Empirische Sonderpädagogik, 2015, Nr. 2, S. 75-98
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Direct Behavior Rating (DBR) als Methode zur Verhaltensverlaufsdiagnostik in der Schule: Ein systematisches Review von Methodenstudien

Christian Huber[1] *& Christian Rietz*[2]

[1] *Universität Potsdam*
[2] *Universität zu Köln*

Zusammenfassung

Die Messung von Verhaltensentwicklungen in der Schule ist eine große methodische Herausforderung. Der vorliegende Beitrag stellt zunächst das Direct Behavior Rating (DBR) als einen möglichen Ansatzpunkt zur Verhaltensverlaufsdiagnostik in der Schule vor. Auf dieser Grundlage werden die Ergebnisse eines systematischen Reviews zu empirischen Forschungsarbeit über diese Methodik skizziert. Die Forschungsergebnisse werden in acht Schwerpunkten eingeordnet und jeweils diskutiert. Das Review zeigt insgesamt moderate bis gute Befunde für Kriteriums-Validität und Interrater-Reliabilität von DBR-Skalen. Die Arbeiten zeigen überdies, dass die Beobachtungsgüte in einen akzeptablen Bereich steigt, wenn jeweils mindestens fünf Messdaten zu einem Wert zusammengefasst werden und das Beobachtungsziel möglichst global formuliert wird. Ferner zeigen verschiedene Studien, dass die Beobachtungsgüte je nach beobachtetem Verhalten schwankt. Die Ergebnisse des Reviews werden zusammenfassend diskutiert.

Schlagwörter: Verhaltensdiagnostik, Verhaltensbeobachtung, Verlaufsdiagnostik, Prozessdiagnostik, Direct Behavior Rating

Developmental behavior assessment in school by using Direct Behavior Rating (DBR): a review

Abstract

The developmental assessment of behavior problems in school is an important but challenging task for psychologists, teachers and other experts. The current paper (first) gives a short summary of conventional methods of behavior assessment. Second the Method of Direct Behavior Rating (DBR) as an alternative method of developmental behaviour assessment is introduced. Based on this, a review of empirical works with focus on the test quality of DBR is given. 17 studies are discussed. Results of the reviewed studies suggest a moderate to good criterion-related validity and interrater reliability. Further the studies found that test-quality could be improved, when a set of five DBR-measurements are resumed and the target behavior is defined in a more global way. Further, findings suggest that test quality is varying with the target behavior. The results of all studies are summarized and their relevance for the developmental assessment of behavior in daily school life is discussed.

Keywords: behavior assessment, developmental diagnostics, direct behavior rating

Die Umsetzung der UN-Behindertenrechtskonvention und der damit verbundene „inklusive Wandel" erfordern neue Ansätze zur Förderung und Diagnostik in der Schule. So zeichnen sich viele Länder, deren Schulsysteme im Hinblick auf den inklusiven Wandel bereits weiter fortgeschritten sind als Deutschland, nicht nur durch den Abbau von separativen Schulstrukturen aus, sondern eben auch durch den Aufbau von gestuften Fördersystemen, in denen die Individualisierung der Förderung durch Lern- oder Entwicklungsverlaufsdiagnostik unterstützt wird. Beispiele für solche Strukturen sind das finnische „part time special education system" (Kivirauma & Ruoho, 2007), das US-amerikanische „response-to-intervention-Modell" (Fuchs & Fuchs, 1986) oder das „system-monitoring" nach kanadischem Vorbild (Sliwka, 2010). Ziele dieser neueren und inklusionsorientierten Konzepte sind die Prävention von Lern- und Entwicklungsproblemen und eine Individualisierung der Förderung.

Alle genannten Modelle eint eine veränderte Sichtweise auf die pädagogische Diagnostik. Während in einem selektionspädagogischen Paradigma die „Etikettierung" von Schülerinnen und Schülern ein zentraler Auftrag der Diagnostik war, gerät in einem inklusionspädagogischen Paradigma die Evaluation von individuellen Unterstützungsmaßnahmen in den Mittelpunkt der diagnostischen Zielsetzung. Eine Evaluation individueller Entwicklungsverläufe und der Wirksamkeit von Fördermaßnahmen setzt diagnostische Verfahren voraus, die über einen längeren Zeitraum zu vielen Messzeitpunkten Verhaltensdaten erheben und analysieren (Grosche & Huber, in Druck).

Verlaufsdiagnostik

Im Folgenden wird das Spektrum vorhandener Verfahren der Verlaufsdiagnostik von curriculumsbasierten Messungen bis zur Verhaltensbeobachtung und –bewertung skizziert. Abschließend wird mit dem Direct Behavior Rating (im Folgenden DBR) bzw. der direkten Verhaltensbeurteilung (Chafouleas, Riley-Tillman & Christ, 2009; Steege, Davin & Hathaway, 2001) ein alternativer Ansatz zur Verhaltensverlaufsdiagnostik vorgestellt, der auch im Mittelpunkt der vorliegenden Arbeit steht.

Diagnostik kognitiver Eigenschaften: Curriculumsbasierte Messungen

Instrumente zur kontinuierlichen Leistungsmessung werden als curriculumsbasierte Messinstrumente („curriculum-based measurements", CBM) bezeichnet (Deno & Fal, 2003). Dabei handelt es sich in der Regel um kurze Tests (Durchführungszeit zwischen 1 und 3 Minuten), die meist auf einen bestimmten (kleinen) Teil des Lernverlaufs bezogen sind (Klauer, 2006). Aufgrund ihrer Charakteristika als Verfahren zur Verlaufsdiagnostik bestehen CBM-Instrumente aus einer Vielzahl vergleichbarer „Paralleltest" (z.B. 20 oder 50) mit vergleichbaren Testgütekriterien (Hosp, Hosp & Howell, 2007; Klauer, 2006; Strathmann & Klauer, 2010). Insofern kann man davon ausgehen, dass sich bei Verwendung von CBM in Bezug auf die Testgütekriterien Objektivität, Reliabilität und Validität keine konzeptuell nennenswerten Unterschiede zwischen den Testsituationen ergeben. CBM können sowohl auf der Basis von Einzelfällen als auch auf der Basis von Aggregaten (z.B. Schulklassen) valide ausgewertet werden (zur Auswertung in Einzelfällen vgl. Wilbert, 2014).

Im englischen Sprachraum gibt es für nahezu alle Fachbereiche schulischer Förderung zahlreiche CBMs. Auch im deutschen Sprachraum erscheinen mittlerweile erste Instrumente, die insbesondere die Bereiche Lesen (Diehl & Hartke, 2012; Walter, 2008) oder Mathematik (Müller & Hartmann, im Druck; Strathmann & Klauer, 2012) abdecken. Lehrkräfte und Schulpsychologen können mit den zur Verfügung stehenden Verfahren regelmäßig und ohne viel Aufwand in kürzeren Abständen (z.B. alle 1-2 Wochen) prüfen, ob die eingeleitete Unter-

stützung bzw. Förderung bei einem Kind „wirkt" oder ob sie verändert werden muss (Huber & Grosche, 2012). Während sich die Diskussion bislang vorwiegend auf kognitive Variablen (z.B. Lesekompetenzen, Rechtschreibung, Rechenleistung) fokussierte, bleibt die Frage offen, wie Interventionen bei Verhaltensproblemen analog, engmaschig und ökonomisch evaluiert werden können.

Gängige Methoden der Verhaltensbeobachtung

Bislang wurden zur Diagnostik von Verhaltensproblemen vor allem zwei grundsätzliche Methoden eingesetzt: 1) die systematische Verhaltensbeobachtung und 2) die Verhaltensbeurteilung (Schmidt-Atzert, 2012). Jedoch liegt das Hauptaugenmerk der Anwendung dieser Methoden in der Praxis vornehmlich auf der einmaligen Einschätzung von Verhaltensproblemen und damit der Statusdiagnostik (Amelang & Schmidt-Atzert, 2006; Schmidt-Atzert, 2012; Wittchen & Hoyer, 2011).

Betrachtet man diese Situation vor dem Hintergrund des inklusiven Wandels, fehlen somit prozessdiagnostische Instrumente für die größte zu integrierende Gruppe: Schülerinnen und Schülern mit Förderbedarf in ihrer emotional-sozialen Entwicklung (Chafouleas, Riley-Tillman & Christ, 2009; Christ, Riley-Tillman & Chafouleas, 2009). Insbesondere für diese Kinder und Jugendlichen ist es wichtig zu wissen, inwieweit aufwändige Unterstützungsmaßnahmen zu einem gewünschten Erfolg führen. Profitiert ein Schüler oder eine Schülerin von einer Unterstützung nicht in der erhofften Weise, können (z.B. analog zur Förderung im Bereich des Lesens oder der Mathematik) so die eingesetzten personellen oder materiellen Ressourcen weiter optimiert und an die Bedürfnisse der Schülerinnen und Schüler angepasst werden. Dieses Prinzip entspricht damit einem förderdiagnostischen Vorgehen, wie es zum Beispiel von Kooij (2004) beschrieben wird.

Aufbauend auf einem kurzen Überblick über die klassischen Ansätze zur Diagnostik von Verhalten (direkte systematische Verhaltensbeobachtung und Verhaltensbeurteilung) soll im Folgenden auf das im englischen Sprachraum seit mehreren Jahren diskutierte „Direct Behavior Rating" (im Folgenden DBR bzw. DBR-Rater) eingegangen werden. Im Gegensatz zu den traditionellen Methoden der Verhaltensbeobachtung ist DBR eher auf die Diagnose von Verhaltensentwicklungen und somit weniger auf eine statusdiagnostische Zielsetzung ausgerichtet.

Die direkte systematische Verhaltensbeobachtung

Die direkte systematische Verhaltensbeobachtung (engl. Systematic Direct Observation) kann als exakteste und differenzierteste Form der Verhaltensbeobachtung betrachtet werden (Amelang & Schmidt-Atzert, 2006). Im Folgenden wird analog zum englischen Fachbegriff die Abkürzung SDO (bzw. SDO-Rater) verwendet. Eine direkte systematische Verhaltensbeobachtung fokussiert in der Regel einen bestimmten Verhaltensausschnitt (z.B. Aggressivität) durch die Erfassung mehrerer Zeit- oder Ereignisstichproben, bei der zumeist Auftretenshäufigkeiten gezählt und somit quantifizierbar gemacht werden (Hussy, Schreier & Echterhoff, 2013). Die systematische Beobachtung erfordert dabei in der Regel eine genaue Operationalisierung des einzuschätzenden Verhaltens und somit eine intensive Schulung der Personen, die die Verhaltensbeobachtung durchführen sollen. Schmidt-Atzert (2012) unterscheidet bei der systematischen Verhaltensbeobachtung Zeichensysteme von Kategoriensystemen. Bei Zeichensystemen erfasst die beobachtende Person nur einzelne gezielte Verhaltensweisen, die für eine bestimmte Fragestellung relevant sind. Im Gegensatz dazu wird bei Kategoriensystemen ein breiterer Bereich eines Verhaltens erfasst und mehr oder weniger klar voneinander abgegrenzten Kategorien zu-

geordnet. Die beobachtende Person muss dabei der Beobachtungssituation direkt oder indirekt beiwohnen. Das bedeutet, sie muss entweder selbst in der Situation (aktiv oder passiv) anwesend sein oder das Verhalten anhand eines Videomittschnitts bewerten (Amelang & Schmidt-Atzert, 2006; Schmidt-Atzert, 2012).

