Empirische Sonderpädagogik, 2017, Nr. 1, S. 19-35
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Monitoring der sozial-emotionalen Situation von Grundschülerinnen und Grundschülern – Ist der SDQ ein geeignetes Verfahren?

Stefan Voß[1] *& Markus Gebhardt*[2]

[1] *Universität Rostock*
[2] *Technische Universität Dortmund*

Zusammenfassung

Der Strength and Difficulties Questionnaire (SDQ) ist ein in Forschung und Praxis etabliertes Screeninginstrument zur Diagnostik von Verhaltensstörungen. Im vorliegenden Artikel wird neben der statusdiagnostischen Eignung der Lehrkraftversion des SDQ über die vier Jahre der Grundschule zu je einem Messzeitpunkt hinweg die Einsetzbarkeit des Verfahrens als Instrument zur Verlaufsmessung geprüft. Um die Skalierung über die vier Messzeitpunkte zu untersuchen, wird die Problemwertskala des SDQ mittels des Raschmodells an einer Schuljahreskohorte einer deutschen Kleinstadt im Längsschnitt analysiert.
Die Ergebnisse zeigen, dass der SDQ Gesamtproblemwert mit wenigen Ausnahmen messinvariant über die Zeit ist. Im Raschmodell über vier Messzeitpunkte weicht ein Item vom eindimensionalen Modell ab und wird für weitere Berechnungen entfernt. Anhand von Mehrebenenregressionen erkennt man, dass die Personenwerte der Grundschulkinder im Gesamtproblemwert über die Schuljahre leicht ansteigen.
Eine Weiterentwicklung des SDQ Richtung eines Instrumentes zur Verlaufsmessung ist möglich, hierzu sollten weitere Items im schwierigen und leichten Bereich konstruiert werden.

Schlüsselwörter: SDQ, Gesamtproblemwert, Verlaufsmessung, Verhaltensentwicklung, Raschmodell

Monitoring of the social emotional situation of elementary school students – Is the SDQ a suitable instrument?

Abstract

The Strength and Difficulties Questionnaire (SDQ) is an established screening tool in research and practice for the purpose of diagnosing behavioral disorders. In this article the teacher version of the SDQ is analyzed in light of its status diagnostic suitability. Longitudinal data of a school year cohort from a small German town was collected to investigate if the total difficulties score over time meets the requirements of the Rasch model and thus if it's appropriate for progress monitoring purposes too.
The results support measurement equivalence of the SDQ total difficulties score over time but with few exceptions. Only one item failed the criteria of the one-dimensional Rasch model and had to be removed from further calculations. Multilevel regressions indicate a slight increase of the total difficulties score over the school years.

A further development of the SDQ towards a progress monitoring measurement is possible. For this purpose, additional items should be constructed to reach a better targeting of the instrument.

Key words: SDQ, total difficulties score, monitoring, course measurement, behavioral development, Rasch model

Verhaltensstörungen im Grundschulalter

Für Kinder mit Verhaltensstörungen sollte das protektive Potential frühzeitig einsetzender Präventionsmaßnahmen (u. a. Beelmann, 2008; Beelmann & Lösel, 2007; Brezinka, 2003; Wiedebusch & Petermann, 2011) genutzt werden, um ungünstige persönliche und schulische Entwicklungen abzumildern bzw. zu verhindern (z. B. Frostad & Pijl, 2007; Huber, 2006; Linderkamp & Grünke, 2007; Reef, Diamantopoulou, van Meurs, Verhulst & van der Ende, 2011; Steinhausen, 2010; Wiedebusch & Petermann, 2011). Nationale sowie internationale Prävalenzstudien deuten darauf hin, dass zwischen 10 % und 20 % aller Kinder und Jugendlichen klinische bzw. quasiklinische Verhaltensstörungen aufzeigen (Costello, Mustillo, Erkanli, Keeler & Angold, 2003; Ihle & Esser, 2008; Petermann, 2005). Neben diesem hohen Anteil von Verhaltensstörungen im Kindes- und Jugendalter spricht auch der Fakt, dass diese Problematiken mit einem gesteigerten Risiko einhergehen, sich im Verlauf der Entwicklung zu manifestieren bzw. weitere Beeinträchtigungen zu entwickeln (Beelmann & Raabe, 2007; Ihle & Esser, 2008) für die Notwendigkeit präventiven Handelns.

