagogische Förderung (30 Jahre) im Co-Teaching unterrichtet. Die Regelschullehrkraft verfügte zum Zeitpunkt der Datenerhebung über 13 Jahre Berufserfahrung, die Lehrkraft für sonderpädagogische Förderung wies fünfeinhalb Jahre an Berufserfahrung auf. Die Zusammenarbeit im Co-Team erfolgte seit dreieinhalb Jahren.

In die vorliegende Studie einbezogen wurden die Schülerinnen und Schüler mit externalisierenden Verhaltensproblemen, d. h. mit ausagierenden, nach „außen" gerichteten Verhaltensweisen, wie z. B. Unaufmerksamkeit, Aggressivität und Impulsivität (Linderkamp & Grünke, 2007; Myschker & Stein, 2014). Eine Fokussierung wurde deshalb vorgenommen, da Verhaltensstörungen ein sehr heterogenes Phänomen darstellen, das in sehr vielen Ausprägungen auftreten kann. Häufig weisen Kinder und Jugendliche allerdings schwerpunktmäßig Probleme in einem spezifischen Verhaltensbereich auf, wenngleich es hohe Komorbiditätsraten gibt (Conroy, Stichter, Daunic & Haydon, 2008). Dementsprechend setzen pädagogische Handlungsmöglichkeiten auch an diesen konkreten Problembereichen an (Lewis, 2016). Kazdin (2005) empfiehlt daher bei der Testentwicklung im Bereich der schulischen Diagnostik von Verhaltensproblemen eine Fokussierung auf einen spezifischen Bereich.

Die Identifikation der Schülerinnen und Schüler mit externalisierendem Problemverhalten erfolgte über die Anwendung eines universellen Verhaltensscreenings, das von den beiden Lehrkräften bearbeitet wurde (s. u.). Insgesamt wurden fünf Jungen mit grenzwertigem oder auffälligem externalisierenden Problemverhalten identifiziert (Tabelle 1).

Erhebungsinstrumente

Strengths and Difficulties Questionnaire (SDQ) – Lehrkraftversion

Zur Identifikation der Schülerinnen und Schüler mit externalisierendem Problemverhalten wurde die deutsche Version des Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) eingesetzt. Insgesamt setzt sich der SDQ aus fünf Skalen (Hyperaktivität, Verhaltensprobleme, Verhaltensprobleme mit Gleichaltrigen, Emotionale Probleme, Prosoziales Verhalten) zusammen, die zur Identifikation von Schülerinnen und Schülern mit Verhaltensproblemen genutzt werden können. Jede Skala besteht aus fünf Items, die von der Lehrkraft auf einer dreistufigen Likert-Skala (0 = nicht zutreffend, 1 = teilweise zutreffend, 2 ein-

Tabelle 1: Schüler mit externalisierendem Problemverhalten in der Klasse (gemessen mit dem SDQ)

ID	Geschlecht	Alter	MH	SFB	SDQ EXT	
					RL	LSF
1	m	10	ja	L	9	13
2	m	11	ja	-	12	10
3	m	9	nein	-	12	7
4	m	10	ja	-	11	7
5	m	10	ja	L	9	11

Anmerkungen: ID = Identifikationsnummer der Schüler; MH = Migrationshintergrund; SFB = sonderpädagogischer Förderbedarf; SDQ EXT = Summenwert der Skala „externalisierendes Problemverhalten" des SDQ; RL = Regelschullehrkraft; LSF = Lehrkraft für sonderpädagogische Förderung; m = männlich; L = Lernen.

deutig zutreffend) eingeschätzt werden. Die vier Problemskalen beziehen sich zum einen auf externalisierendes (Hyperaktivität, Verhaltensprobleme) und zum anderen auf internalisierendes Problemverhalten (Verhaltensprobleme mit Gleichaltrigen, Emotionale Probleme). Die bisherigen Evaluationen der deutschsprachigen Lehrkraftversion weisen auf eine angemessene interne Konsistenz sowie eine akzeptable Passung des 5-Faktorenmodells hin (Bettge, Ravens-Sieberer, Wietzker & Hölling, 2002; Saile, 2007). Untersuchungen hinsichtlich der prädiktiven Validität zeigen, dass sich der SDQ gut dazu eignet, jene Schülerinnen und Schüler als problematisch im Verhalten zu klassifizieren, die auch mit einem eher klinischen Maß, wie der Child Behavior Checklist (CBCL; Döpfner, Plück & Kinnen, 2015) als problematisch im Verhalten identifiziert werden (Becker, Woerner, Hasselhorn, Banaschewski & Rothenberger, 2004; Bettge et al., 2002). Damit eignet sich das Instrument zur Identifikation von Schülerinnen und Schüler mit Problemen im Verhalten.

Im Rahmen der vorliegenden Studie wurden die Skalen zum externalisierenden Problemverhalten (Verhaltensprobleme, Hyperaktivität) genutzt. Dieses Vorgehen eignet sich für ein erstes Screening zur Identifikation von Risikofällen in universellen Stichproben besser als eine gezielte Nutzung einzelner Skalen, da das interessierende Konstrukt durch die größere Bandbreite an Verhaltensweisen besser abgedeckt wird (Goodman & Goodman, 2009). Als problematisch im Verhalten werden demzufolge die Schülerinnen und Schüler klassifiziert, die von der beurteilenden Lehrkraft mit einem Cut-Off-Wert von ≥ 9 (grenzwertiges Problemverhalten) beurteilt wurden. In die vorliegende Studie wurden die Schülerinnen und Schüler einbezogen, deren Verhalten von mindestens einer der beiden Lehrerinnen über dem Cut-Off eingeschätzt wurde.