Die Verhaltensbeurteilung

Die Verhaltensbeurteilung ist im Gegensatz zur systematischen Verhaltensbeobachtung eine abstrahierende Methode, die den Beobachter nach seiner Einschätzung im Hinblick auf ein zuvor definiertes Verhalten in einem bestimmten Zeitraum befragt (Schmidt-Atzert, 2012). Diese Einschätzung wird in der Regel über mehrstufige Ratingskalen erfasst (Schmidt-Atzert, 2010b; Stemmler, 2010). Im Gegensatz zu einer systematischen Beobachtung werden bei der Verhaltensbeurteilung keine konkreten Verhaltensweisen gezählt, sondern es wird auf die Erfahrung bzw. Wahrnehmung einer Person, die einer Beobachtungssituation beigewohnt hat, im Nachhinein zurückgegriffen. Verhaltensbeurteilungen können somit auch Tage und Wochen nach dem Ende einer Beobachtungssituation angewendet werden oder eine Verhaltenseinschätzung über mehrere längere Beobachtungszeiträume (z.B. mehrere Unterrichtsstunden, Schulfächer oder Wochen) zusammengefasst abfragen (Schmidt-Atzert, 2012; Ziegler & Bühner, 2012). Christ, Riley-Tillman und Chafouleas (2009) unterscheiden in diesem Zusammenhang die Einschätzung des Verhaltens durch die beobachtende Person selbst von einer indirekten Erfassung des Verhaltens. Bei Letzterem wird die beobachtende Person durch eine weitere Person (z.B. eine sonderpädagogische Fachkraft oder einen Schulpsychologen) im Hinblick auf einen bestimmten Verhaltensausschnitt im Rahmen eines Interviews befragt. Diese Form der Verhaltenserfassung kann wiederum strukturiert anhand von spezifischen Leitfadeninterviews oder offen bzw. narrativ geführt werden (Christ, Riley-Tillman & Chafouleas, 2009; Fydrich, 2012). Immer wieder diskutiert wird, ob Verhaltensbeurteilungen besonders anfällig für Urteilsfehler sind, da sie sehr unterschiedliche Verhaltensausschnitte mit einer stärkeren zeitlichen Verzögerung zusammenfassen (Schmidt-Atzert, 2010b).

Insgesamt ermöglichen direkte systematische Verhaltensbeobachtungen zwar einerseits eine exaktere Erfassung des Verhaltens, sie erfordern jedoch andererseits auch einen wesentlichen höheren Einsatz von Zeit und eine umfassende Schulung der beobachtenden Person. Verhaltensbeurteilungen sind ökonomischer in der Durchführung, sie sind jedoch mit einem deutlich höheren Risiko für Urteilsfehler behaftet.

Direct Behavior Rating (DBR)

Die Methode des Direct Behavior Ratings (Direkte Verhaltensbeurteilung) stellt eine Kombination aus einer direkten systematischen Verhaltensbeobachtung und einer Verhaltensbeurteilung mit Hilfe einer Ratingskala dar (Chafouleas, 2011; Christ, Riley-Tillman & Chafouleas, 2009), wobei diese Ratingskala nominalskalierte, ordinalskalierte oder intervall- bzw. verhältnisskalierte Merkmale umfassen kann.

Erstmalig dargestellt wurde das Prinzip des DBR von Steege et al. (2001) bzw. Chafouleas, Riley-Tillman und Sassu (2002). Ursprüngliches Ziel war die Verhaltensbeurteilung im Rahmen von Tagesbeurteilungslisten in der Schule (engl. daily behavior repord cards), die wiederum an Tokensysteme und Verstärker-Programme gebunden waren (Chafouleas et al., 2002; Christ, Riley-Tillman & Chafouleas, 2009; Huber, 2013). Direkte Verhaltensbeurteilung bedeutet zunächst, dass die Beurteilung eines Verhaltens immer in und/ oder direkt nach der Situation erfolgt, in der das Verhalten beobachtet werden sollte. Die Beobachtungszeiträume können nach Christ, Riley-Tillman und Chafouleas (2009) zwischen wenigen Sekunden bis hin zu einem Tag andauern. Demnach unterliegt die „Direkt-

heit“ der Methode einem gewissen Interpretationsspielraum, der in der Praxis je nach erforderlicher Ökonomie und Auftretenswahrscheinlichkeit des Verhaltens angepasst werden kann.

Ziel einer direkten Verhaltensbeurteilung ist es, die Verzögerung zwischen dem Auftreten eines Verhaltens und dessen Beurteilung gegenüber einer klassischen Verhaltensbeurteilung stark zu reduzieren und somit eine (vom Aufwand her ökonomische) Einschätzung mit einer hohen „situativen Validität“ verbinden zu können.

Direkte Verhaltensbeurteilungen orientieren sich in der Regel an Zeichensystemen. Das bedeutet, dass ausschließlich ein zuvor definiertes Zielverhalten bewertet wird und andere Verhaltensbereiche nicht erfasst werden. Im Unterschied zu Zeichen- oder Kategoriensystemen erfolgt die Messung dabei jedoch nicht durch explizites Zählen oder Ausmessen von Zielverhaltensweisen innerhalb eines definierten Zeitraums (z.B. Peter hat sich innerhalb von 10 Minuten 4 Minuten am Unterricht beteiligt), sondern (wie bei einer Verhaltensbeurteilung) mit Hilfe einer Ratingskala (z.B. von 0 = geringe Ausprägung bis 10 = starke Ausprägung) (Chafouleas, 2011; Christ, Riley-Tillman & Chafouleas, 2009).

Das besondere Kennzeichen der direkten Verhaltensbeurteilung betrifft ihre Einsatzmöglichkeiten. Während systematische Verhaltensbeobachtungen und Verhaltensbeurteilungen in erster Linie zur Statusdiagnostik im Bereich Verhalten genutzt werden und mit einem entsprechenden Aufwand verbunden sind, kann der primäre Einsatzbereich einer direkten Verhaltensbeurteilung in der Prozess- oder Modifikationsdiagnostik gesehen werden (Chafouleas, 2011; Christ, Riley-Tillman & Chafouleas, 2009; Riley-Tillman, Christ, Chafouleas, Boice Mallach & Briesch, 2011). Daher können DBR-Skalen in präventiven gestuften Förderprogrammen wie z.B. response-to-intervention (Huber & Grosche, 2012; Johnson, Fuchs & McKnight, 2006) einfach und effizient eingesetzt werden. DBR-Skalen eignen sich somit in besonderer Weise zur Evaluation von Förderhypothesen und Interventionen.

Abbildung 1 nimmt eine Einordnung der direkten systematischen Verhaltensbeob-

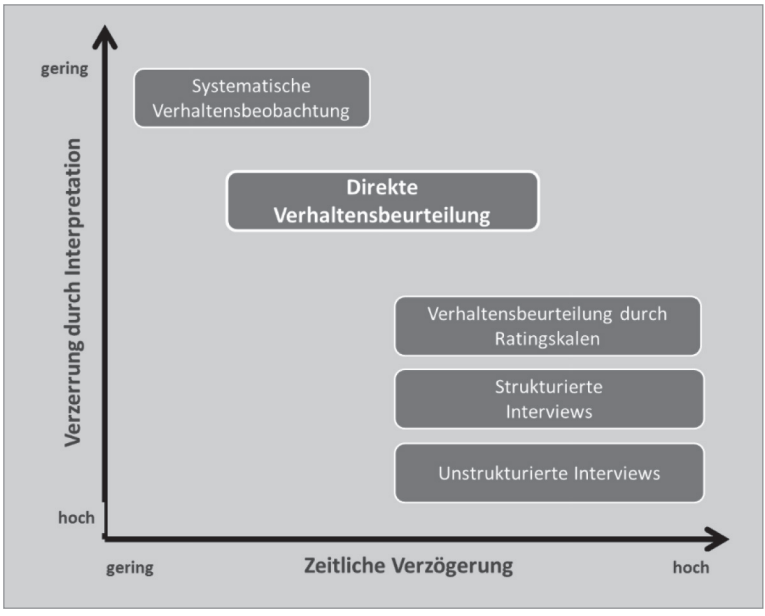

Abbildung 1: Einordnung der direkten Verhaltensbeurteilung

achtung und der Verhaltensbeurteilung anhand der Bereiche „Anfälligkeit für Interpretationsfehler“ und „zeitliche Verzögerung der Erfassung“ in Anlehnung an Christ, Riley-Tillman und Chafouleas (2009) vor.

Im Folgenden sollen nun die bislang vorliegenden empirischen Befunde für die Testgüte und die Verwendung von direkten Verhaltensbeurteilungen zusammengefasst werden.

Systematisches Review methodisch-empirischer Studien zum DBR

Generell unterliegt die „Qualität“ von Verhaltensbeobachtungen Einschränkungen, z.B. durch Wahrnehmungs- und Urteilsfehler (Schmidt-Atzert, 2010a, Schmidt-Atzert, 2012; Westhoff et al., 2010). Unabhängig von diesen grundsätzlichen Einschränkungen gibt es eine intensive Diskussion darüber, inwieweit die klassischen Testgütekriterien (wie z.B. Objektivität, Reliabilität und Validität) auf den Bereich der Verhaltensbeobachtung übertragbar sind. Während einige Autoren argumentieren, dass die Güte von Verhaltensbeobachtungen nicht durch die klassischen Testgütekriterien der Psychometrie abgebildet werden kann (z.B. Nelson, Hay & Hay, 1977), diskutiert Cone (1988), dass es allenfalls konzeptuelle, aber keine methodischen Unterschiede zwischen Verhaltensbeobachtungen und psychometrischen Tests gäbe und somit Objektivität, Reliabilität und Validität einer Verhaltensbeobachtung relevante Hauptgütekriterien zur Bestimmung ihrer Eignung sind. Diese „Übertragbarkeit“ leitet z.B. Cone (1988) aus den Annahmen ab, dass Verhaltensbeobachtungen a) möglichst genau mit dem tatsächlich gezeigten „wahren“ Verhalten übereinstimmen sollten, b) eine Person das gleiche Verhalten zu zwei Zeitpunkten auch gleich beurteilt und c) dass das gleiche Verhalten in unterschiedlichen Settings stabil bewertet wird. Amelang und Schmidt-Atzert (2006) weisen darauf hin, dass die – wie auch immer zu definierende – Güte einer Verhaltensbeobachtung in jeder Untersuchung neu bewertet werden müsse, da sich Beobachtungsgegenstand, Beobachtungssituation und beobachtende Personen fortlaufend ändern.

Bortz und Döring (2006) gehen bei der Diskussion der Güte von Verhaltensbeobachtungen pragmatisch dazu über, geeignete Maße zur Bestimmung der Beobachterübereinstimmung (Intrarater-Reliabilität und Interrater-Reliabilität) zu diskutieren. Dem voraus geht die bisher noch nicht hinreichend diskutierte Frage, inwieweit bei einem Instrument der Verhaltensverlaufsdiagnostik eine intersubjektive Übereinstimmung mit einer Norm oder einem Kriterium gewährleistet werden muss, wenn nur die Entwicklung des Verhaltens über die Zeit abgebildet werden soll. So wäre es bei einem Kind, dass häufig den Unterricht stört, unerheblich, ob störende Verhalten noch als normal, überdurchschnittlich oder stark überdurchschnittlich bewertet werden kann, wenn sich die verantwortlichen Akteure (z.B. Lehrkräfte, Eltern, Kind, Schulpsychologen, etc.) einig sind, dass eine Abnahme der Störungen im Unterricht hilfreich wäre. Weiterhin stellt sich die Frage nach dem Stellenwert der Interrater-Reliabilität, wenn davon ausgegangen wird, dass die Ratings immer konstant von der gleichen Person oder sogar dialogisch zwischen den gleichen zwei Akteuren (z.B. Lehrkraft und Kind) vorgenommen werden. Generell ist festzuhalten, dass die messtheoretische Fundierung für die Methode DBR bislang nicht hinreichend diskutiert wurde, sodass die klassischen Konzepte der Testtheorie vorerst nur unter Vorbehalten verwendet werden sollten.