Da a) viele persistierende psychiatrische Störungen ihren Ursprung in der Kindheit bzw. der Adoleszenz zu haben scheinen (z. B. Costello, Egger & Angold, 2005), b) der geringe Anteil betroffener Kinder, die tatsächlich Behandlung erhalten (Costello et al., 2005; Petermann, 2005), u. a. für eine mangelnde Erkennungsgüte in der Praxis sprechen und c) Verhaltensstörungen im Kindes- und Jugendalter häufig mit deutlichen Beeinträchtigungen im schulischen oder sozialen Umfeld assoziiert sind (Huber, 2006; Prince et al., 2007; Reef et al., 2011), ist der Bedarf an Instrumenten groß, die das frühzeitige Erkennen emotional-sozialer Störungen und damit auch den zeitnahen Einsatz entsprechender Interventionsmöglichkeiten ermöglichen. Neben dem punktuellen Einsatz von Screeninginstrumenten zur Identifikation von Verhaltensschwierigkeiten gibt es im schulischen Bereich den Ansatz, Verhalten im Längsschnitt zu messen und Lehrkräften eine Rückmeldung hinsichtlich der eingesetzten pädagogischen Maßnahmen zu ermöglichen. Eine derartige formative Evaluation des Verhaltens hat sich als effektiv erwiesen (Volpe & Fabiano, 2013), jedoch mangelt es an Instrumenten, die ein Monitoring des Verhaltens über die Zeit reliabel ermöglichen. Daher ist zu prüfen, inwieweit bestehende Verfahren dies bereits zulassen.

Verlaufsdiagnostik als Element schulischer Prävention

Erfolgreiche schulische Prävention ist an verschiedene Anforderungen geknüpft. So erfordert sie zum einen universelles Wissen der schulischen Akteure, d. h. allgemeines Wissen über Entwicklungsprozesse, zentrale Meilensteine in der Entwicklung, Einflussfaktoren auf das Lernen und Maßnahmen für einen guten Unterricht (Hartke, 2005). Zum anderen ist darüber hinaus spezifisches Wissen von Nöten, das sich auf die Art und das Ausmaß verschiedener Störungen, deren Ursachen sowie Verlauf ohne Intervention, verschiedene Handlungsmöglichkeiten und deren Zielgruppen bezieht. Als ein ebenfalls zentrales Element schuli-

scher Prävention postuliert Hartke (2005) darüber hinaus das Monitoring des betroffenen Problemfeldes über die Zeit. Monitoring meint hierbei eine fortlaufende Beobachtung und Dokumentation der Entwicklung mit dem Ziel der Adaption der eingesetzten Maßnahmen, sofern avisierte Entwicklungsergebnisse ausbleiben. Damit verbunden sind verschiedenste Methoden mit stark variierenden Graden an Strukturierung bzw. Systematik sowie Standardisierung (Bell & Cowie, 2001). Allen gemein ist, dass durch das wiederholte Erheben von Schülerdaten Entwicklungsverläufe, je nach Frequenz der Datenerhebung mehr oder weniger kurzfristig, abgebildet und verfolgt werden können, auf deren Grundlage Feedback für die Lehrperson aber auch die Kinder selbst abgeleitet werden kann. Dabei wird vor allem die individuelle Bezugsnorm als Vergleichsmaßstab zur Abschätzung von Erfolgen herangezogen, also die eigene Entwicklung im Vergleich zu einem früheren Zeitpunkt anstatt des sozialen Vergleichs mit anderen Schülerinnen und Schülern (Rheinberg, 2001).

Da Monitoring-Verfahren regelmäßig im schulischen Alltag eingesetzt werden sollen, ist die Gewährleistung der diagnostischen Nebengütekriterien der Nützlichkeit, Praktikabilität sowie Ökonomie von entscheidender Bedeutung. Den aktuellen Diskurs zur Verlaufsdiagnostik zusammenfassend, lassen sich überdies weitere Forderungen an verlaufsdiagnostische Verfahren stellen, welche in erster Linie auf ihre psychometrische Güte bezogen sind (u. a. Fuchs, 2004; Voß, 2014; Voß, Sikora & Hartke, 2017; Wilbert, 2014; Wilbert & Linnemann, 2011):

- Das Instrument muss den psychometrischen Eigenschaften der Statusdiagnostik entsprechen. Die Güte eines Verfahrens wird zu einem festen Zeitpunkt geprüft. Es wird hierzu die Höhe der Objektivität, Reliabilität und Validität geschätzt.
- Das Instrument muss den psychometrischen Eigenschaften der Verlaufsmessung genügen und auch änderungssensibel den Entwicklungsverlauf darstellen.
- Das Instrument muss ökonomisch in der schulischen Praxis einsetzbar sein und den schulischen Unterricht positiv beeinflussen. Diese Anforderung lässt sich nur in quasiexperimentaler Forschung im Feld nachprüfen.