Integrated Teacher Report Form – German language version (ITRF-G)

Zur Systematisierung des unterrichtlichen Problemverhaltens der Schülerinnen und Schüler mit externalisierendem Problemverhalten wurde die deutsche Version der Integrated Teacher Report Form (ITRF; Volpe & Fabiano, 2013) eingesetzt. Bei der ITRF handelt es sich um ein universelles Verhaltensscreening, das im Lehrkrafturteil die problematische Ausprägung von konkreten Verhaltensweisen im Unterricht erfasst. Die deutsche Version (Volpe et al., angenommen) umfasst 47 Items, die sich in den Faktoren „Probleme im lernförderlichen Verhalten" (z. B. *Beginnt mit der Aufgabenbearbeitung nicht selbstständig*) und „oppositionelles/störendes Verhalten" (z. B. *Hat Konflikte mit Mitschülerinnen und Mitschü-*

lern) abbilden lassen. Die Einschätzung erfolgt auf einer vierstufigen Likert-Skala (0 = Verhalten ist nicht problematisch, 1 = Verhalten ist leicht problematisch, 2 = Verhalten ist mäßig problematisch, 3 = Verhalten ist stark problematisch). Die Auswertung erfolgt über die Bildung von Summenscores, so dass man a) im normorientierten Vergleich die Schülerinnen und Schüler mit dem problematischsten Verhalten identifizieren kann und b) im intraindividuellen Vergleich die konkreten problematischen Verhaltensweisen für eine Schülerin/einen Schüler identifizieren kann.

Die psychometrischen Eigenschaften wurden sowohl für die US-amerikanische Originalversion als auch für eine Kurzvariante (16 Items) der deutschsprachigen Version positiv evaluiert. So weisen beide Sprachversionen hohe interne Konsistenzen für alle Skalen, eine akzeptable Test-Retest-Reliabilität sowie eine hohe externe Validität (konvergente und divergente Validität, Klassifikationsgenauigkeit) im Vergleich zu anderen Screeningverfahren auf (Daniels, Volpe, Briesch & Fabiano, 2014; Daniels, Volpe, Fabiano & Briesch, 2017; Volpe et al., angenommen; Volpe et al., unter Begutachtung).

Direct Behavior Rating (DBR)

Zur Erfassung der Verhaltensverläufe im Unterricht wurden in der vorliegenden Studie die DBR für externalisierendes Problemverhalten angewendet. Es wurde sowohl eine SIS als auch eine MIS genutzt. Externalisierendes Problemverhalten wurde relativ weit definiert als all jene beobachtbaren Verhaltensweisen im Unterricht, die andere Schülerinnen und Schüler stören oder das eigene Lernen sowie das Lernen anderer beeinträchtigen. Die Definition wurde jeweils auf dem Beurteilungsbogen der SIS für die Lehrkraft aufgelistet. Die Beurteilung der Auftretenshäufigkeit dieser Verhaltensweisen in einer spezifischen Unterrichtssituation erfolgte auf einer sechsstufigen Likertskala mit jeweils quantitativen (0 bis 5) und qualitativen (0 = nie, 1 = selten, 2 = manchmal, 3 = oft, 4 = sehr oft, 5 = immer) Ankern.

Zur Identifikation relevanter Items für die MIS zum externalisierenden Problemverhalten im Unterricht wurde ein schrittweises Vorgehen in Anlehnung an Hyman et al. (1998) gewählt. Zuerst wurden mittels ITRF die fünf problematischsten Items für jeden Verhaltensbereich (störendes/oppositionelles Verhalten, Probleme im lernförderlichen Verhalten) identifiziert. Im zweiten Schritt wurden Interviews mit beiden Lehrkräften geführt, um zu überprüfen, ob diese Items für die jeweiligen Schüler im Unterricht problematisch und grundsätzlich beobachtbar sind. Auf dieser Grundlage wurden drei Items mit Bezug zum lernförderlichen Verhalten und zwei Items mit Bezug zum störenden/oppositionellen Verhalten beibehalten (Tabelle 2). Auch für die Beurteilung durch die MIS wurde die genannte sechsstufige Likert-Skala genutzt.

Tabelle 2: Auswahl der Items für die MIS

Störendes/Oppositionelles Verhalten
Stört andere
Lässt sich vom negativen Verhalten anderer ablenken
Hat Konflikte mit Mitschülerinnen und Mitschülern
Bevormundet andere
Streitet mit Mitschülerinnen und Mitschülern
Probleme im lernförderlichen Verhalten
Bearbeitet Unterrichtsaufgaben ungenau
Bearbeitet Unterrichtsaufgaben unvollständig
Bearbeitet Unterrichtsaufgaben nicht in der dafür vorgesehenen Zeit
Korrigiert seine eigenen Aufgaben nicht
Beginnt mit der Aufgabenbearbeitung nicht selbstständig

Vorgehensweise

Vor dem Beginn der Datenerhebung wurde mit beiden Lehrkräften ein Training zur Durchführung von DBR auf Basis der Empfehlung von Chafouleas (2011) durchgeführt. Die Datenerhebung erfolgte innerhalb einer Schulwoche an vier konsekutiven Schultagen. Beide DBR-Formen (MIS, SIS) wurden zweimal täglich in der ersten und dritten Unterrichtsstunde in individuellen Stillarbeitsphasen von ca. zehn Minuten für die fünf Schüler ausgefüllt. Diese Unterrichtsphase wurde ausgewählt, da a) alle fünf Schüler zu diesen Phasen in der Klasse anwesend waren und keine externen Förderangebote in Anspruch nahmen und b) beide Lehrerinnen individuelle Stillarbeitsphasen als am problematischsten in Bezug auf das Schülerverhalten einschätzten. Beide Lehrkräfte bearbeiteten die DBR unmittelbar im Anschluss an die Situation unabhängig voneinander ohne sich abzusprechen.