Im Folgenden werden die zentralen empirischen Arbeiten zum DBR hinsichtlich der Gütekriterien Validität und Reliabilität im Rahmen eines systematischen Reviews analysiert und diskutiert. Bei einem systematischen Review handelt es sich um ein eine Technik der Sekundärforschung, die einer klassischen Metaanalyse dahingehend

ähnelt, dass verschiedene Studien systematisch vergleichend betrachtet werden. Im Gegensatz zur Metaanalyse wird in Bezug auf die Zielvariable („die Güte von DBR") keine vergleichende Metrik gewählt, da dies (wie sich noch zeigen wird) in Anbetracht der vielfältigen Fragestellungen, Herangehensweisen und Auswertungsmethoden nicht möglich ist. Der Vergleich der verschiedenen Studien und die Identifikation von Wirkgrößen werden also nicht statistisch, sondern auf hermeneutisch-qualitativer Ebene vorgenommen. Dieser Ansatz erlaubt die Berücksichtigung der zahlreichen qualitativen wie methodischen Unterschiede zwischen den verschiedenen Studien (vgl. zum systematischen Review und zur Abgrenzung zur Metaanalyse auch Bortz und Döring (2006)). Abschließend werden aus den Ergebnissen des systematischen Reviews allgemeine Ableitungen zur Validität und Reliabilität von DBR-Skalen vorgenommen.

Auswahl der Studien

Zur Recherche der in dieses Review einfließenden Studien sind insgesamt vier internationale Datenbanken (Psyndex, PsychArticles, Psychinfo und ERIC) unter dem Stichwort „direct behavior rating" und „behavioral recording" durchsucht worden. Zudem wurde die deutschsprachige Datenbank FIS-Bildung mit Hilfe der direkten Übersetzung (direkte Verhaltensbeurteilung) durchsucht. Aufgenommen in die folgende Diskussion wurden ausschließlich deutsch- oder englischsprachige Studien, die in Journalen mit Peer-Review-Verfahren erschienen sind und sich mit der Verhaltensbeurteilung im pädagogisch-psychologischen Kontext befassten. So wurden zum Beispiel alle Studien, die Instrumente zur Verhaltensanalyse im Rahmen von Tierversuchen evaluierten, nicht in die vorliegende Arbeit aufgenommen.

Insgesamt ergaben die Recherchen zum „direct behavior rating" 25 und zum „behavioral recording" 24 Treffer. Darunter waren insgesamt 19 empirische Studien zur Verhaltensdiagnostik und drei Review-Artikel zu finden. Zwei Arbeiten waren in spanischer und italienischer Sprache und wurden daher nicht aufgenommen. Fünf Arbeiten haben sich eher mit konzeptionellen Fragen der Implementation von DBR in ein gestuftes präventives Förderkonzept (wie z.B. response-to-intervention) befasst. Die im Rahmen dieser Übersichtsarbeit eingeflossenen 17 Studien sind in Tabelle 1 chronologisch und mit einem Hinweis zu ihrer inhaltlichen Schwerpunktsetzung aufgeführt.

Im Folgenden werden die Ergebnisse dieser Studien zusammengefasst und vor dem Hintergrund der praktischen Anwendung in der Verlaufsdiagnostik diskutiert. Mit Ausnahme einer Studie untersuchten alle recherchierten Arbeiten die Beobachtungsgüte von Single-Item-Scales (SIS). Bei diesem Skalenformat wird jeweils nur ein einziges Item wiederholt beurteilt. Daher sind alle hier skizzierten Befunde in erster Linie für diesen Skalentyp gültig. Nur eine Studie (Volpe & Briesch, 2012) hat SIS und Multi-Items-Scales (MIS) im Hinblick auf die Beobachtungsgüte miteinander verglichen.

Bewertung der Beobachtungsgüte in den recherchierten Studien

Die Beobachtungsgüte wurde in den recherchierten Studien an drei Kriterien festgemacht:

Erstens die Übereinstimmung mit einem „wahren" Verhaltenswert. Dieses Kriterium diente vornehmlich der Prüfung der Validität. Dabei wurde jeweils geprüft, inwieweit die DBR-Messungen mit einem vermeintlich besseren Indikator für das Verhalten der Kinder übereinstimmen. Zwei Studien (Kilgus, Chafouleas, Riley-Tillman & Welsh, 2012; Kilgus, Riley-Tillman, Chafouleas, Christ & Welsh, 2014) setzten als Indikator für das „wahre Verhalten" ein standardisiertes Verfahren zur Verhaltensbeurteilung ein. In acht Studien wurde der „wahre Verhaltenswert" über eine direkte systematische Verhaltensbeobachtung durch zwei professionelle und geschulte Rater operationalisiert (Chafoule-

Tabelle 1: Studien im Review

Nr.	Autoren	Titel	Jahr	Schwerpunkt
1	Steege, Davin & Hathaway	Reliability and Accuracy of a Performance-Based Behavioral Recording Procedure.	2001	1
2	Chafouleas, Christ, Riley-Tillman, Briesch & Chanese	Generalizability and Dependabilty of Direct Behavior Rating to Assess Social Behavior of Preschoolers	2007	1
3	Riley-Tillman, Chafouleas, Sassu, Chanese & Glazer	Examining the Agreement of Direct Behavior Rating and Systematic Direct Observation Data for On-Task and Disruptive Behavior	2008	4
4	Schlientz, Riley-Tillman, Briesch, Walcott & Chafouleas	The Impact of Training on the Accuracy of Direct Behavior Ratings (DBR)	2009	9
5	Chafouleas, Christ & Riley-Tillman	Generalizability of Scaling Gradients on Direct Behavior Ratings	2009	1, 2
6	Riley-Tillman, Chafouleas, Christ, Briesch & LeBel	The Impact of Item Wording and Behavioral Specificity on the Accuracy of Direct Behavior Rating (DBRs)	2009	4, 5, 6
7	Briesch, Chafouleas & Riley-Tillman	Generalizability and Dependability of Behavior Assessment Methods to Estimate Academic Engagement: A Comparison of Systematic Direct Observation and Direct Behavior Rating	2010	1, 7
8	LeBel, Kilgus, Briesch & Chafouleas	The Impact of Training on the Accuray of Teacher-Completed Direct Behavior Ratings (DBR)	2010	9
9	Christ, Riley-Tillman, Chafouleas & Boice	Direct Behavior rating (DBR): Generalizability and Dependability Across Raters and Observations	2010	1, 2, 7
10	Chafouleas, Briesch, Riley-Tillman, Christ, Blak & Kilgus	An Investigation of the generalizability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students	2010	1
11	Riley-Tillman, Christ, Chafouleas, Boice Mallach & Briesch	The Impact of Observation Duration on the Accuracy of Date Obtained From Direct Behavior Rating (DBR)	2011	2, 4, 8
12	Christ, Riley-Tillman, Chafouleas & Jaffery	Direct Behavior Rating: An Evaluation of Alternate Definitions to Assess Classroom Behaviors	2011	4, 5
13	Chafouleas, Kilgus, Riley-Tillman, Jaffery & Harrison	Preliminary evaluation of various training components on accuracy of Direct Behavior Rating	2012	9
14	Volpe & Briesch	Generalizability and Dependability of ingle-Item and Multiple-Item Direct Behavior Rating Scales for Engagement and Disruptive Behavior	2012	1, 2, 7, 9
15	Briesch, Kilgus, Chafouleas, Riley-Tillman & Christ	The Influence of Alternative Scale Formats on the Generalizability of Data Obtained from Direct Behavior Rating Single-Item Scales (DBR-SIS)	2012	2, 9
16	Kilgus, Chafouleas, Riley-Tillman & Welsh	Direct behavior rating scales as screeners: A preliminary investigation of diagnostic accuracy in elementary school	2012	9
17	Chafouleas, Jaffery, Riley-Tillman, Christ & Sen	The Impact of Target, Wording, and Duration on Rating Accuracy for Direct Behavior Rating	2013	4, 6

Anmerkungen: 1 = Beobachtungsgüte; 2 = Skalendesign; 3 = Anzahl der Items; 4 = Wahl des Beobachtungsziels; 5 = Formulierung des Beobachtungsziels; 6 = Valenz der Zielformulierung; 7 = Anzahl der Messzeitpunkte; 8 = Länge der Verhaltensstichproben; 9 = Auswirkungen eines Beobachtertrainings

as, Briesch, Riley-Tillman, Christ, Black & Kilgus, 2010; LeBel, Kilgus, Briesch & Chafouleas, 2009; Riley-Tillman et al., 2011; Riley-Tillman, Chafouleas, Christ, Briesch & LeBel, 2009; Riley-Tillman, Chafouleas, Sassu, Chanese & Glazer, 2008a; Schlientz, Riley-Tillman, Briesch, Walcott & Chafouleas, 2009). Die professionellen Rater haben dabei entweder das Verhalten eines oder mehrerer Schülerinnen und Schüler direkt im Feld beurteilt oder haben die Verhaltensbeobachtung auf der Grundlage von Experimentalvideos vorgenommen. In den Studien wurden Bedingungen als günstiger bewertet, wenn die Übereinstimmung mit dem „wahren Wert" höher war.

Zweitens der Anteil der Varianz in den Messdaten, der durch verschiedene Facetten erzeugt wird. Sieben Studien basieren auf der Generalisierungstheorie (Briesch, 2010; Briesch, Kilgus, Chafouleas, Riley-Tillman & Christ, 2012; Chafouleas, Christ, Riley-Tillman, Briesch & Chanese, 2007; Chafouleas et al., 2009; Chafouleas et al., 2010; Christ, Riley-Tillman, Chafouleas & Boice, 2010a; Volpe & Briesch, 2012). Der Ansatz basiert auf der Idee, dass Varianzen in den Messwerten auf unterschiedliche Einflussfaktoren (Facetten) der Messung zurückführbar sind. In den Studien geht es dabei meist um die Untersuchung der Interrater-Reliabilität (Varianzanteil, der auf die verschiedenen Rater zurückzuführen ist) von DBR-Instrumenten.

Das dritte Kriterium zur Bewertung der Beobachtungsgüte war die Stabilität der DBR-Beurteilungen desselben Experimentalvideos zu zwei unterschiedlichen Zeitpunkten. Dieses Kriterium kann analog zur Test-Retest-Reliabilität und damit zur Intrarater-Reliabilität interpretiert werden. Unter den recherchierten Arbeiten untersuchte nur eine Studie (Christ, Riley-Tillman, Chafouleas & Boice, 2010b) diesen Indikator.

Das folgende Review ist in zwei Teile unterteilt. Im ersten Teil sind Studien zusammengefasst, die sich grundsätzlich mit der Beobachtungsgüte von DBR-Skalen befasst haben. Hier sind insbesondere Studien beschrieben, die unter Verwendung der Generalisierungstheorie die Frage untersuchen, wieviel Prozent der Varianz von DBR-Messreihen auf die Facetten von Ratern, Situationen und beobachteten Personen zurückführbar sind. Im zweiten Teil des Reviews sind Studien zusammengefasst, die eine Optimierung der Beobachtungsgenauigkeit durch eine systematische Veränderung der Rahmenbedingung einer DBR-Messung untersuchten. Aus den Arbeiten im zweiten Teil lassen sich nach Auffassung der Autoren insgesamt acht Untersuchungsschwerpunkte bilden. Da fast alle Studien gleichzeitig mehrere Fragestellungen prüften, werden die Arbeiten zum Teil unter mehreren Untersuchungsschwerpunkten zitiert.