Da der SDQ hinsichtlich seiner statusdiagnostischen Eignung bereits hinlänglich untersucht wurde, steht insbesondere der zweite Punkt im Fokus dieses Artikels. Drei notwendige Voraussetzungen müssen dabei gelten, damit die Skalierung einer Lernverlaufsdiagnostik angemessen ist (Wilbert, 2014). Die erste Voraussetzung ist, dass auch bei wiederholten Messungen stets dasselbe homogene Konstrukt gemessen wird. Weiterhin wird vorausgesetzt, dass der wiederholt eingesetzte Test jeweils die gleiche Schwierigkeit aufweist und faire Schätzungen der untersuchten Schülerinnen und Schüler ermöglicht. Schließlich ist zu zeigen, dass die Tests änderungssensibel sind und der Verlauf aller gemessenen Schülerdaten dargestellt werden kann (Klauer, 2014). Um diese drei Voraussetzungen zu erfüllen, schlagen Wilbert und Linnemann (2011) explizit für die Verlaufsdiagnostik eine Skalierung nach der Item Response Theory (IRT) vor. Nach dieser Theorie wird die latente Personeneigenschaft bei der Auswertung der Tests berücksichtigt. Dies geschieht einerseits durch die Ausprägung der Person auf der latenten Eigenschaft (Personenparameter) und andererseits anhand der Schwierigkeit der Aufgabe (Itemschwierigkeit). Die Wahrscheinlichkeit der Lösung einer Testaufgabe steht mit den beiden Parametern in einer psychologisch plausiblen probabilistischen Beziehung (Rost, 2004). Für dieses Raschmodell sind notwendige Voraussetzungen zur Modellgültigkeit, dass die Eindimensionalität der Skala und die stichprobeninvariante Anordnung der Items nach ihrer Schwierigkeit gegeben sind. Erst wenn diese Annahmen gelten, ist der Sum-

menwert aussagekräftig hinsichtlich des Antwortverhaltens der getesteten Personen. Damit die Veränderung der Summenwerte auf eine Veränderung des untersuchten Merkmals zurückgeführt werden kann, müssen die zu den einzelnen Messzeitpunkten eingesetzten Tests nicht nur dasselbe Konstrukt erfassen, sondern zudem über die Zeit messinvariant sein (Gebhardt, Heine, Zeuch & Förster, 2015; Klauer, 2014).

Lernverlaufsmessung im Unterschied zur Verhaltensverlaufsmessung

Diese Verlaufsmessung ist in Deutschland vor allem im Bereich der akademischen Leistungsdiagnostik bekannt (u. a. Klauer, 2006; Hasselhorn, Schneider & Trautwein, 2014) und spielt jüngst auch im Bereich der emotional-sozialen Entwicklung eine markante Rolle (Casale, Hennemann, Huber & Grosche, 2015; Wiedebusch & Petermann, 2011). Während hinsichtlich der Lernverlaufsmessung das jeweils unterrichtete Curriculum die entscheidenden Inhalte vorgibt, bezüglich derer alle Schülerinnen und Schüler angehalten sind, sich über die Zeit zu verbessern, sind zu erreichende Ziele im Bereich des Verhaltens nicht schulorganisatorisch geregelt, sondern ergeben sich in erster Linie kontextuell-situativ. Ob ein Verhalten angemessen oder als störend erlebt wird, kann nur jeweils subjektiv, bezogen auf spezifische schulische Situationen sowie vor dem Hintergrund von unterschiedlichen Variablen bezüglich der Klassensituation und Lehrperson entschieden werden. Im Rahmen der Schule ist vor allem das Verhalten von Relevanz, welches in direkter Verbindung zum Lernerfolg der Kinder steht. Daher geht es vor allem um die Erfassung des Arbeits- und Sozialverhaltens in der Klasse. Ziel ist, möglichst frühzeitig negative Verhaltenstendenzen auszumachen, um adäquate Maßnahmen einzuleiten, die einer ungünstigen Entwicklung entgegenwirken. Die besondere Betonung der frühen Förderung emotional-sozialer Kompetenzen zur Prävention zukünftiger manifester und damit meist nur schwer als auch kostenintensiv handhabbarer Störungen ist bereits vielfach im Bereich präventionsbezogener Forschung diskutiert und akzeptiert (u. a. Beelmann, 2008; Beelmann & Lösel, 2007; Brezinka, 2003; Garner, 2010; Wiedebusch & Petermann, 2011). Hierzu ist ein regelmäßiger Einsatz von Instrumenten zur Status- sowie Verlaufsdiagnostik angezeigt (Wiedebusch & Petermann, 2011). Aktuell mangelt es an Instrumenten, die ein Monitoring des Verhaltens über die Zeit reliabel ermöglichen. Es erscheint sinnvoll, verfügbare Verfahren dahingehend zu prüfen.