Datenanalyse

Der Studie liegt ein vollständig gekreuztes 2-Facetten-Design zugrunde, was bedeutet, dass zwei Faktoren, die als die DBR-Messung beeinflussend angenommen werden, modelliert werden. Bei der ersten Facette handelt es sich um die Regelschullehrkraft sowie die Lehrkraft für sonderpädagogische Förderung (Facette Rater). Da beide Lehrkräfte dem Verständnis der GT zufolge eine repräsentative Auswahl des Universum aller zulässigen Lehrkräfte der Regel- oder Förderschule darstellen, handelt es sich hier in der GT-Terminologie um eine zufällige Facette, d. h. bei den Varianzanalysen in der G-Studie wird diese Facette als zufälliger Faktor modelliert (Briesch, Swaminathan, Welsh & Chafouleas, 2014). Bei der zweiten Facette handelt es sich um die insgesamt acht Messungen in den individuellen Stillarbeitsphasen über vier Schultage (Facette Messzeitpunkt). Da auch diese Messungen lediglich eine Auswahl aller potenziell möglichen Messungen aus dem Universum aller zulässigen Bedingungen darstellen, wird auch diese Facette als zufällig angenommen. Bei den fünf Schülern handelt es sich um die Differenzierungsfacette, also die Bedingungsgruppe, über die die Aussagen generalisiert werden sollen. Damit wurden in der vorliegenden Studie insgesamt 80 Datenpunkte generiert (5 Schüler x 2 Rater x 8 Messungen).

Im ersten Schritt wurden die Varianzanteile jeder einzelnen Facette sowie deren Interaktionen geschätzt. Dazu wurde die Varianz des beobachteten Werts in seine Bestandteile zerlegt:

$$\sigma^2(Y) = \sigma^2(s) + \sigma^2(r) + \sigma^2(m) + \sigma^2(sr) + \sigma^2(sm) + \sigma^2(rm) + \sigma^2(srm, e),$$

wobei $\sigma^2(Y)$ die Varianz des beobachtbaren Wertes und $\sigma^2(s)$, $\sigma^2(r)$ und $\sigma^2(m)$ die Varianzen der Facetten Schüler, Rater und Messungen sowie deren Interaktionen darstellen. Hierfür wurden bei der SIS die Rohwerte und bei der MIS die Summenwerte über alle fünf Items genutzt. Für die Varianzanalysen wurden ANOVA Typ III Quadratsummen berechnet (Brennan, 2001). Im zweiten Schritt wurde eine Simulationsstudie durchgeführt, um herauszufinden, wie viele Messzeitpunkte ein einzelner Rater benötigt, um zuverlässige DBR-Befunde zu erzielen. Hierfür wurde die Anzahl an Messzeitpunkten von eins bis 20 variiert und das Modell für jede dieser Bedingung berechnet. Schließlich wurden für jede Analyse sowohl $\varepsilon\rho$ als auch Φ berechnet, um den Schwellenwert für zuverlässige Messungen für normorientierte ($\varepsilon\rho$) bzw. intraindividuelle (Φ) praktische Entscheidungen zu identifizieren. Der G-Koeffizient $\varepsilon\rho$ berechnet sich wie folgt:

$$\varepsilon\rho^2 = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma_\delta^2}$$

Er setzt sich also aus der Varianz der universalen Werte der Personen (σ_ρ^2) in Beziehung zur Summe dieser Varianz und der relativen

Fehlervarianz (σ^2_δ), also der Varianz aus den gemessenen Werten von mehreren Personen, zusammen. Der Zuverlässigkeitsindex wird wie folgt berechnet:

$$\Phi = \frac{\sigma^2_\rho}{\sigma^2_\rho + \sigma^2_\Delta}$$

Hier wird also die Varianz der universalen Werte der Personen ins Verhältnis zur Summe dieser Werte und der absoluten Fehlervarianz der gemessenen Werte einer Person zu den acht Messzeitpunkten gesetzt. Für beide Koeffizienten wurde ein Kriteriumswert von ≥ .70 als akzeptabel für Daten der Verlaufsdiagnostik angelegt (Salvia, Ysseldyke & Witmer, 2012).

Ergebnisse

Deskriptive Befunde

Tabelle 3 zeigt die deskriptiven Befunde über alle acht Messzeitunkte sowohl für jeden Schüler als auch für alle fünf Schüler aus den DBR beider Lehrkräfte. Schüler 2 zeigt insgesamt im Urteil beider Lehrkräfte am häufigsten externalisierendes Problemverhalten. Schüler 1 und Schüler 4 zeigen insgesamt im Urteil beider Lehrkräfte das geringste externalisierende Problemverhalten. Im SIS-Urteil der Regelschullehrkraft tritt das Problemverhalten insgesamt häufiger auf als im SIS-Urteil der Lehrkraft für sonderpädagogische Förderung. Bei den Ratings mit der MIS ist dieser Trend genau gegenläufig. Die *lag*-1-Autokorrelationen der Messzeitpunkte weisen größtenteils ein ver-

Tabelle 3: Deskriptive Ergebnisse über alle acht Messzeitunkte für jeden Schüler sowie insgesamt aus den DBR beider Lehrkräfte

	Regelschullehrkraft									
	DBR-SIS					**DBR-MIS[a]**				
Schüler	**Min**	**Max**	**M**	**SD**	**AC**	**Min**	**Max**	**M**	**SD**	**AC**
1	0	1	0.50	0.53	0	0	10	3.45	1.77	0.11
2	0	2	1.33	1.16	-0.13	10	15	12.50	2.50	-0.04
3	0	2	1.25	1.04	-0.16	0	15	9.38	5.48	-0.20
4	0	1	0.17	0.41	-0.63	0	15	6.25	6.07	0.54
5	0	2	1.50	0.76	-0.19	8	15	12.20	2.48	-0.02
Gesamt	**0**	**2**	**0.94**	**0.90**		**0**	**15**	**8.33**	**5.55**	
	Lehrkraft für sonderpädagogische Förderung									
	DBR-SIS					**DBR-MIS[a]**				
Schüler	**Min**	**Max**	**M**	**SD**	**AC**	**Min**	**Max**	**M**	**SD**	**AC**
1	0	3	1.50	1.60	-0.28	0	8	3.45	3.00	-0.07
2	0	3	2.00	1.73	-0.05	0	15	10.00	8.65	-0.07
3	0	4	1.88	1.36	-0.38	0	18	8.54	6.40	-0.50
4	0	2	0.80	1.10	0.55	0	18	5.50	2.95	0.04
5	2	4	3.00	0.76	0.50	0	13	7.20	4.10	0.48
Gesamt	**0**	**4**	**1.91**	**1.42**		**0**	**18**	**6.94**	**5.02**	