Teil 1: Grundlegende Studien zur Untersuchung der Beobachtungsgüte von DBR

Grundsätzlich ist davon auszugehen, dass eine zufriedenstellende Interrater-Reliabilität unter Verwendung von DBR schwieriger zu erreichen ist, als bei SDO-Messungen. In einer Generalisierungsstudie zeigten Briesch, Chafouleas und Riley-Tillman (2010), dass sich bei beiden Methoden zwischen 47% und 48% der Varianz der Beurteilungen durch die Varianz des Schülerverhaltens erklären lassen. In der Gruppe der Lehrkräfte, die DBR als Methode nutzten, war die restliche Varianz jedoch zu einem großen Teil (20%) durch eine Interaktion zwischen der beobachtenden Lehrkraft und beurteiltem Zielschüler erklärbar. Demgegenüber war in der Gruppe, die eine direkte systematische Verhaltensbeobachtung als Methode nutzten, weniger als 1% der Varianz durch diesen Interaktionseffekt erklärbar (Briesch et al., 2010). Dieses Phänomen wird durch mehrere vergleichbar aufgebaute Studien bestätigt (Chafouleas et al., 2007; Chafouleas et al., 2009; Chafouleas et al., 2010; Christ et al., 2010a; Hintze & Matthews, 2004).

Obwohl bei beiden Methoden ein vergleichbarer Anteil der Varianz durch das

Verhalten des Zielschülers erklärbar ist, sind direkte Verhaltensbeurteilungen demnach deutlich stärker durch die beobachtende Person geprägt. Insgesamt drei Forschergruppen analysierten die Urteilsfehler genauer. Eine der ersten Studien führten Steege et al. (2001) durch. In einer recht kleinen Untersuchung mit fünf Schülerinnen und Schülern zeigten die Autoren, dass die jeweiligen Lehrkräfte die aktive Teilnahme und das stereotype Verhalten der Schülerinnen und Schüler vergleichbar beurteilten wie zwei professionelle Rater, die zu 30 Zeitpunkten eine direkte systematische Verhaltensbeobachtung vornahmen. Beide Ratergruppen sollten auf einer 6-fach unterteilten Likert-Skala angeben, wieviel Zeit einer 15-minütigen Beobachtungseinheit die Schülerinnen und Schüler das Zielverhalten zeigten. In der Studie nahmen DBR-Rater und SDO-Rater in 94% bzw. 95% der Situationen identische Ratings vor (Steege et al., 2001). Dieser Befund spricht wiederum für eine insgesamt gute Kriterumsvalidität und eine vergleichsweise hohe Interraterübereinstimmung von DBR-Messungen. Riley-Tillman et al. (2008a) und Christ, Riley-Tillman, Chafouleas und Jaffery (2011) analysierten die Urteilsfehler differenzierter und zeigten, dass die Beobachtungsgüte stark von dem beobachteten Verhaltensaspekt abhängig ist. Die erste Autorengruppe untersuchte die Beobachtungsgüte für die Bereiche Unterrichtsbeteiligung und störendes Verhalten (siehe auch unten). Die Befunde machten deutlich, dass die DBR-Werte in 67% bis 93% der Fälle nur um +/- 1 Punkt von dem Wert abwichen, der durch eine direkte systematische Verhaltensbeurteilung zustande kam. Sogar 100% der DBR-Rater lagen innerhalb einer 2-Punkte-Abweichung zu den SDO-Ratern. Christ, Riley-Tillman, Chafouleas und Jaffery (2011) betrachteten die durchschnittliche Richtung der Abweichung. Sie stellten bei ungeschulten DBR-Ratern und unter Verwendung einer SI-Skala eine Überschätzung des beobachteten Verhaltens um 1.5 Punkte fest. Dies bedeutet, dass die Probanden die Zeit, in der der Zielschüler störte oder sich am Unterricht beteiligte um ca. 15% gegenüber den geschulten Beobachtern (SDO) überschätzten (Riley-Tillman et al., 2011). Die bislang vorliegenden Studien lassen demnach keine grundsätzlichen Ableitungen zur Beobachtungsgüte von DBR-Beurteilungen zu. Die Befunde deuten generell darauf hin, dass DBR-Beurteilungen stärker durch die beobachtende Person selbst beeinflusst sind als dies bei einer direkten systematischen Verhaltensbeobachtung der Fall ist. Damit wären statusdiagnostische Entscheidungen, die auf der Grundlage von DBR-Messungen beruhen mit einer etwas höheren Unsicherheit verbunden als statusdiagnostische Entscheidungen, die auf Daten mittels SDO-Messungen fußen.

Ein weiteres Kriterium zur Bestimmung der Beobachtungsgüte von DBR-Skalen ist die Test-Retest-Reliabilität. Bislang untersuchten nur Riley-Tillman und Kollegen (2011) diesen Aspekt der Beobachtungsgüte. Im Rahmen dieser videogestützten Experimentalstudie untersuchten die Autoren die Übereinstimmung der DBR-Beurteilungen von 88 Studierenden zu zwei Messzeitpunkten. Die Probanden sollten hierzu die Teilnahme am Unterricht und die Unterrichtsstörungen von Zielschülern in zwei identischen Unterrichtsvideos in einem Abstand von einer Woche mit Hilfe einer elfstufigen DBR-Skala (Prozent der Zeit) beurteilen. Die Befunde ergaben, dass die Test-Retest-Reliabilität zwischen den Testzeitpunkten schwankte. Die höchste Test-Retest-Reliabilität ergab sich mit Werten von $r > 0.7$ für den Mittelwert von vier Einzelbeobachtungen jeweils im Test und im Retest (Riley-Tillman et al., 2011). Diese vergleichsweise hohe Übereinstimmung zeigt, dass selbst die unerfahrenen studentischen Rater mit Hilfe einer DBR-Skala zu zeitlich überdauernden Urteilen kommen und somit keiner Willkür unterliegen. Insgesamt interpretieren Christ, Riley-Tillman und Chafouleas (2009) und auch später Chafouleas (2011) die vorliegenden Befunde so, dass sie DBR sowohl für prozessdiagnosti-

sche Fragestellungen als auch für den statusdiagnostischen Einsatz empfehlen.

In der Vergangenheit zeigten verschiedene Arbeitsgruppen, dass die Beobachtungsgüte durch verschiedene Faktoren weiter erhöht werden kann. Die bislang untersuchten Faktoren zur Erhöhung der Beobachtungsgüte von DBR-Messungen werden nun im zweiten Teil des Reviews skizziert.

Teil 2: Studien zur Untersuchung spezifischer Aspekte der Methode DBR

Im zweiten Teil des Reviews sollen nun Studien zusammengefasst werden, die gezielt verschiedene Aspekte der Methode DBR untersuchten. Analog zu den in Tabelle 1 aufgeführten inhaltlichen Studienschwerpunkten lässt sich folgende Untergliederung formulieren:

1) Skalendesign
2) Anzahl der Items
3) Anzahl der Messzeitpunkte
4) Wahl des Beobachtungsziels
5) Formulierung des Beobachtungsziels
6) Valenz der Zielformulierung
7) Länge der Verhaltensstichproben
8) Auswirkungen eines Beobachtertrainings

Skalendesign. DBR-Messungen erfolgen über Ratingskalen. Mit Blick auf die Validität und die Reliabilität der durchgeführten Messungen stellt sich hier – wie aber auch sonst in der empirischen Forschung - die Frage nach der „richtigen" Zahl der Ausprägungen der jeweiligen Skalen (z.B. dreistufig, fünfstufig, elfstufig) und auch nach der Ausgestaltung der Skalen (numerische, verbale oder grafische Stützung). Prototypisch für das verwendete Stimulusmaterial sind in Abbildung 2 zwei Präsentationsformen dargestellt.

Briesch et al. (2012), Chafouleas et al. (2009) sowie Christ et al. (2010b) konnten bislang auf Grundlage verschiedener Studien, die alle mit dem messtheoretischen Ansatz der Generalisierbarkeitstheorie ausgewertet wurden, bei Designs mit sechs-, zehn- und 14- stufigen Skalen keinen nennenswerten Einfluss der Skalenbreite auf die durch die Lehrkräfte aufgeklärte Varianz nachweisen. Auch die Länge der Skalen hatte in den genannten Studien keinen signifikanten Einfluss auf die Beobachtungsgüte. Das bedeutet, dass die beurteilenden Lehrkräfte durch die Veränderungen der Skalenbreite nicht gleichzeitig zu genaueren (bzw. ungenaueren) Ratern wurden. Auf der Basis

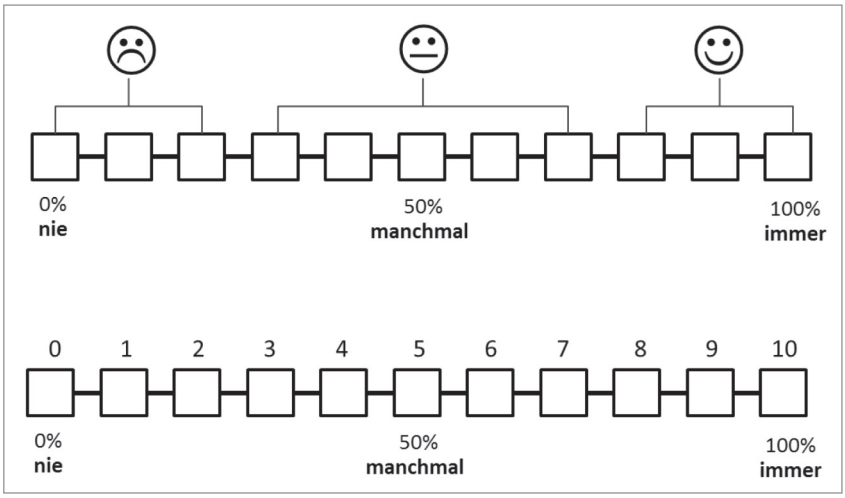

Abbildung 2: Elfstufige Skalen mit numerischer, grafischer und verbaler Stützung bzw. nur mit numerischer und verbaler Stützung

einer Sekundäranalyse empfehlen Christ, Riley-Tillman und Chafouleas (2009) die Verwendung von mindestens sechsstufigen Skalen, wobei die empirische Grundlage dieser Empfehlung offen bleibt. Riley-Tillman et al. (2011) untersuchten in einer weiteren Studie den Einfluss der Skalenbeschriftung (absolute Zeit in Sekunden vs. prozentualer Zeitanteil der Erreichung eines Verhaltensziels pro Zeiteinheit) auf die Beobachtungsgüte. Auch hier fanden die Autoren keinen signifikanten Einfluss auf die Interrater-Reliabilität. Chafouleas (2011) empfiehlt zusammenfassend eine prozentuale Skala, die in Zehnerschritten von 0% bis 100% reicht. Dieser Empfehlung folgend würde ein Rater mit Hilfe einer DBR-Skala einschätzen, wieviel Prozent der Beobachtungszeit eine Person ein zuvor definiertes Verhalten gezeigt hat. Allerdings lässt sich auch diese Schlussfolgerung, nach der ein prozentuales Maß anderen denkbaren Maßen überlegen ist, anhand der hier diskutierten Studien nicht empirisch absichern. So war auch unter den 17 betrachteten Veröffentlichungen keine Studie zu finden, die die zeitliche Verankerung der Beurteilung gegenüber anderen Verankerungen (z.B. einer subjektiven Einschätzung der Zielübereinstimmung) in einem vergleichenden Design absicherten. Letztendlich sind weitere Studien durchzuführen, die diese und weitere Fragestellungen in Bezug auf das Skalendesign systematisch untersuchen.

Anzahl der Items. Als weiterer relevanter Aspekt zur Erhöhung der Beobachtungsgüte wurde die Frage nach der Anzahl der Items, die pro Messung eingeschätzt werden sollten, *untersucht*. Hierbei lassen sich zwei grundsätzliche Varianten von DBR-Instrumenten unterscheiden. Zum einen die Einschätzung des Zielverhaltens über ein einzelnes Item (Single-Item-Scale, im Folgenden SIS). Zum anderen die Einschätzung des Zielverhaltens über mehrere (spezifischere) Items (Multi-Item-Scale, im Folgenden MIS). Bei einer MIS ist ein globales Verhaltensziel durch mehrere spezifische Items operationalisiert (siehe Tabelle 2). Im Rahmen des Reviews fand sich eine Arbeit, die unter der Verwendung der Generalisierbarkeitstheorie die Kriteriums-Validität von SIS und MIS vergleichend untersuchte (Volpe & Briesch, 2012).