Der Strength and Difficulties Questionnaire

Zur Diagnostik von Verhaltensstörungen wird eine multimodale sowie multiinformante Vorgehensweise empfohlen (Amelang & Zilinski, 2004; Döpfner & Petermann, 2008). Neben standardisierten Interviews, systematischen Verhaltensbeobachtungen oder psychologischen Testverfahren sind Rating-Verfahren zur Beurteilung des Verhaltens gängig. Da sie in der Regel ökonomisch durchführbar sind, haben Rating-Verfahren großen Zuspruch. Beim Verhaltens-Rating wird ein Verhalten nicht direkt erfasst, sondern retrospektiv, in der Regel über einen festgesetzten Zeitraum extern beurteilt (Döpfner & Petermann, 2008). Zwar wird stellenweise die Verzerrung der Ergebnisse durch sozial erwünschtes Antwortverhalten moniert (Beelmann & Raabe, 2007), dennoch weisen Ratings in der Regel eine hohe Objektivität auf.

Es gibt verschiedene Ratingverfahren zur Einschätzung der emotional-sozialen Situation von Kindern und Jugendlichen, von welchen die Child Behavior Checklist (CBCL; Achenbach, 1991) als die wohl bekannteste wie auch bedeutsamste angesehen wird (Stone, Otten, Engels, Vermulst & Janssens, 2010). Neben der CBCL hat jedoch der SDQ (Goodman, 1997, 2001) in den letzten Jahrzehnten zunehmend an Be-

deutung gewonnen. Dies liegt vor allem darin begründet, dass er mit vergleichsweise wenigen Items Aussagen hinsichtlich des Verhaltens von Kindern und Jugendlichen im Alter von vier bis 16 Jahren zu ermöglichen versucht. Er wurde als Screeninginstrument entwickelt und besteht aus fünf Dimensionen. Jede der Dimensionen *Emotionale Probleme, Verhaltensprobleme, Hyperaktivität, Verhaltensprobleme mit Gleichaltrigen* sowie *Prosoziales Verhalten* besteht aus fünf Items. Der Anwender bzw. die Anwenderin bewertet rückwirkend für die vorangegangenen sechs Monate, das jeweilige Item durch Auswählen einer der Kategorien „nicht zutreffend", „teilweise zutreffend" oder „eindeutig zutreffend".

Die fünf Dimensionen sind in den verschiedenen Ländernormierungen mittels explorativer und konfirmatorischer Faktorenanalysen bestätigt worden (u. a. Koglin, Barquero, Mayer, Scheithauer & Petermann, 2007; Lohbeck, Schultheiß, Petermann & Petermann, 2015). Es gibt aber auch eine Studie, welche zeigt, dass die Aufteilung in die drei Dimensionen *internalisierendes* und *externalisierendes* sowie *prosoziales Verhalten* ebenfalls möglich ist (Goodman, Lamping & Ploubidis, 2010). Des Weiteren hat sich eine Annahme eines Bi-Faktormodells als reliabel erwiesen, welches neben den 5 Dimensionen einen generellen Problemfaktor sowie einen Faktor zum prosozialen Verhalten zugrunde legt (Kóbor, Takács & Urbán, 2013).