Anmerkungen: Min = minimaler Wert; Max = maximaler Wert; M = Mittelwert; SD = Standardabweichung; AC = lag-1-Autokorrelation der acht Messzeitpunkte. [a]Die DBR-MIS bestand aus fünf Items und kann Werte zwischen 0 (mindestens) und 25 (maximal) annehmen.

tretbares Ausmaß auf. Bei der Beurteilung durch die Regelschullehrkraft liegen für Schüler 4 (SIS: -0.63, MIS: 0.54) hohe Autokorrelationen vor. Bei der Beurteilung durch die Sonderschullehrkraft sind für Schüler 3 (MIS: -0.50), Schüler 4 (SIS: 0.55) und Schüler 5 (SIS: 0.50) ebenfalls hohe Werte feststellen.

Varianzkomponentenschätzung (G-Studie)

Die Ergebnisse der G-Studie (Tabelle 4) zeigen, dass sowohl bei den SIS- als auch bei den MIS-Ratings die Schüler den größten Anteil der Gesamtvarianz aufklären. Das bedeutet, dass die fünf Schüler externalisierendes Problemverhalten in unterschiedlicher Ausprägung zeigen. Dieser Befund wird auch durch die deskriptiven Befunde in Tabelle 3 gestützt. In Bezug auf die Facetten Rater und Messzeitpunkt lässt sich konstatieren, dass sowohl die Rater (18.1%) als auch die Messzeitpunkte (17.9%) einen substantiellen Anteil der Gesamtvarianz der SIS-Ratings aufklären. Bei den MIS-Ratings bleibt dieser Anteil sowohl bei den Ratern (2.5%) als auch bei den Messzeitpunkten (5.2%) im kleinen bis moderaten Bereich. In Bezug auf die Rater bedeutet dies, dass die Verhaltensbeurteilungen der beiden Lehrkräfte mit den SIS sehr unterschiedlich sind und sich die MIS diesbezüglich nur geringfügig unterscheiden. In Bezug auf die Messzeitpunkte bedeutet dies, dass die Verhaltensbeurteilungen über alle fünf Schülerinnen und Schüler mit den SIS zu den acht Arbeitsphasen sehr unterschiedlich sind und bei den MIS-Urteilen moderate Unterschiede bestehen. Die Auftretenshäufigkeit des Verhaltens wird mit den MIS also a) von beiden Lehrkräften ähnlich und b) als weniger sprunghaft eingeschätzt.

In Bezug auf die Interaktionseffekte der personen- und zeitbezogenen Faktoren lassen sich sowohl mit den SIS (2.3%) als auch mit den MIS (4.6%) nur geringe Schüler-Rater-Interaktionen feststellen. Dieser Befund impliziert, dass beide Lehrkräfte die Schüler im relativen Vergleich ähnlich einschätzen. Die Interaktion zwischen den Schülern und den Messzeitpunkten (S x M) klärt hingegen sowohl bei den SIS (19.8%) als auch bei den MIS (22.1%) einen substanziellen Anteil der Gesamtvarianz auf. Dies bedeutet, dass die Auftretenshäufigkeit des externalisierenden Problemverhaltens der gleichen Schüler an den verschiedenen Messzeitpunkten unterschiedlich beurteilt wurde. Bezüglich der Rater-Messzeitpunkt-Interaktion (R x M) sind die Befunde für SIS und MIS uneinheitlich. Während sich für die SIS ein kleiner Interaktionseffekt feststellen lässt (2.3%), stellt diese Interaktion bei den MIS die zweitgrößte Varianzquelle dar (26.8%). Dieses Ergebnis bedeutet, dass beide Rater zu den acht Messungen die Verhaltensausprägung mit den SIS über alle fünf Schüler sehr ähnlich beurteilen, wohingegen die differenzierteren Beurteilungen mit den MIS sich zwischen beiden Lehrkräften zu den acht Messungen stark unterscheiden.

Tabelle 4: Ergebnisse der G-Studie: Varianzkomponentenschätzungen (%)

Facette	**Externalisierendes Problemverhalten**	
	DBR-SIS	**DBR-MIS**
	VC (%)	**VC (%)**
Schüler (S)	25.3	28.5
Rater (R)	18.1	2.5
Messzeitpunkt (M)	17.9	5.2
S x R	2.3	4.6
S x M	19.8	22.1
R x M	2.3	26.8
S x R x M + res	13.9	10.3
$\varepsilon\rho$	**.76**	**.76**
Φ	**.54**	**.66**

Anmerkung: DBR = Direct Behavior Rating; MIS = Multiple Item Skala; SIS = Single Item Skala; VC (%) = Varianzkomponentenschätzung in Prozent; res = Residualvarianz; $\varepsilon\rho$ = Generalisierbarkeitskoeffizient; Φ = Zuverlässigkeitskoeffizient.

Bedingungssimulation (D-Studie)

Aufbauend auf den Befunden der Varianzkomponentenschätzung wurden im Rahmen der Simulationsstudie die Bedingungen innerhalb der Facette Messzeitpunkte so variiert, dass die Modellschätzung für jeweils einen bis 20 Messzeitpunkte für einen Rater erfolgte. Als kritischer Minimalwert für pädagogisch-praktische Entscheidungen wurde eine Varianzaufklärung von mindestens 70% festgelegt. Die Ergebnisse (Abbildung 1) zeigen, dass sowohl die SIS (vier MZP) als auch die MIS (fünf MZP) nach relativer kurzer Zeit den kritischen Wert für relative Entscheidungen (εp) überschreiten. Dies bedeutet, dass nach vier bzw. fünf Messungen mit DBR ein zuverlässiger Vergleich zwischen mehreren Schülern mit Verhaltensproblemen möglich ist. In Bezug auf intraindividuelle Entscheidungen (Φ) zeigt sich, dass das externalisierende Problemverhalten mit den MIS nach 13 Messzeitpunkten erfasst werden kann. Die Simulation mit SIS bleibt selbst nach 20 Messungen unter dem kritischen Wert. Dies bedeutet, dass sich nach 13 Messungen mit DBR-MIS zuverlässige Aussagen über die intraindividuelle Verhaltensentwicklung eines Schülers machen lassen. Für die Beurteilung mit SIS konnte der kritische Wert in dieser Studie nicht überschritten werden.