Grundsätzlich handelt es sich bei der Generalisierbarkeitstheorie um Konzept zur Analyse der Messgenauigkeit eines Messinstruments. Der Ansatz geht dabei davon aus, dass sich alle Messfehler vollständig in verschiedene Komponenten aufteilen und einzelnen Faktoren (Facetten) zuordnen lassen. Die Analyse der Messfehler erfolgt in der Regel über eine Varianzanalyse. Dabei werden in einem ersten Schritt (eine sogenannte G-Study) verschiedene Varianzkomponenten ermittelt. In einem zweiten Schritt (eine sogenannte D-Study) wird wiederum geprüft unter welchen Bedingungen die Facetten der G-Studie unter dem Aspekt der Generalisierbarkeit optimal ausgeprägt wären (vgl. zum Vorgehen z.B. Eisend, 2007).

Volpe und Briesch (2012) haben in ihrer Studie pro zu beurteilendem Verhaltensbereich fünf Items formuliert, wobei unklar bleibt, auf welcher Datenbasis sie sich für diese Anzahl entschieden haben. Die Autoren zeigten im Rahmen ihrer Generalisierungsstudie (G-Study) mit zwei Beobachtern und sieben Zielschülern, dass bei einer SIS ein weitaus höherer Anteil Beurteilungsvarianz unerklärt blieb (ca. 33%) als bei einer MIS (zwischen 5% und 26%). Wenn die Verhaltensbeurteilung direkt im Feld (d.h. direkt in der Schulklasse und nicht in einer Laborsituation) vorgenommen wurde, stieg der Anteil der unerklärten Varianz bei der Verwendung einer SIS stärker an als bei einer MIS (Volpe & Briesch, 2012). In einer Entscheidungsstudie-Studie (D-Study) untersuchten die Autoren, ob es einen Unterschied zwischen einer SIS und einer MIS in Bezug auf die Anzahl der Messzeitpunkte bzw. Wiederholungsmessungen gibt, die notwendig sind, um eine akzeptable Messgenauigkeit zu erreichen. Vergleichsmaßstab war in dieser Studie immer das Ergebnis einer direkten systematischen Verhaltensbeobachtung durch zwei geschulte Ra-

ter. Als Maß für die Übereinstimmung legten Volpe und Briesch (2012) einen Wert von $\rho^2 = .8$ fest. Die Ergebnisse zeigten, dass dieses Kriterium bei der Verwendung von MI-Skalen mit weniger Messzeitpunkten erreicht wurde als bei der Verwendung einer SI-Skala. Die Autoren leiten aus diesen Befunden insgesamt eine höhere Messgenauigkeit von MI-Skalen ab. Sie diskutierten zudem, ob eine SIS im Vergleich zu einer MIS „sensitiver" ist und eher im Bereich der Veränderungsmessung verwendet werden kann. Einschränkend muss hier jedoch bedacht werden, dass diese Interpretation der Befunde problematisch ist. So handelt es sich bei einer Messung mit Hilfe einer MIS mit vier Items strenggenommen um vier Messungen. Da Mittelwerte aus mehreren Messungen in der Regel stabilere Kennwerte liefern als Werte von Einzelmessungen, ist es unklar, ob die bessere Messgüte wirklich eine Folge der spezifischeren Operationalisierung der Items ist oder ob es sich lediglich um das Resultat eines stabileren Mittelwerts handelt, welches auch bei der Verwendung von vier SIS-Messungen erzielt worden wäre.

Anzahl der Messzeitpunkte. Eine der ersten Studien zu dieser Fragestellung führten Hintze und Matthews (2004) durch. Im Mittelpunkt standen dabei nicht DBR-Skalen, sondern die Güte von einer direkten systematischen Verhaltensbeobachtung. Bei einer sehr konservativen Interpretation der Befunde kamen die Autoren zu dem Schluss, dass schon allein für diese aufwändige Form der Verhaltensdiagnostik zwischen sieben und 20 Beurteilungssequenzen notwendig sein könnten, um eine ausreichende Beurteilerübereinstimmung zu erreichen. Andere Autorengruppen teilten diese konservative Einschätzung nicht. Die umfänglichsten Studien in Bezug auf direkte Verhaltensbeurteilungen führten Briesch et al. (2010) sowie Christ et al. (2010b) durch. Unter Verwendung der Generalisierungstheorie zeigten die Autorengruppen im Rahmen von Entscheidungsstudien, dass für den Verhaltensbereich „Teilnahme am Unterricht" in der Regel rund fünf Beurteilungen à ca. 15 Minuten erforderlich sind, um einen reliabilitätsequivalenten Koeffzienten von $\rho^2 > .6$ zu erhalten und so eine akzeptable Übereinstimmung mit SDO-Ratern zu erreichen. Dabei machte es nur einen geringfügigen Unterschied, ob die Beurteilungen an einem einzigen Tag oder über mehrere Tage hinweg vorgenommen wurden (Briesch et al., 2010). Wichtige pädagogische Entscheidung sollten jedoch besser und auf einem höheren Level von $\rho^2 > .80$ abgesichert werden. Christ et al. (2010b) sowie Volpe und Briesch (2012) kamen für SI-Skalen zu dem Ergebnis, dass dieses strenge Kriterium erst nach weitaus mehr Messungen erreicht werden kann. Allerdings divergieren die Angaben zur Anzahl der notwendigen Messzeitpunkte über die verschiedenen Studien zwischen sieben Messungen für die „Teilnahme am Unterricht" (Volpe & Briesch, 2012) und bis zu 20 Messungen für „störendes Verhalten" (Volpe & Briesch, 2012) oder die aktive Beschäftigung mit einem Puzzle (Christ et al., 2010a). Insgesamt sind die Untersuchungsdesigns der Studien jedoch so unterschiedlich, dass die Befunde nur sehr eingeschränkt miteinander verglichen werden können. Beispielsweise ließen Briesch et al. (2010) jeweils drei Sequenzen zwischen neun und 17 Minuten für zwölf Kindergartenkinder beurteilen, Volpe und Briesch (2012) acht Siebtklässler in drei zehnminütigen Sequenzen und Christ et al. (2010b) 18 einminütige Clips mit Grundschülern.

Christ, Riley-Tillman und Chafouleas (2009) fassen den Stand der Forschung zusammen und folgern, dass zumindest für die gut erforschten Beobachtungsziele (Teilnahme am Unterricht und störendes Verhalten) bei Verwendung einer DBR-SI-Skala mindestens fünf Beurteilungssituationen (quasi als „Kalibrierungsphase") zur Erreichung einer zufriedenstellenden Validität und Reliabilität ausreichend sind. Zur Erreichung einer möglichst hohen diagnostischen Güte

sollten darüber hinaus eher längere Verhaltensstichproben als kurze Sequenzen eingeschätzt werden.

Wahl des Beobachtungsziels. Gleich mehrere Studien untersuchten die Messgenauigkeit von DBR-Skalen in Abhängigkeit von der Wahl des beobachteten Verhaltens. Die Hypothese dieser Studien bestand darin, dass nicht jedes Verhalten im Unterricht gleich exakt mit Hilfe von DBR-Skalen beobachtet werden kann. Die vorliegenden Befunde bestätigen diese Hypothese. So scheint die Beobachtung den Probanden in verschiedenen Studien immer dann mit größerer Genauigkeit (höhere Validität und höherer Interraterübereinstimmung) zu gelingen, wenn sie die „Teilnahme am Unterricht" (engl. academic engagement) oder das „störende Verhalten" (engl. disruptive behavior) einer Schülerin oder eines Schülers beobachten sollten (Chafouleas, Jaffery, Riley-Tillman, Christ & Sen, 2013; Christ, Riley-Tillman, Chafouleas & Jaffery, 2011; Riley-Tillman et al., 2009).

Christ, Riley-Tillman, Chafouleas und Jaffery (2011) zeigten, dass Korrelationen zwischen ungeschulten Verhaltensbeurteilern und geschulten Verhaltensbeobachtern für die Bereiche „Teilnahme am Unterricht" und „störendes Verhalten" stabil zwischen r = .67 und r=.78 variieren (Christ, Riley-Tillman, Chafouleas & Jaffery, 2011). Riley-Tillman et al. (2008a) ergänzen diese Befunde durch eine Feldstudie mit 15 Lehrkräften, die jeweils einen Grundschüler im Rahmen ihres eigenen Unterrichts beurteilen sollten. Die Autoren zeigten, dass die Einzelkorrelationen zwischen DBR- und SDO-Ratern in dem von Christ, Riley-Tillman, Chafouleas und Jaffery (2011) beschriebenen Bereich bleiben, wenn Beurteilungen direkt im Unterricht durch die jeweiligen Lehrkräfte erfolgen. Die Korrelationen stiegen weiter auf r=.81 (Teilnahme am Unterricht) und r=.87 (störendes Verhalten), wenn die Werte für die drei Beobachtungszeitpunkte in einem Mittelwert verrechnet wurden (Riley-Tillman et al., 2008b).

Etwas größere Diskrepanzen zwischen DBR und SDO wurden sichtbar, wenn die Probanden „angemessenes Verhalten" (engl. compliance) oder respektvolles Verhalten (engl. respectful behavior) beurteilen sollten. Für andere Beobachtungsbereiche (z.B. „Interaktion mit der Lehrkraft", „Interaktion mit Klassenkameraden" sowie für motorisches und sprachliches Verhalten) gaben die Autoren weniger robuste Korrelationen von r<.4 an. In einer weiteren Studie kamen Riley-Tillman et al. (2011) zu einem vergleichbaren Schluss. Tabelle 2 stellt die in den Untersuchungen verwendeten Definitionen der Beobachtungsziele zusammenfassend dar.

Formulierung des Beobachtungsziels. Christ, Riley-Tillman, Chafouleas und Jaffery (2011) untersuchten den Einfluss von verschiedenen Formulierungen des Beobachtungsziels bei SI-Skalen auf die Validität und die Reliabilität der direkten Verhaltensbeurteilung. Im Rahmen der Studie wurden 88 Probanden (Studierende) gebeten, das Verhalten von vier Schülerinnen und Schülern in fünf zweiminütigen Videoclips zu beurteilen. Das Kriterium zur Bestimmung der Validität war dabei die Übereinstimmung der DBR-Ratings mit zwei geschulten SDO-Ratern. Die Autoren zeigten, dass eine globale Zielformulierungen nahezu über alle Verhaltensbereiche hinweg einer spezifischen Zielformulierung überlegen war (Christ, Riley-Tillman, Chafouleas & Jaffery, 2011). Analog zu diesen Befunden war auch die Interrater-Reliabilität bei globalen Zielformulieren ebenfalls höher (zwischen r=.61 und r=.81), während sie bei spezifischen Zielformulierungen auf Werte zwischen r=.09 und r=.60 absank (Christ, Riley-Tillman, Chafouleas & Jaffery, 2011). Zu vergleichbaren Ergebnissen kamen auch Riley-Tillman, et al. (2009): In einem experimentellen Design mit 145 studentischen Probanden zeigte die Autorengruppe, dass die Genauigkeit der DBR-Messungen im Vergleich zum „wahren" Wert steigt, wenn die Zielformulierung global (z.B. Teilnahme am Unter-

Tabelle 2: Operationalisierung von SIS und MIS

Bereich	Operationalisierung SI-Skala[1]	Operationalisierung MI-Skala[2]
Teilnahme am Unterricht (engl. academic engagement)	*Academically engaged:* actively or passively participating in the classroom activity. *Examples:* writing, raising his or her hand, answering a question, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials.	– Finishes work on time – Actively participates in class – Raises hand when appropriate – Works hard – Stays on task
Störendes Verhalten (engl. disruptive behavior)	*Disruptive:* action that interrupts regular school or classroom activity. *Examples:* out of seat, fidgeting, calling out, talking/ yelling about things that are unrelated to classroom instruction, acting aggressively, playing with objects	– Calls out – Noisy – Clowns around – Talks to classmates when inappropriate – Out of seat/area
Nicht-Respektvolles Verhalten (engl. disrespectful behavior / compliance)	*Disrespectful:* noncompliant, defiant, insubordinate, or socially rude behavior in response to adult directions and/or interactions with peers and adults. *Examples:* refusal to follow teacher directions, talking back, eyerolling, inappropriate gesture, inappropriate language and/or social interactions, disruption with negative tone/connotation.	