Als wichtigster Wert wird der Gesamtproblemwert des SDQ angesehen, der sich als Summe aus den Dimensionen *Emotionale Probleme, Verhaltensprobleme, Hyperaktivität* und *Verhaltensprobleme mit Gleichaltrigen* ergibt. Für diesen Wert liegen internationale Normen vor, die eine Klassifikation in eine der Kategorien „normal", „grenzwertig" oder „auffällig" erlauben.

Neben einer Version für Lehrkräfte existieren auch Auskunftsbögen für Eltern oder die betroffenen Kinder selbst, bei denen geringfügige Adaptionen vorgenommen wurden.

Während die CBCL ein sehr umfassendes Bild über Störungen im emotional-sozialen Bereich eines Kindes oder Jugendlichen ergibt, ist der SDQ als ein Screeninginstrument zur Identifikation von Problemfeldern als auch Ressourcen in diesem Feld zu verstehen, welchem im Verdachtsfall differenzierte Diagnoseprozesse nachgeschaltet werden sollten, um so zu einem umfassenderen sowie reliableren Bild zu gelangen.

Fragestellung

Der SDQ ist ein sehr verbreitetes Sreeninginstrument, welches in den verschiedenen Klassenstufen der Grundschule häufig verwendet wird und an dessen Gesamtproblemwert pädagogische und schulrelevante Entscheidungen gefällt werden. Das Ziel dieses Beitrages ist es, zu analysieren, inwieweit sich der SDQ zur Verlaufsmessung eignet. Anhand einer deutschen Gesamtstichprobe (kleinstädtisch) im Längsschnitt von Klasse 1 bis 4 sollen die psychometrischen Kriterien des SDQ Gesamtproblemwertes dargestellt werden. Es ergeben sich folgende Fragen:

1. Wie fallen die interne Konsistenz und die Interkorrelation des Gesamtproblemwertes zu den Klassenstufen 1 bis 4 aus?
2. Kann der Gesamtproblemwert messinvariant über vier Messzeitpunkte erfasst werden und entspricht er den Anforderungen des Raschmodells?
3. Verändert sich der Gesamtproblemwert über die Zeit und welchen Einfluss hat hierbei das Geschlecht?

Methode

Stichprobe

Zur Untersuchung der Fragestellung wurden Lehrereinschätzungen anhand des SDQ für einen gesamten Einschulungsjahrgang über den Zeitraum der ersten vier Jahre in der Grundschule, jeweils zum Ende des

Schuljahres erhoben. Der SDQ wurde zu allen Messzeitpunkten von jeweils der Grundschullehrkraft ausgefüllt, welche als Klassenlehrkraft in der Klasse unterrichtete. Insgesamt handelte es sich um 17 Personen. Der Stichprobenumfang der Kinder über die Messzeitpunkte variiert (Gründe: Zu- und Wegzüge, Umschulungen von Kindern, vereinzelte Elternverweigerungen, nicht auswertbare Daten aufgrund fehlender Angaben). Die deutliche Zunahme des Stichprobenumfangs vom ersten zum zweiten Erhebungszeitpunkt erklärt sich durch einen Datenausfall in drei Klassen. Vergleicht man jedoch die Werte für diese Kinder mit denen der Gesamtstichprobe zu den nachfolgenden Testzeitpunkten, ergeben sich keinerlei Abweichungen, sodass hier keine systematischen Verzerrungen zu erwarten sind.

Zur Übersicht über die Stichprobe sind die Angaben zu der untersuchten Kohorte sowie Mittelwerte in den Hauptskalen des SDQ Tabelle 1 zu entnehmen.

Die empirischen Richtwerte der Normierung für den SDQ weisen 80 % als verhaltensunauffällige, 10 % als grenzwertige und 10 % als auffällige Personen aus (Goodman, 2001; Koglin et al., 2007). Die hier vorliegenden Verteilungen ähneln der deutschen Norm des SDQ (Koglin, et al, 2007). Dort wird der höchste Wert der Kategorie „normal“ mit 13 und der Kategorie „grenzwertig“ mit 17 Punkten im Gesamtproblemwert angegeben, Werte darüber gelten als „auffällig“. In der hier untersuchten Stichprobe liegt die obere Grenze bei 11 (MZP 1), 12 (MZP 2 und 3) bzw. 14 Punkten (MZP 4) für die Kategorie „normal“. Der Wert 15 Punkte markiert zum ersten Messzeitpunkt die obere Grenze der Kategorie „grenzwertig“, für die Messzeitpunkte 2 und 3 ist dies der Wert 16 Punkte, für den vierten Messzeitpunkt 18 Punkte.