Diskussion

Inhaltliche Diskussion

In der vorliegenden Studie wurde gezeigt, dass die Rater bei der Verhaltensverlaufsdiagnostik mit DBR nur dann einen substantiellen Einfluss auf die Zuverlässigkeit haben, wenn die interessierenden Verhaltensweisen global mit einer SIS erfasst werden. Bei den spezifischeren und konkreteren MIS lässt sich kein substantieller Einfluss feststellen, wenn die Rater isoliert betrachtet werden. Dies könnte damit zu erklären sein, dass die Zuverlässigkeit von Verhaltensbeurteilung mit der Ausprägung der Inferenz der Items, die zur Beurteilung genutzt werden, zusammenhängt. Generell lässt sich sagen, dass die Beurteilung mit hoch-inferenten Items weniger zuverlässig erfolgt als

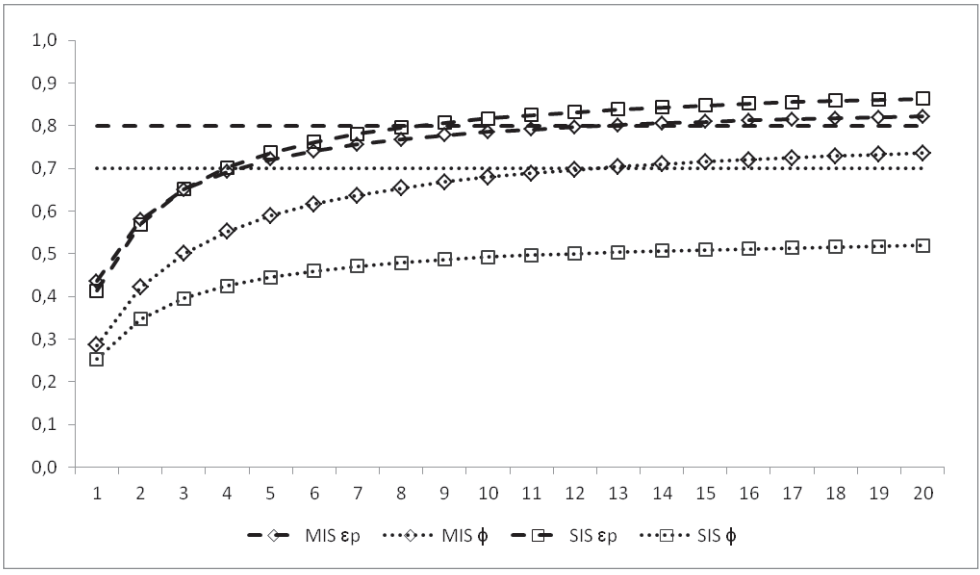

Abbildung 1: Ergebnisse der D-Studie: Entwicklung der Reliabilitätskoeffizienten über eine variierende Anzahl an Messungen

mit niedrig- oder mittel-inferenten Items (Cone, 1977). Für die Verhaltensverlaufsdiagnostik wurde ein angemessener Grad an Inferenz der Items in zwei Arbeiten als Gütekriterium für die entsprechenden Verfahren angelegt (Casale et al., 2015[b]; Christ et al., 2010). Auch konnte in einer US-amerikanischen Studie nachgewiesen werden, dass Lehrkräfte die tatsächliche Auftretenshäufigkeit von Verhaltensproblemen ziemlich zuverlässig einschätzen, wenn sie genau wissen, wie sich die Verhaltensweisen äußern (Conley, Marchant & Caldarella, 2014). Die vorliegende Studie könnte Hinweise darauf liefern, dass die Konkretisierung der Verhaltensweisen in den Items einen entscheidenden Einfluss auf die Zuverlässigkeit der DBR-Befunde hat. Da es allerdings Studien gibt, die darauf hinweisen, dass die Formulierung der Items keinen substanziellen Einfluss auf die Zuverlässigkeit von DBR-SIS hat (Riley-Tillman, Chafouleas, Christ, Briesch & LeBel, 2009), sollten zukünftige Studien diesen Aspekt gezielt in den Blick nehmen.

Ein weiterer Befund der vorliegenden Studie weist auf einen starken Einfluss der einzelnen Messzeitpunkte hin. Dies bedeutet im Umkehrschluss, dass die Zuverlässigkeit der Ergebnisse aus DBR davon abhängen, in welcher Situation die Ratings durchgeführt werden. Hier schlagen sowohl bei den SIS (19.8%) als auch bei den MIS (22.1%) die Interaktionseffekte zwischen Messzeitpunkten und Schülern zu Buche, was die inhaltliche Aussage impliziert, dass die gleichen Schüler zu den verschiedenen Messzeitpunkten unterschiedlich beurteilt wurden. Dies könnte mit der hohen Variabilität und Sprunghaftigkeit des konkreten Verhaltens im Unterricht von Schülerinnen und Schüler mit externalisierendem Problemverhalten zu erklären sein (Hayling, Cook, Gresham, State & Kern, 2008). Ob und in welchem Ausmaß ein bestimmtes Verhalten in einer bestimmten Situation gezeigt wird, hängt von vielen Faktoren, wie z. B. der persönlichen Motivation, dem Interesse am Lerngegenstand oder umweltbezogenen Aspekten, die zu einer sprunghaften Verhaltensveränderungen führen, ab (Fox & Conroy, 1995). Im letzten Fall spricht man auch von sog. „setting events" (Fox & Conroy, 1995; S. 130). Insbesondere Schülerinnen und Schüler mit externalisierendem Problemverhalten erleben häufig aversive psychologische und physiologische Umwelten, so dass die Wahrscheinlichkeit des Einflusses der hier erlebten Reize auf das gezeigte Verhalten besonders hoch ist (Wettstein, Bryjová, Faßnacht & Jakob, 2011).