Anmerkungen: [1]übernommen aus Chafouleas et al. (2013). Für die Studie wurde eine Skala von 0% bis 100% eingesetzt (siehe Abbildung 2); [2]übernommen aus Volpe und Briesch (2012). Für die Studie wurde eine Skala von 1 (Verhalten trat nie auf) bis 6 (Verhalten trat während des gesamten Beobachtungszeitraums auf) eingesetzt.

richt) und nicht spezifisch (z.B. „hebt die Hand", „folgt den Anweisungen" oder „sitzt angemessen") vorgegeben wurde.

Valenz der Zielformulierung. Ein weiteres relevantes Kriterium zur Erreichung einer guten Beobachtungsgenauigkeit könnte die Valenz der Zielformulierung sein. Beobachtungsziele können grundsätzlich immer eher ein positives erwünschtes Verhalten oder ein negatives unerwünschtes Verhalten fokussieren. Riley-Tillman et al. (2009) und Chafouleas et al. (2013) untersuchten den Einfluss positiver und negativer Zielformulierungen bei SI-Skalen im Rahmen eines Vergleichs der Messdaten von DBR- und SDO-Ratern. Auch hier wurden N = 145 bzw. N = 113 Studierende gebeten, das Verhalten eines Zielschülers in kurzen Videoclips zu beurteilen. Die Probanden wurden dabei in den Studien zufällig in zwei Experimentalgruppen aufgeteilt, die unterschiedliche Formulierungen (positiv vs. negativ) des Zielverhaltens erhielten. Beide Arbeitsgruppen kamen zu dem Ergebnis, dass eine positive (globale) Zielformulierung bei der Beobachtung der „Teilnahme am Unterricht" zu einer höheren Beobachtungsgenauigkeit führt und eine negative Zielformulierung die Genauigkeit für den Bereich „störendes Verhalten" verbessert. Demnach war es für die Probanden einfacher korrekt zu beurteilen, wie lange sich ein Zielschüler am Unterricht beteiligt hat, als zu beobachten, wie lange sich der gleiche Schüler nicht am Unterricht beteiligt hat. Andersherum war es für die Probanden einfacher, den zeitlichen Anteil des „störenden Verhaltens" in einer Unterrichtssequenz korrekt zu beurteilen als den Anteil „nicht störenden Verhaltens". Uneindeutig waren die Befunde für den Bereich respektvolles Ver-

halten. Hier empfehlen Chafouleas et al. (2013) eine negative Zielformulierung, also den Anteil „nicht respektvollen Verhaltens". Letztendlich kann aus der Befundlage keine Empfehlung in Bezug auf die „positive" oder „negative" Zielexplikation abgeleitet werden. Insgesamt zeigen die einzelnen Arbeiten, dass die Ergebnisse der Studien nicht direkt von einem Beobachtungsbereich auf einen anderen übertragbar sind, sondern für die jeweiligen Beobachtungsbereiche getrennt betrachtet werden müssen.

Länge der Verhaltensstichproben. Riley-Tillman et al. (2011) untersuchten den Einfluss der Länge einer Verhaltensstichprobe auf die Beobachtungsgüte in einem experimentellen Design. Im Rahmen ihrer Studie ließen die Autoren 81 studentische Probanden verschiedene Videoclips in verschiedenen Längen zweimal mit Abstand von einer Woche beurteilen. Das Forschungsteam zeigte im Rahmen dieser Studie, dass die Test-Retest-Reliabilität stark zwischen den beobachteten Situationen und der Beobachtungszeit variierte. Grundsätzlich fiel die Test-Retest-Reliabilität für die zehnminütige Beobachtungssequenzen mit Korrelationen zwischen $r_{tt} = .31$ und $r_{tt} = .56$ etwas geringer aus als für 20-minütige Beobachtungssequenzen, wo die Spannweite zwischen $r_{tt} = .31$ und $r_{tt} = 1.00$ lag (Riley-Tillman et al., 2011). Es ergaben sich zudem Hinweise darauf, dass ein Mittelwert von mehreren Ratings zu einer höheren Reliabilität führen könnte als einzeln vorgenommene Ratings (Riley-Tillman et al., 2011; Volpe & Briesch, 2012). So lagen die Test-Retest-Reliabilitätswerte höher, wenn eine 20-minütige Videosequenz in vier 5-minütige Verhaltensstichproben unterteilt und viermal getrennt mit Hilfe einer DBR-Skala beurteilt wurde. Es zeigt sich also auch hier, dass für die Beurteilung der gleichen Verhaltensstichprobe der Mittelwert aus mehreren einzelnen kurzen Beurteilungen insgesamt ein robusterer Kennwert für eine Verhaltensbeurteilung ist als eine einzelne DBR-Messung über einen längeren Zeitraum. Kritisch muss bei dieser Studie jedoch angemerkt werden, dass die Videos in Test und Retest zu unterschiedlichen Längen geschnitten waren, was zu einer Verzerrung der Ergebnisse geführt haben könnte. Insgesamt sind die hier ermittelten Test-Retest-Reliabilitäten zudem eher unbefriedigend und DBR-Messungen sollten in weiteren Studien differenzierter im Hinblick auf Optimierungspotenziale überprüft werden.

Auswirkungen des Beobachtertrainings. Mit Blick auf den Einsatz von DBR-Instrumenten in der schulischen Praxis stellt sich die Frage, inwieweit Beobachter-Trainings die Messgenauigkeit dieser Methode erhöhen können. Während einige Experimentalstudien schon bei unerfahrenen Studierenden ohne Beobachter-Training zu vergleichsweise befriedigenden Ergebnissen kamen (Kilgus et al., 2012; Riley-Tillman et al., 2011), nahmen vor allem Autorengruppen, die ihre Studien mit kleineren Probandenzahlen durchführten, zumeist ein kurzes Beobachter-Training vor (Briesch et al., 2012; Volpe & Briesch, 2012). Die einzigen Studien, die den Effekt von Beobachterschulungen auf die Messgenauigkeit von DBR systematisch untersuchten stammen von Schlientz et al. (2009) und LeBel et al. (2009). Erstere Autoren prüften in einer Experimentalstudie mit 59 Studierenden, inwieweit ein praktisches Beobachtertraining die Messgenauigkeit bei DBR verbessern konnte. Die eine Hälfte der Studierenden wurde dabei zufällig einer Trainingsgruppe zugeordnet, die eine ca. 23-minütige Einführung zur Verwendung von DBR-Instrumenten erhielt. Am Ende dieses Trainings sollten die Studierenden sechs Videosequenzen mit Hilfe einer DBR-SIS bewerten, wobei den Studierenden direkt nach jeder Beurteilung die Ergebnisse von geschulten Verhaltensbeobachtern zurückgemeldet wurden und sie somit ein unmittelbares korrektives Feedback erhielten. Die andere Hälfte der Probanden erhielt nur eine kurze Einführung. Schlientz et al. (2009) zeigten, dass die Messgenauigkeit der trainierten

Studierenden signifikant höher war als die Beobachtungen der Kontrollgruppe. Die höhere Genauigkeit zeigte sich in einer höheren Interrater-Übereinstimmung und einer geringeren durchschnittlichen Abweichung der trainierten Beurteiler von den Ergebnissen einer direkten systematischen Verhaltensbeobachtung (Schlientz et al., 2009). Zu etwas anderen Ergebnissen kamen LeBel et al. (2009) in ihrer Studie. Die Autoren teilten die Probanden in drei Experimentalgruppen auf, wobei eine Gruppe untrainiert blieb und zwei weitere Gruppen ein unterschiedlich intensives Beobachtertraining erhielten. Die Befunde zeigten, dass ein besonders intensives Training mit Übungsphasen im Vergleich zu einem kurzen einführenden Training ohne Trainingsphasen keine signifikante Verbesserung der Beobachtungsgüte brachte (LeBel et al., 2009). Die Befunde deuten darauf hin, dass möglicherweise eine kurze Schulung im Umgang mit DBR-Skalen ausreicht, um eine bestmögliche Genauigkeit von DBR-Messungen zu erhalten. Zu grundsätzlich vergleichbaren Befunden kamen auch Chafouleas, Kilgus, Riley-Tillman, Jaffery und Harrison (2012) in ihrer Experimentalstudie. Die Autoren zeigten, dass ein intensiveres Training nur dann zu einer höheren Beobachtungsgenauigkeit führt, wenn es sich um weniger eindeutiges Verhalten in einem qualitativ mittleren Ausprägungsgrad handelte, wie z.B. eine qualitativ mittelmäßige Beteiligung am Unterricht (Chafouleas et al., 2012).

Diskussion

Im ersten Teil der Arbeit wurde Direct Behavior Rating als alternative Methode zur Diagnose von Verhaltensentwicklungen in der Schule dargestellt. Dabei wurde DBR als Hybridform aus einer direkten systematischen Verhaltensbeobachtung und einer Verhaltensbeurteilung beschrieben. Im Gegensatz zu einer sehr viel aufwändigeren direkten systematischen Verhaltensbeobachtung sollen die Bobachtungswerte nicht durch das exakte Auszählen einzelner Verhaltensstichproben, sondern, wie bei einer Verhaltensbeurteilung, auf einer mehrstufigen Likert-Skala angegeben werden. Im Unterschied zu einer Verhaltensbeurteilung werden die Messungen nicht einmalig zu einem bestimmten Zeitpunkt vorgenommen, sondern hochfrequent mehrfach am Tag und direkt im Anschluss an die Beobachtungssituation. Durch die hohe Anzahl an Messwerten lassen sich einerseits Entwicklungsverläufe abbilden. Andererseits erhoffen die derzeit tätigen Arbeitsgruppen, dass durch eine intensive Beforschung von DBR als Methode grundsätzliche Ableitungen zur Beobachtungsgüte unabhängig von der beobachtenden Person möglich sind und somit die aufwändige Überprüfung von Interrater-Reliabilität oder Kriteriums-Validität, wie es im Rahmen einer direkten systematischen Verhaltensbeobachtung empfohlen wird, nicht mehr zwingend erforderlich sind. Um die Eignung der Methode für die praktische Diagnostik in Unterricht und Schule besser abschätzen zu können, wurden die bislang vorliegenden Befunde im Rahmen eines Reviews ausgewertet.

Die Ergebnisse des systematischen Reviews geben Hinweise darauf, dass die Beobachtungsgüte von DBR-Messungen im Gegensatz zu SDO-Messungen geringer ist und stärker von der beobachtenden Person abhängt. Es lassen sich jedoch auf der Grundlage der bislang vorliegenden Forschungsarbeiten einige Hinweise zur Erhöhung der Beobachtungsgüte ableiten.