Vorgehen

Die Analysen werden mit dem Statistikprogramm R (R Core Team, 2013) mithilfe des Pakets *pairwise* (Heine, 2014) durchgeführt. Hierbei wird die Methode der expliziten Berechnung der Itemparameter im Raschmodell durch den paarweisen Itemvergleich (Choppin, 1968; Wright & Masters, 1982) angewendet. Diese Methode eignet sich insbesondere zur Bestimmung der stichprobeninvarianten Itemparameter für die Kalibrierung eines gegebenen Itempools (Choppin, 1968). Der pairwise-Schätzer eignet sich auch bei kleinen Stichproben oder Datensätzen mit fehlenden Werten (Wright & Masters, 1982; Heine & Tarnai, 2015), wie es in der vorliegenden Stichprobe der Fall ist.

Zuerst wird die Messinvarianz über die vier Zeitpunkte mittels des grafischen Modelltests geprüft. Dann werden die Itemparameter über alle vier Messzeitpunkte berechnet (Rost, 2004) und die Passung des Modells zu allen Messzeitpunkten mittels Mean-Square-Fit-Statistiken (*Infit* und *Outfit*) bestimmt.

Die Personenparameter werden für die jeweiligen Messzeitpunkte mittels der

Tabelle 1: Deskriptive Angaben zur Stichprobe

	N	**Anteil Mädchen *N* (%)**	**Alter *M* (*SD*)**	**SDQ Gesamtproblemwert *M* (*SD*)**	**SDQ Prosoziales verhalten *M* (*SD*)**
Klasse 1	289	159 (55.0)	7;7 (0;3)	6.92 (5.84)	8.37 (2.82)
Klasse 2	342	179 (52.3)	8;8 (0;4)	7.10 (6.46)	8.00 (2.02)
Klasse 3	370	189 (51.1)	9;8 (0;4)	7.02 (6.26)	8.25 (1.94)
Klasse 4	375	195 (52.0)	10;8 (0;4)	8.51 (7.10)	7.48 (2.39)

Weighted-Maximum-Likelihood-Methode (WLE; Warm, 1989) geschätzt. Für die gemeinsamen Itemparameter werden jeweils die punktbiserialen Korrelationen mit dem Skalenwert (WLE-Schätzer) als Trennschärfe für den jeweiligen Messzeitpunkt berichtet.

Die Analyse der Verläufe über die Zeit erfolgt anhand eines hierarchisch-linearen Modells (HLM, Bryk & Raudenbush, 1992; Level 1 Zeitebene, Level 2 Schülerebene), da dieses eine genauere Schätzung im Umgang mit Missings erzielt, als es bspw. eine ANOVA mit Messwiederholung ermöglicht. Im Rahmen der ersten Ebene wird die Entwicklung des SDQ über die vier Klassenstufen betrachtet. Die anhand des Raschmodells geschätzten Personenparameter hinsichtlich des Gesamtproblemwerts des SDQ bilden dabei die abhängige Variable, während die Angabe der Klassenstufe als unabhängige Variable dient (zentriert auf das Ende der ersten Klassenstufe). Auf der zweiten Ebene wird das schülerspezifische Geschlecht berücksichtigt, da verschiedene Studien hier auf Unterschiede zwischen Jungen und Mädchen hinweisen (Costello at el., 2003; Ihle & Esser, 2008; Petermann, 2005).

Ergebnisse

Reliabilität

Die interne Konsistenz des SDQ Gesamtproblemwertes (α_{MZP1} = .87, α_{MZP2} = .90, α_{MZP3} = .90, α_{MZP4} = .90) und der Skala Prosoziales Verhalten (α_{MZP1} = .72, α_{MZP2} = .80, α_{MZP3} = .81, α_{MZP4} = .87) ist über alle vier Messzeitpunkte zufriedenstellend.

Die Interkorrelationsmatrix der Skala Prosoziales Verhalten über die vier Messzeitpunkte weist mittlere Zusammenhänge auf. Insgesamt erzeugen die Daten eine Simplexstruktur (je näher die Messzeitpunkte bei einander liegen, desto höher der Kennwert und umgekehrt). Für den Gesamtproblemwert ergeben sich ähnliche Ergebnisse, jedoch liegen die Werte deutlich höher (vgl. Tabelle 2). Der Gesamtproblemwert fällt somit über die Zeit stabiler aus als die Skala Prosoziales Verhalten. Nachfolgende Analysen sind ausschließlich auf den Gesamtproblemwert beschränkt.