Da die Verhaltensverlaufsdiagnostik insbesondere im Rahmen von mehrstufigen Fördersystemen relativ zeitnah die Datengrundlage für die Entscheidung über den Erfolg oder Misserfolg von Verhaltensfördermaßnahmen schafft, ist es wichtig zu wissen, nach wie vielen Messungen die DBR-Ergebnisse zuverlässig interpretierbar sind (Fan & Hansmann, 2015). Dieser Frage wurde in der vorliegenden Untersuchung durch die Durchführung einer Simulationsstudie nachgegangen. Berechnet wurde die Zuverlässigkeit für normorientierte und intraindividuelle Entscheidungen. In Bezug auf die normorientiere Einordnung des Verhaltens eines Schülers liefern sowohl die SIS (nach vier Messungen) als auch die MIS (nach fünf Messungen) sehr zeitnah Ergebnisse, die zuverlässig interpretierbar sind. Für die intraindividuelle Einordnung eines Verhaltenswertes konnte in der vorliegenden Studie nur mittels MIS in einem ökonomischen Zeitraum (13 Messungen) der kritische Wert erreicht werden. Bei täglichen DBR-Messungen würde dies bedeuten, dass nach ungefähr drei Wochen eine zuverlässige Interpretation der intraindividuellen Verhaltensverläufe einer Schülerin/ eines Schülers erfolgen kann. Auch wenn ein kürzeres Zeitintervall wünschenswert wäre, würde die Methode Verhaltensverläufe immer noch in einem Zeitraum abbilden, in dem eher traditionelle diagnostische Verfahren, wie z. B. Beobachtungen, Verhaltensveränderungen nicht sensitiv abbilden können (Casale et al., 2015[b]).

Methodische Diskussion

Der methodische Zugang erfolgte in der vorliegenden Studie über den Ansatz der Generalisierbarkeitstheorie. Durch die Möglichkeiten, (a) mehrere Messfehlerquellen und deren Interaktionen simultan zu schätzen, (b) die Bedingungen innerhalb dieser Quellen zu simulieren und (c) Reliabilitätskoeffizienten sowohl für relative als auch absolute Vergleiche zu berechnen, eignet sich dieser Ansatz gut für die Entwicklung verlaufsdiagnostischer Testverfahren (Hintze, Owen, Shapiro & Daly, 2000). Allerdings sollte die methodische Herangehensweise insbesondere in Bezug auf die einbezogenen Facetten, die Autokorrelation bei zeitbezogenen Facetten, die Stichprobengröße sowie die Generalisierbarkeit der Befunde diskutiert werden.

Im Rahmen der GT werden verschiedene Faktoren, die einen Einfluss auf die Zuverlässigkeit haben, als Facetten modelliert und gezielt hinsichtlich ihres Einflusses analysiert. Dies erfordert eine theoretisch und empirisch fundierte Begründung für die Fokussierung auf genau diese Facetten (Brennan, 2001). Dies bedeutet gleichzeitig, dass andere Facetten, die durchaus auch einen Einfluss auf die Zuverlässigkeit der Messungen haben können, bewusst vernachlässigt werden (sog. „versteckte Facetten"; Briesch et al., 2014). In Bezug auf DBR wurden bereits zahlreiche Studien durchgeführt, die neben den Ratern und den Situationen weitere Facetten (z. B. das Skalenformat, die Itemformulierung) überprüft haben (z. B. Briesch, Kilgus, Chafouleas, Riley-Tillman & Christ, 2013; Chafouleas, Jaffery, Riley-Tillman, Christ & Sen, 2013; Riley-Tillman et al., 2009). Ausgehend von dem aktuellen Forschungsstand zu DBR-SIS kann davon ausgegangen werden, dass es vor allem die personen- und zeitspezifischen Facetten sind, die die Messgenauigkeit beeinflussen (Huber & Rietz, 2015b). Aus diesem Grund fokussiert sich die vorliegende Studie auf diese Facetten. Die recht kleinen Varianzanteile der nicht-interpretierbaren Dreifachinteraktion (S x R x M + res) sowohl für SIS (13.9%) als auch für MIS (10.3%) weisen darauf hin, dass versteckte Facetten die hier erzielten Ergebnisse nicht besonders stark beeinflusst haben. Gleichwohl erscheint es zukünftig sinnvoll, vor allem auch für MIS weitere Facetten in den Blick zu nehmen, da sich der aktuelle Forschungsstand zu DBR vor allem auf SIS bezieht.

Ein Kritikpunkt, der häufig im Kontext der Erhebung von Zeitreihen diskutiert wird, betrifft den der Autokorrelation der Daten beim Einbezug zeitbezogener Facetten (Lei, Smith & Suen, 2007). Im Rahmen der GT wird mit Varianzschätzungsprozeduren gearbeitet, die grundsätzlich eine Unabhängigkeit der Datenstruktur unterstellen (Brennan, 2001). Allerdings korrelieren Daten aus Zeitreihenanalysen häufig stark untereinander, was im Rahmen der Varianzkomponentenschätzung zu einer Unterschätzung des Messfehlers und einer Überschätzung des erklärbaren Varianzanteils führen kann (Rowley, 1989). Dies bedeutet, die Reliabilität würde systematisch überschätzt. Wenngleich dieser Punkt in der jüngeren Zeit verstärkt diskutiert wird (Lei et al., 2007), gibt es derzeit noch keinen tragfähigen Vorschlag zur Lösung des Problems. Daher können wir diese Herausforderung nicht kontrollieren, möchten aber darauf hinweisen, dass die Autokorrelationen in den meisten Fällen nicht besonders hoch waren (Tabelle 3).