So scheinen Validität und Reliabilität in einem erheblichen Maße von dem zu beobachtenden Zielverhalten abzuhängen. Die Beobachtungsgüte stieg in den vorliegenden Studien an, wenn ein Zielverhalten beurteilt werden sollte, das im schulischen Kontext gut zu beobachten und zu operationalisieren war. Die beste Beobachtungsgenauigkeit lag für die Bereiche „Teilnahme am Unterricht“ und „störendes Verhalten“ vor. Eine weiterhin akzeptable Beobachtungsgüte lag für den Bereich des „respekt-

vollen Verhaltens“ vor. Chafouleas (2011) bezeichnet diese drei Beobachtungsbereiche als die „Big 3“ der Methode DBR.

Bei der Gestaltung der Itemskalen scheinen zwei verschiedene Strategien zu einer Erhöhung der Kriteriums-Validität zu führen. Im Rahmen der ersten Strategie wird die Übereinstimmung mit dem „wahren Verhalten“ durch das Rating mehrerer spezifischer Verhaltensausschnitte erhöht. Dies ist bei Multi-Item-Scales der Fall. Pro Beobachtungseinheit sind somit nicht nur eine, sondern mehrere spezifische Beurteilungen vorzunehmen. Trotz der hier in der Tendenz höheren Stabilität und Übereinstimmung mit der systematischen Verhaltensbeobachtung bleibt zu beachten, dass die Komplexität bzw. Schwierigkeit der Bewertungsaufgabe für den Rater mit jedem zusätzlichen Item steigt.

Die zweite Strategie bezieht sich auf den Einsatz von SI-Skalen. Zur Erhöhung der Beobachtungsgüte sollte das Zielverhalten bei diesem Skalentyp möglichst positiv und global formuliert sein und eher die Anwesenheit eines Verhaltensbereiches und weniger die Abwesenheit eines sehr spezifischen Verhaltensausschnitts definieren. Die verwendeten Skalenbreiten und Verankerungen hatten in den hier zusammengefassten Arbeiten hingegen keinen messbaren Einfluss auf die Beobachtungsgüte.

Die Angaben über die zur Erlangung einer akzeptablen Beobachtungsgüte erforderliche Anzahl an Beurteilungen schwanken stark. Die hier zusammengefassten Studien legen nahe, dass bei weniger als fünf Messzeitpunkten die Messwerte zu stark zwischen den Ratern streuen und somit mehr Varianz durch die Person des Raters aufgeklärt wird als durch die Verhaltensunterschiede der beobachteten Personen. Bei entscheidungsrelevanten Fragestellungen empfehlen einige Autoren sicherheitshalber sogar die Zusammenfassung von bis zu 20 Messwerten. Die vorliegenden Studien deuten darauf hin, dass Validität und Reliabilität von DBRs weiter steigen, wenn die Rater zuvor ein kurzes Training durchlaufen haben. Hinweise auf eine Verbesserung der Beobachtungsgüte durch intensive Trainingsphasen gab es bislang nicht.

Insgesamt stellt sich bei einer Diskussion über die Eignung von DBR für die Diagnostik bei Verhaltensproblemen die Frage nach der Zielsetzung und der Fragestellung ihres Einsatzes. So lassen sich die bislang vorliegenden Befunde dahingehend interpretieren, dass direkte Verhaltensbeurteilungen, die nur auf wenigen Beobachtungen beruhen, auch bei ansonsten optimierten Einsatzbedingungen je nach Beobachtungsziel mit einer erheblichen interpersonalen Varianz versehen sind. Eine zuverlässige statusdiagnostische Einschätzung sollte mindestens auf 20 Messdaten beruhen und bietet auch dann nur einen sehr eng umgrenzten Erkenntnisgewinn, da er sich bei Verwendung einer einzelnen SI-Skala nur auf ein einzelnes Merkmal (z.B. Unterrichtsengagement) bezieht. Der Sinn und der Nutzen von DBR für die Statusdiagnostik wären demnach nicht besonders hoch und die Ökonomie der Methode wäre vor diesem Hintergrund gering.

Bewertet man die Befunde vor dem Hintergrund einer prozessdiagnostischen Zielsetzung verändert sich dieses Bild. So zeigen die Daten zunächst, dass 20 Messzeitpunkte gemeinsam ein solides diagnostisches Bild ergeben. Betrachtet man die Ökonomie und die Geschwindigkeit, mit der eine einzelne DBR-Messung insbesondere bei Verwendung einer SI-Skala erfolgt, lassen sich im Laufe einer Schulwoche viele Messungen vornehmen. Würde eine Lehrkraft zum Beispiel nach jeder Schulstunde eine Beurteilung vornehmen, ließen sich auf dieser Grundlage pro Woche leicht 20-30 Messdaten gewinnen. Auf diese Weise entsteht im Verlauf einzelner Wochen schnell eine solide Datengrundlage für eine prozessdiagnostische Betrachtung eines Verhaltensausschnitts in der Schule.

Die prozessdiagnostische Eignung von DBR-Skalen wird jedoch unter anderem durch die zeitliche Stabilität der Raterurteile begrenzt. Die bislang vorliegenden Be-

funde zeigen, dass die Test-Retest-Reliabilität von direkten Verhaltensbeurteilungen vergleichsweise hoch ist, sofern sie auf mehreren (dort fünf) DBR-Messungen beruht (Riley-Tillman et al., 2011). Allerdings muss hier kritisch angemerkt werden, dass eine Korrelation zwischen einem Test und ihrem Retest nur eine Aussage darüber zulässt, inwieweit die Raterprofile über eine Reihe von Messdaten übereinstimmen. Ein Korrelationsmaß lässt hingegen keine Aussage darüber zu, inwieweit die absoluten Messwerte oder die Mittelwerte von einer Test- und einer Retestreihe tatsächlich übereinstimmen. Letzteres wäre jedoch ein relevantes Kriterium für die Verwendung von DBR-Instrumenten in der Praxis. Insbesondere wenn ein und dieselbe Situation zu zwei Zeitpunkten von einem Rater vergleichbar eingeschätzt wird, sind DBR-Messungen ein guter Indikator für Verhaltensentwicklungen. Bestätigen sich die bislang vorliegenden Befunde in weiteren Studien, bedeutete dies, dass die intrapersonelle Einschätzung eines Verhaltens im Unterricht vergleichsweise stabil sein könnte. Dies wäre eine wichtige Voraussetzung für den prozessdiagnostischen Einsatz von DBR bei der Beurteilung von Verhaltensentwicklungen. Weitere Forschungsarbeiten müssen in diesem Zusammenhang Klarheit schaffen.

Im Gegensatz zu einer statusdiagnostischen Zielsetzung von DBR-Messdaten wäre es bei ihrem prozessdiagnostischen Einsatz zudem weniger relevant, inwieweit die Lehrkraft das Verhalten als „überdurchschnittlich", „unterdurchschnittlich" oder „normal" einschätzt. Entscheidend für einen prozessdiagnostischen Einsatz von DBR ist vor allem die Frage, ob es aus Sicht der beobachtenden Person eine Veränderung des Verhaltens über die Zeit gibt. DBR-Messungen in der Schule können damit insbesondere eingesetzt werden, um die Wirksamkeit pädagogischer Maßnahmen in Schule und Familie auszuwerten. Sie bilden damit das Fundament für evidenzbasierte Entscheidungsprozesse in multiprofessionellen Teams, wie sie beispielsweise im response-to-intervention-Konzept empfohlen werden (Reschley & Bergstrom, 2009). Stellt man die Anzahl der erreichten Punkte pro Woche zum Beispiel als Kurve dar, könnten professionelle Helferinnen und Helfer, aber auch Eltern und ein Schüler oder eine Schülerin selbst ohne großen Aufwand die Verhaltensentwicklung über das Schuljahr erkennen und gemeinsam Rückschlüsse über die Wirksamkeit der eingeleiteten Maßnahmen ziehen. Insbesondere die Auswertung eines Entwicklungsverlaufs als „Response" auf zuvor eingeleitete pädagogische Maßnahmen in Schule oder Umfeld bietet dabei neue Ansatzpunkte für die weitere pädagogische Arbeit. Direct Behavior Rating muss vor diesem Hintergrund also in erster Linie als förderdiagnostisches Instrument betrachtet werden, in deren Mittelpunkt die Wirkung einer Intervention und weniger die Etikettierung eines Schülers oder einer Schülerin steht.

Mit Blick auf die praktische Umsetzung von DBR im Schulalltag stellt sich die Frage, wer die DBR-Messungen tatsächlich durchführen soll. Da DBR auf hochfrequenten Messungen beruht, wäre es auch für die Praxis unerlässlich, dass die Messungen von Personen vorgenommen werden, die täglich bzw. regelmäßig in Kontakt mit den betreffenden Kindern und Jugendlichen kommen. Betrachtet man die Studien in diesem Review unter diesem Aspekt, scheinen Lehrkräfte hier tatsächlich die bevorzugte Anwendergruppe für diese Methode zu sein. Aber auch andere pädagogische Fachkräfte wie Erzieher (z.B. im Kindergarten) oder Sozialpädagogen (z.B. im offenen Ganztag) oder gar die Eltern kommen grundsätzlich als mögliche Anwender dieser Technik in Frage. Insbesondere, da die in diesem Review diskutierten Studien gezeigt haben, dass selbst Laienbeobachter, die eine kurze Schulung erhalten haben, zufriedenstellende Einschätzungen liefern können. Wichtig erscheint jedoch, dass diese Schulungen im Alltag auch tatsächlich stattfinden. Ferner muss auch für den Alltag eine Qualitätssicherung im Hinblick auf DBR diskutiert werden, bei der eine ausrei-

chende Beobachtungsgüte der Rater nicht einfach unterstellt, sondern von Zeit zu Zeit auch geprüft wird.

Einschränkend muss gesagt werden, dass die hier diskutierte Befundlage noch zu gering ist, um die prozessdiagnostische Eignung von DBR seriös einschätzen zu können. Mit Blick auf die erhebliche Bandbreite an Verhaltensauffälligkeiten in der Schule muss geprüft werden, inwieweit DBR tatsächlich für alle Verhaltensausschnitte eine valide und reliable Methode der Prozessdiagnostik ist oder der Einsatz der Technik auf einzelne Verhaltensbereiche beschränkt bleiben sollte. So weisen die Studien darauf hin, dass insbesondere sehr auffällige Verhaltensausprägungen leichter zu beurteilen sind als weniger auffällige. Ferner liegen bislang keine Erkenntnisse zum Einsatz von DBR bei internalisierenden Verhaltensauffälligkeiten vor. Insgesamt bleibt also bei einer Gesamtbetrachtung der Befunde unklar, inwieweit eine gute Test-Retest-Reliabilität auch in anderen Studien mit anderen Zielverhaltensweisen replizierbar ist.

Das Review umfasst viele qualitativ sehr unterschiedliche Einzelbefunde, so dass (neben der generellen Übertragbarkeit auf den deutschen Sprachraum) eine integrierende Betrachtung nur bedingt möglich ist. Einige Parameter sind zudem bisher überhaupt nicht untersucht worden: So fehlen beispielsweise Untersuchungen zur „Schwierigkeit" der diagnostischen Situation. So gibt es inhaltliche, methodische und didaktische Rahmenbedingungen in der Schule (z.B. ein interessanter Unterrichtsstoff oder eine spannende Methode der Lehrkraft), die das Auftreten eines bestimmten Verhaltens wahrscheinlicher machen und andere, die es eher unwahrscheinlicher machen. Völlig ungeklärt ist, wie Verhaltensentwicklungen vor dem Hintergrund schwankender Situationsschwierigkeiten interpretiert werden sollen.