Analysen zur Eignung für die Verlaufsmessung

Um eine Verlaufsmessung zu entwickeln ist es notwendig, eine Skala mit einem latenten Konstrukt für einen längeren Entwicklungszeitraum zu konstruieren. Daher wurde der SDQ Gesamtproblemwert mit 20 Items gewählt. In der vorliegenden Studie wurden die Kategorien „teilweise zutreffend“ und „eindeutig zutreffend“ hierbei zusammengenommen, denn ein teilweises Auftreten eines Störverhaltens kann von ungeschulten Lehrkräften als eine Abweichung vom Wunschzustand bewertet werden. Die ursprüngliche Kodierung führte in den vorliegenden Daten zu dem Problem, dass einzelne Ausprägungen bei den Items nur sehr selten oder gar nicht angekreuzt wurden. Häufig wurde die Mittelkategorie „teilweise zutreffend“ kaum verwendet. Des Weiteren

Tabelle 2: Interkorrelation der Daten der vier Messzeitpunkte nach Pearson

	Prosoziales Verhalten				**Gesamtproblemwert**			
	Kl. 1	**Kl. 2**	**Kl. 3**	**Kl. 4**	**Kl. 1**	**Kl. 2**	**Kl. 3**	**Kl. 4**
Kl. 1	1	.44	.45	.38	1	.69	.63	.58
Kl. 2		1	.42	.53		1	.70	.65
Kl. 3			1	.65			1	.77
Kl. 4				1				1

könnten bei einer Betrachtung des SDQ als Verlaufsmessung Items mit mehreren Kategorien zu Interpretationsschwierigkeiten des Summenwerts führen.

Insbesondere für die Messung des Verlaufs ist es wichtig, dass sich die Itemparameter über die Messzeitpunkte als konstant, d. h. messinvariant über die Zeit erweisen. Um dies zu prüfen, wurden die Itemparameter für den jeweiligen Messzeitpunkt berechnet und im Rahmen eines grafischen Modelltests jeweils an der X- und Y-Achse abgetragen (Abbildung 1). Wenn die Itemparameter über die Messzeitpunkte konstant sind, verlaufen sie entlang der Winkelhalbierenden. Die Konfidenzintervalle (95 %) werden anhand der Ellipsen angedeutet.

Insgesamt kann man die Itemparameter als annähernd konstant betrachten. In Abbildung 1 ist der grafische Modelltest zwischen jeweils zwei Messzeitpunkten dargestellt. Nur wenige Items haben zwischen den Messzeitpunkten kleinere Abweichungen (Klasse 1 zu 2: „Unruhig“, „Unglücklich“, „Zappelig“, „Bedacht“; Klasse 2 zu 3: „Konzentration“, „Wutanfälle“; Klasse 3 zu 4 „Beliebtheit“, „Gehänselt“; Klasse 1 zu 4: „Unruhig“, „Zappelig“, „Einzelgänger“, „Gehänselt“). Hier urteilen die Lehrkräfte