Ein weiterer Kritikpunkt an der Nutzung der GT zur Analyse von Verhaltensverlaufsdiagnostik ist die relativ geringe Stichprobengröße in den Studien. Diese kann zu negativen Varianzkomponentenschätzungen führen, da im Rahmen der GT Varianzanalysen durchgeführt werden und diese grundsätzlich auf zu kleine Stichproben nicht anwendbar sind (Eisend, 2007). Im Kontext der Anwendung der GT wird allerdings häufig argumentiert, dass die einzelnen Faktoren und deren Bedingungen, pragmatisch in Bezug auf die realen Bedingungen konstituiert werden sollten (Shavelson

& Webb, 1991). Daher ist die Frage nach der für G-Studien richtigen Stichprobengröße noch nicht vollends geklärt (Briesch, Swaminathan, Welsh & Chafouleas, 2014). In der vorliegenden Studie kann davon ausgegangen werden, dass die Stichprobengröße ausreichend war, da keine negativen Varianzkomponentenschätzungen erzielt wurden. Zudem ist unsere Stichprobengröße analog zu den bisherigen G-Studien über DBR.

Die Generalisierbarkeitstheorie wird auch als „Stichprobentheorie" bezeichnet (Shavelson & Webb, 1991; S. 57). Daher sind die Befunde streng genommen nur über die in der Studie gewählten Facetten und Facettenstufen generalisierbar. Im Fall der vorliegenden Studie wären dies: a) männliche Schüler mit externalisierendem Problemverhalten (Facette Schüler), b) weibliche Regellehrkräfte und Lehrkräfte für sonderpädagogische Förderung (Facette Rater) sowie c) individuelle Stillarbeitsphasen im Unterricht (Facette Messzeitpunkt). Es leuchtet intuitiv ein, dass die Reichweite der Generalisierbarkeit dieser Befunde auf diesen Bedingungskreis beschränkt ist. Diese Reichweite kann über die Durchführung mehrerer G-Studien mit unterschiedlichen Schwerpunkten gesteigert werden. Vor dem Hintergrund der bereits zahlreichen G-Studien zur Evaluation von DBR kann die vorliegende Studie einen Beitrag hierzu leisten.

Praktische Implikationen

Aus den hier berichteten Befunden lassen sich zwei zentrale Implikationen für die praktische Anwendung von Verhaltensverlaufsdiagnostik ableiten. Zum einen scheint mit DBR ein Instrument vorzuliegen, das zeitnah eine Datengrundlage für pädagogisch-praktische Entscheidungen über den Erfolg von Fördermaßnahmen schaffen kann. Damit erfüllt es eine zentrale Funktion im Rahmen mehrstufiger Fördersysteme zur Prävention von manifesten Verhaltensstörungen, nämlich die engmaschige und zeitnahe Überprüfung des individuellen Fördererfolgs von Schülerinnen und Schüler (Huber & Grosche, 2012). Zum anderen lässt sich auf Basis der hier erzielten Ergebnisse ableiten, dass für die zuverlässige Beurteilung der intraindividuellen Entwicklung einer Schülerin oder eines Schülers in einem überschaubaren Zeitraum die Nutzung der konkreten und spezifischen MIS den SIS vorzuziehen ist. Für relative Vergleiche zwischen mehreren Schülerinnen und Schüler können sowohl SIS als auch MIS zeitnah zuverlässige Ergebnisse liefern.

Literatur

Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T. & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample. *European Child and Adolescent Psychiatry*, *13*, 11–11.

Bettge, S., Ravens-Sieberer, U., Wietzker, A. & Hölling, H. (2002). Ein Methodenvergleich der Child Behavior Checklist und des Strengths and Difficulties Questionnaire. *Das Gesundheitswesen*, *64*, 119–124. https://doi.org/10.1055/s-2002-39264

Brennan, R. L. (2001). *Generalizability theory*. New York, NY, US: Springer-Verlag Publishing.

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., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C. & Christ, T. J. (2013). The Influence of Alternative Scale Formats on the Generalizability of Data Obtained from Direct Behavior Rating Single-Item Scales (DBR-SIS). *Assessment for Effective Intervention*, *38*, 127–133.

Briesch, A. M., Swaminathan, H., Welsh, M. & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design,

implementation, and interpretation. *Journal of School Psychology*, *52*, 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K. & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73,* 757–783.

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

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. & Hernandez, P. (2009). Using Direct Behavior Rating (DBR) to Screen for School Social Risk: A Preliminary Comparison of Methods in a Kindergarten Sample. *Assessment for Effective Intervention*, *34*, 214–223. https://doi.org/10.1177/1534508409333547

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

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C. & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology*, *48*, 219–246. https://doi.org/10.1016/j.jsp.2010.02.001

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

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. & Boice, C. H. (2010). Direct Behavior Rating (DBR): Generalizability and Dependability Across Raters and Observations. *Educational and Psychological Measurement*, *70*, 825–843. https://doi.org/10.1177/0013164410366695

Christ, T. J., Nelson, P. M., Van Norman, E. R., Chafouleas, S. M. & Riley-Tillman, T. C. (2014). Direct Behavior Rating: An evaluation of time-series interpretations as consequential validity. *School Psychology Quarterly*, *29*, 157–170. https://doi.org/10.1037/spq0000029

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, *8*, 411–426. https://doi.org/10.1016/S0005-7894(77)80077-4

Cone, J. D. (1998). *Psychometric considerations: Concepts, contents, and methods.* Allyn & Bacon. Abgerufen am 17.07.2014 unter http://psycnet.apa.org/psycinfo/1997-36895-002

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

Conroy, M. A., Stichter, J. P., Daunic, A. & Haydon, T. (2008). Classroom-Based Research in the Field of Emotional and Behavioral Disorders: Methodological Issues and Future Research Directions. *Journal of Special Education*, *41*, 209–222.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measures. Theory of Generalizability of Scores and Profiles*. New York: Jon Wiley & Sons.

Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. (1963). Theory of Generalizability: A Liberalization of Reliability Theory? *British Journal of Statistical Psychology*, *16*, 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Daniels, B., Volpe, R. J., Briesch, A. M. & Fabiano, G. A. (2014). Development of a problem-focused behavioral screener linked to evidence-based intervention. *School Psychology Quarterly*, *29*, 438–451. https://doi.org/10.1037/spq0000100

Daniels, B., Volpe, R. J., Fabiano, G. A., & Briesch, A. M. (2017). Classification accuracy and acceptability of the Integrated Screening and Intervention System Teacher Rating Form. *School Psychology Quarterly, 32*(2), 212-225. http://dx.doi.org/10.1037/spq0000147

Dever, B. V., Dowdy, E., Raines, T. C. & Carnazzo, K. (2015). Stability and Change of Behavioral and Emotional Screening Scores. *Psychology in the Schools*, *52*, 618–629. https://doi.org/10.1002/pits.21825

Döpfner, M., Plück, J. & Kinnen, C. (2015). *CBCL/6-18R, TRF/6-18R, YSR/11-18R Deutsche Schulalter-Formen der Child Behavior Checklist von Thomas M. Achenbach*. Göttingen: Hogrefe.

Fan, C.-H. & Hansmann, P. R. (2015). Applying Generalizability Theory for Making Quantitative RTI Progress-Monitoring Decisions. *Assessment for Effective Intervention*, *40*, 205–215. https://doi.org/10.1177/1534508415573299

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

Goodman, A. & Goodman, R. (2009). Strengths and Difficulties Questionnaire as a Dimensional Measure of Child Mental Health. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*, 400–403. https://doi.org/10.1097/CHI.0b013e3181985068

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

Grosche, M. (2014). Fördermaßnahmen im Prozess überprüfen. In T. Bohl, A. Feindt, B. Lütje-Klose, M. Trautmann & B. Wischer (Hrsg.), *Fördern* (S. 113–115). Velber: Friedrich.

Hawken, L. S., Vincent, C. G. & Schumann, J. (2008). Response to Intervention for Social Behavior: Challenges and Opportunities. *Journal of Emotional and Behavioral Disorders*, *16*, 213–225.

Hayling, C. C., Cook, C., Gresham, F. M., State, T. & Kern, L. (2008). An Analysis of the Status and Stability of the Behaviors of Students with Emotional and Behavioral Difficulties. *Journal of Behavioral Education*, *17*, 24–42.

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

Hintze, J. M., Owen, S. V., Shapiro, E. S. & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly*, *15*, 52–68. https://doi.org/10.1037/h0088778

Hoyt, W. T. & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Huber, C. & Grosche, M. (2012). Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik*, *63*, 312–322.

Huber, C. & Rietz, C. (2015a). Behavior Assessment Using Direct Behavior Rating (DBR) - A Study on the Criterion Validity

of DBR Single-Item-Scales. *Insights on Learning Disabilities*, *12*, 73–90.

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

Hyman, I. A., Wojtowicz, A., Lee, K. D., Haffner, M. E., Fiorello, C. A., Storlazzi, J. J. & Rosenfeld, J. (1998). School-Based Methylphenidate Placebo Protocols: Methodological and Practical Issues. *Journal of Learning Disabilities*, *31*, 581–594. https://doi.org/10.1177/002221949803100609

Iovannone, R., Greenbaum, P. E., Wang, W., Dunlap, G. & Kincaid, D. (2014). Interrater Agreement of the Individualized Behavior Rating Scale Tool. *Assessment for Effective Intervention*, *39*, 195–207. https://doi.org/10.1177/1534508413488414

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*, 1–73. https://doi.org/10.1111/jedm.12000

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

Lei, P.-W., Smith, M. & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools*, *44*, 433–439. https://doi.org/10.1002/pits.20235

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

Linderkamp, F. & Grünke, M. (2007). Lern- und Verhaltensstörungen: Klassifikation, Prävalenz & Prognostik. In F. Linderkamp & M. Grünke (Hrsg.), *Lern- und Verhaltensstörungen. Genese - Diagnostik - Intervention.* (S. 13–28). Weinheim: Beltz.

Miller, F. G., Cohen, D., Chafouleas, S. M., Riley-Tillman, T. C., Welsh, M. E. & Fabiano, G. A. (2015). A Comparison of Measures to Screen for Social, Emotional, and Behavioral Risk. *School Psychology Quarterly*, *30*, 184–196. https://doi.org/10.1037/spq0000085

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

Reschly, D. J. & Bergstrom, M. K. (2008). Response-to-Intervention. In T. B. Gutkin & C. R. Reynolds (Hrsg.), *The Handbook of School Psychology* (4. Aufl., S. 434–460). New York, NY, US: John Wiley & Sons Inc.

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M. & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly*, *24*, 1–12. https://doi.org/10.1037/a0015248

Rowley, G. L. (1989). Assessing Error in Behavioral Data: Problems of Sequencing. *Journal of Educational Measurement*, *26*, 273–284.

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

Salvia, J., Ysseldyke, J. & Witmer, S. (2012). *Assessment: In Special and Inclusive Education*. Cengage Learning.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA, US: Sage Publications, Inc.

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.

Volpe, R. J., Casale, G.,Mohiyeddini, C., Grosche, M., Hennemann, T., Briesch, A. M., & Daniels, B. (angenommen). A universal screener linked to personalized classroom interventions: Psychometric characteristics in a large sample of German schoolchildren. *Journal of School Psychology*.

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

Gino Casale
Universität Paderborn
Institut für Erziehungswissenschaft
Warburger Str. 100
33098 Paderborn
gino.casale@uni-paderborn.de

Erstmalig eingereicht: 03.03.2017
Überarbeitung eingereicht: 20.06.2017
Angenommen: 19.07.2017