Mit Blick auf die vorliegenden Forschungsarbeiten erscheint auch die Interpretation von Befunden, die auf sehr kleinen Stichproben oder wenigen zu bewertenden Verhaltenssequenzen basierten, als schwierig. Hier sind sicherlich Replikationen unter systematischer Bedingungsvariation notwendig. Ebenso fehlen weitere systematische Studien zur Formulierung des Beobachtungsziels, da es durchaus denkbar ist, dass sprachliche Unterschiede bei der Formulierung des Beobachtungsziels auch zu einer Veränderung der Beobachtungsgüte führen. Eine weitere Forschungslücke betrifft die Frage, inwieweit professionelle Beobachter zu einer anderen Einschätzung des Zielverhaltens gelangen als Laienbeobachter. So ist ein Großteil der hier diskutierten Studien mit Hilfe von Studierenden ohne Berufserfahrung entstanden. Generalisierungsstudien bezogen zwar unterschiedliche Berufsgruppen (z.B. Laienbeobachter und Erzieher oder Lehrer) mit ein, hier waren die Anzahl der Beobachter jedoch oft gering.

Unabhängig von der Variation der „unabhängigen Variablen", die einen Einfluss auf die „Qualität" einer DBR-Messreihe haben, muss weitere Forschung die Frage klären, wie eigentlich die „Qualität" dieser spezifischen Form der Messung, bei der im Grunde der Beobachter das „Messinstrument" ist, definiert werden kann. Zumal hier normale testtheoretische Gütekriterien zur Bewertung der Beobachtungsgüte streng genommen nicht ohne weiteres übernommen werden können. Ebenso muss in einem darauf folgenden Schritt die Frage beantwortet werden, mit welchen Verfahren die „Qualität" der DBR statistisch überprüft werden kann.

Trotz dieser zahlreichen Probleme erscheint es den Autoren wichtig, DBR als Methode zur Verhaltensverlaufsdiagnostik in der Schule zu optimieren. In der Vergangenheit haben zahlreiche Studien gezeigt, dass ein Lern- und Entwicklungsfeedback die Effizienz von pädagogischen Interventionen deutlich steigern können (Hattie, 2008; Hattie & Timperley, 2007). Insofern wäre es sowohl für eine verbesserte Förderung von Kindern mit Verhaltensproblemen in der Schule als auch für den effizienten Einsatz schulischer Ressourcen wichtig, valide und

reliable Instrumente zur Verhaltensverlaufsdiagnostik (weiter) zu entwickeln. Der hier diskutierte Ansatz des Direct Behavior Ratings erscheint vor dem Hintergrund der hier zusammengefassten Befunde hierfür als sinnvoller Ausgangspunkt. Allerdings müssen die hier diskutierten offenen Fragen für einen flächendeckenden Einsatz von DBR besser und umfassender beantwortet werden, als es die derzeitige Forschungslage erlaubt. Insbesondere müssen weitere Forschungsarbeiten zeigen, für welchen konkreten Einsatzbereich und für welche pädagogischen Fragestellungen DBR tatsächlich sinnvoll eingesetzt werden kann.

Literaturverzeichnis

Amelang, M. & Schmidt-Atzert, L. (Hrsg.). (2006). *Psychologische Diagnostik und Intervention* (4., Vollst. Uberarb., Aktualisierte Aufl). Berlin: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler* (4., überarbeitete Aufl.). Heidelberg: Springer.

Briesch, A. M. (2010). Generalizability and dependability of behavioral assessment methods: A comparison of systematic direct observation and direct behavior ratings. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 70* (9-A), 3338.

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* (3), 408–421.

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* (2), 127–133.

Chafouleas, S. M. (2011). Direct Behavior Rating: A Review of the Ussues an Research in Its Development. *Education And Treatment of Children, 34* (4), 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* (3), 219–246. Verfügbar unter http://dx.doi.org/10.1016/j.jsp.2010.02.001

Chafouleas, S. M., Christ, T. J. & Riley-Tillman, T. C. (2009). Generalizability of Scaling Gradients on Direct Behavior Ratings. *Educational and Psychological Measurement, 69* (1), 157–173.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M. & Chanese, Julie A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36* (1), 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* (1), 39–53.

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* (3), 317–334. Verfügbar unter http://dx.doi.org/10.1016/j.jsp.2011.11.007

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* (4), 195–200. Verfügbar unter http://dx.doi.org/10.1177/1534508409340391

Chafouleas, S. M., Riley-Tillman, T. C. & Sassu, K. A. (2002). Good, bad or in-between: How does the Daily Report Card rate? *Psychology in the Schools, 39,* 157–169.

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* (4), 201–213.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M. & Boice, C. H. (2010a). Direct Behavior Rating (DBR): Generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70* (5), 825–843. Verfügbar unter http://dx.doi.org/10.1177/0013164410366695

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

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. & Jaffery, R. (2011). Direct behavior rating: An evaluation of alternate definitions to assess classroom behaviors. *School Psychology Review, 40* (2), 181–199.

Cone, J. D. (1988). Psychometric considerations. In A. S. Bellack & M. Hersen (Hrsg.), *Behavioral assessment. A practical handbook* (Pergamon general psychology series, Bd. 65, 3rd ed, S. 38–68). New York: Pergamon Press.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37* (3), 184–192.

Diehl, K. & Hartke, B. (2012). *IEL-1. Inventar zur Erfassung der Lesekompetenz im 1. Schuljahr : ein curriculumbasiertes Verfahren zur Abbildung des Lernfortschritts*. Göttingen: Hogrefe.

Eisend, M. (2007). Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung, WiSt, 36, 494-500.

Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53* (3), 199–208.

Fydrich, T. (2012). Diagnostik in der Klinischen Psychologie. In L. Schmidt-Atzert (Hrsg.), *Psychologische Diagnostik* (Springer-Lehrbuch, 5., vollständig überarbeitete und erweiterte Auflage, S. 503–537). Berlin, Heidelberg: Springer Berlin Heidelberg.

Grosche, M. & Huber, C. (in Druck). Die passgenaue Abstimmung zwischen Lernbedürfnissen und Lernmethoden durch das Konzept Response-to-Intervention (RTI). In K. Reich & D. Asselhoven (Hrsg.), *Eine (inklusive) Schule (neu) erfinden.* Frankfurt a.M.: Campus.

Hattie, J. (2008). *Visible learning. A synthesis of meta-analyses relating to achievement*. London: Routledge.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77* (1), 81–112.

Hintze, J. M. & Matthews, W. J. (2004). The Generalizability of Systematic Direct Observations Across Time and Setting: A Preliminary Investigation of the Psychometrics of Behavioral Observation. *School Psychology Review, 33* (2), 258–270.

Hosp, M. K., Hosp, J. L. & Howell, K. W. (2007). *The ABCs of CBM. A practical guide to curriculum-based measurement*. New York: Guilford Press.

Huber, C. (2013). Token-Systeme – Kriterien wirksamer schulischer Verhaltensmodifikation. In H. Bartnitzky, U. Hecker & M. Lassek (Hrsg.), *Individuell fördern - Kompetenzen stärken. Ab Klasse 3* (Beiträge zur Reform der Grundschule, Bd. 135, neue Ausg, S. 47–54). Frankfurt am Main: Grundschulverband - Arbeitskreis Grundschule.

Huber, C. & Grosche, M. (2012). Das response-to-intervention Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift fur Heilpadagogik, 63* (8), 312–322.

Hussy, W., Schreier, M. & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bache-*

lor. Mit 23 Tabellen (Springer-Lehrbuch, 2., überarb. Aufl). Berlin: Springer.

Johnson, M. E., Fuchs, D. & McKnight, M. A. (2006). *Responsiveness to Intervention (RTI): How to Do It.,* National Research Center on Learning Disabilities. Verfügbar unter http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED496979

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* (1), 41–50. Verfügbar unter http://dx.doi.org/10.1037/a0027150

Kilgus, S. P., Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J. & Welsh, M. E. (2014). Direct behavior rating as a school-based behavior universal screener: Replication across sites. *Journal of School Psychology, 52* (1), 63–82.

Kivirauma, J. & Ruoho, K. (2007). Excellence through special education? Lessons from the Finnish school reform. *International review of education, 53* (3), 283–302.

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

Kooij, R. van der. (2004). Förderdiagnostik als Prozess. In P. Jogschies (Hrsg.), *Neue Entwicklungen in der Förderdiagnostik. Grundlagen und praktische Umsetzungen* (1. Aufl.). Weinheim: Beltz.

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

Müller, C. M. & Hartmann, E. (im Druck). *Lernfortschrittsdiagnostik: Grundrechenarten. 120 Drei-Minuten-Tests für den inklusiven Mathematikunterricht - ZR 1-100*. Hamburg: Persen Verlag.

Nelson, R. O., Hay, L. R. & Hay, W. M. (1977). Comments on Cone's 'The relevance of reliability and validity for behavioral assessment.'. *Behavior Therapy, 8* (3), 427–430. Verfügbar unter http://dx.doi.org/10.1016/S0005-7894(77)80078-6

Reschley, D. & Bergstrom, M. K. (2009). Response to Intervention. In T. B. Gutkin & C. R. Reynolds (Hrsg.), *The handbook of school psychology* (4. Aufl., S. 434–460). Hoboken (N.J.): J. Wiley.

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., Briesch, A. M. & LeBel, T. J. (2009). The impact of wording and behavioral specifity on accuracy of Direct Behavior Rating (DBRs). *School Psychology Quarterly, 24,* 1–12.

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

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

Riley-Tillman, T. 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* (2), 119–128. Verfügbar unterhttp://dx.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 accuracy of Direct Behavior Ratings (DBR). *School Psychology Quarterly, 24* (2), 73–83. Verfügbar unter http://dx.doi.org/10.1037/a0016255

Schmidt-Atzert, L. (2010a). Beobachtungsfehler und Beobachtungsverzerrungen. In K. Westhoff, C. Hagemeister, M. Kersting, F. Lang, H. Moosbrugger, G. Reimann et al. (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN*

33430 (S. 74–78). Lengerich: Pabst Science Publishers.

Schmidt-Atzert, L. (2010b). Ratingverfahren. In K. Westhoff, C. Hagemeister, M. Kersting, F. Lang, H. Moosbrugger, G. Reimann et al. (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (S. 61–68). Lengerich: Pabst Science Publishers.

Schmidt-Atzert, L. (Hrsg.). (2012). *Psychologische Diagnostik* (Springer-Lehrbuch, 5., vollständig überarbeitete und erweiterte Auflage). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sliwka, A. (2010). Chancengerechtigkeit und Exzellenz: Was das deutsche Schulsystem von Kanada lernen kann. In K. Hurrelmann, u.a, M. Speich & D. Deißner (Hrsg.), *Transmission 03. Herkunft und Chance. Wege zu mehr Bildungsgerechtigkeit an Deutschlands Schulen* (S. 38-58). Düsseldorf: Vodafone-Stiftung Deutschland.

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

Stemmler, G. (2010). Beobachtung: Begriff und Verständnis. In K. Westhoff, C. Hagemeister, M. Kersting, F. Lang, H. Moosbrugger, G. Reimann et al. (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (S. 37–42). Lengerich: Pabst Science Publishers.

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42* (2), 111–122.

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

Volpe, R. J. & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41* (3), 246–261.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. *Heilpädagogische Forschung* (2), 62–79.

Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G. et al. (Hrsg.). (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst Science Publishers.

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Bd. 12, 1. Aufl, S. 281–308). Göttingen, Niedersachs: Hogrefe Verlag.

Wittchen, H.-U. & Hoyer, J. (Hrsg.). (2011). *Klinische Psychologie & Psychotherapie* (2., überarbeitete und erweiterte Auflage). Berlin: Springer-Verlag Berlin Heidelberg.

Ziegler, M. & Bühner, M. (Hrsg.). (2012). *Grundlagen der Psychologischen Diagnostik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Prof. Dr. Christian Huber
Universität Potsdam
Professur für Inklusionspädagogik
Förderschwerpunkt emotionale und soziale Entwicklung
Karl-Liebknecht-Straße 24 - 25
14476 Potsdam
chhuber1@uni-potsdam.de

Erstmalig eingereicht: 14.07.2014
Überarbeitung eingereicht: 23.09.2014
Angenommen: 30.09.2014