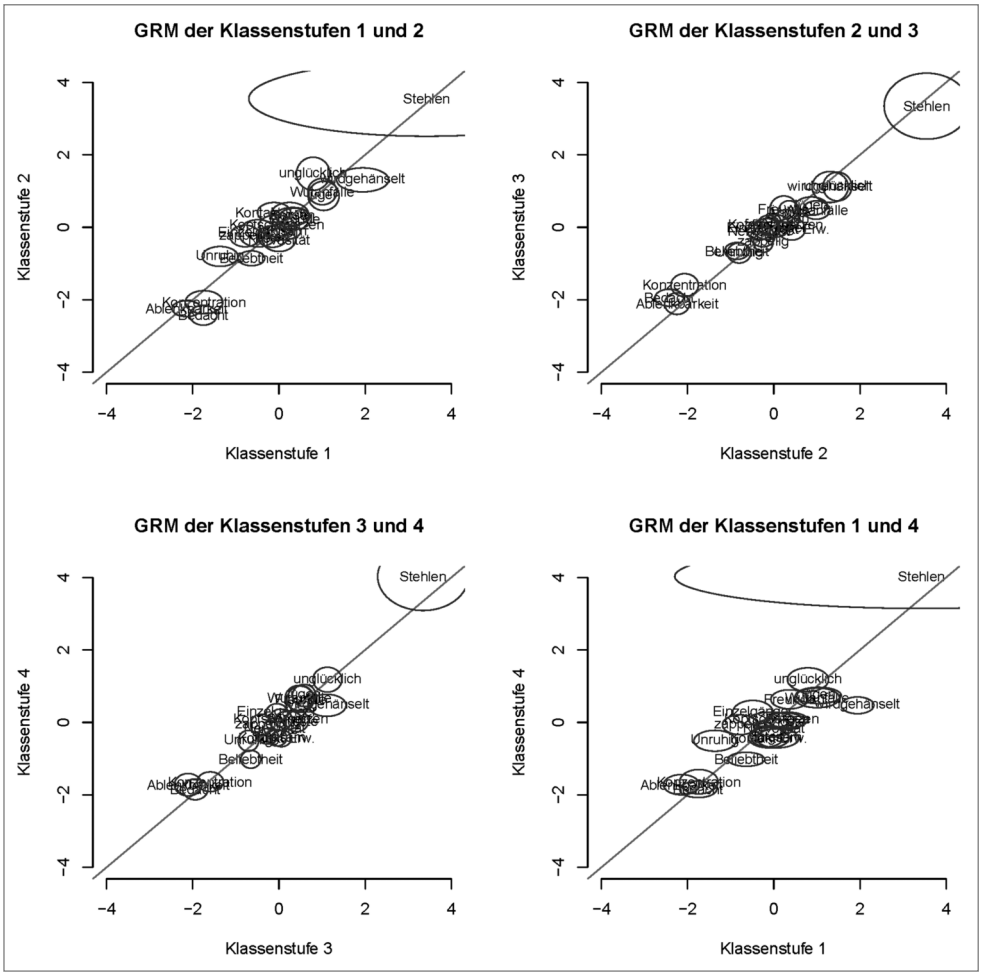

Abbildung 1: Analysen zur Messinvarianz zu den Messzeitpunkten 1 bis 4 in der Skala SDQ Gesamtproblemwert

über die verschiedenen Klassenstufen leicht unterschiedlich (Differential Item Functioning), dennoch erscheint es gerechtfertigt, die Itemparameter zu den verschiedenen Messzeitpunkten als konstant zu betrachten.

Raschmodell über vier Messzeitpunkte

Um den Gesamtproblemwert weiter zu prüfen, wurden für alle Items dieser Skala ein eindimensionales Raschmodell für alle vier Messzeitpunkte gleichzeitig berechnet und anschließend die Personenparameter für die jeweiligen Messzeitpunkte geschätzt (Rost, 2004, S. 287 ff.). Zur Prüfung des Raschmodells nach lokalen Modellverletzungen (d. h. Verletzungen auf Itemebene) werden die Mean-Square-Fit-Statistiken *Infit-* und *Outfit-Wert* herangezogen. Diese sollten nicht signifikant von ihrem Erwartungswert 1 abweichen (Wertebereich zwischen 0.7 und 1.5; Linacre, 2002).

In den vorliegenden Analysen ergaben sich für alle vier Messzeitpunkte zufriedenstellende Infit- und Outfit-Werte. Ausnahmen hierbei bilden die Items „Einzelgänger" und „Kopfschmerzen", die zum ersten Messzeitpunkt einen Outfit-Wert von 1.55 aufweisen, sowie das Item „Kontakt zu Erwachsen", welches zu jedem Messzeitpunkt einen Outfit-Wert über 1.6 erreicht. Gemäß Linacre (2002) sind jedoch Modellverletzun-

Tabelle 3: Items mit Itemparametern und Trennschärfe geordnet nach den Itemparametern

Item	Subskala	Itemparameter (*WLE*)	Trennschärfe			
			Klasse 1	Klasse 2	Klasse 3	Klasse 4
Stehlen	VP	3.59	.17	.25	.24	.27
Unglücklich	EP	1.14	.49	.5	.45	.46
Wird gehänselt	VPG	1.06	.36	.46	.55	.56
Lügen	VP	0.79	.43	.41	.53	.56
Wutanfälle	VP	0.76	.44	.55	.58	.56
Freunde	VPG	0.46	.49	.51	.54	.43
Ängste	EP	0.26	.41	.44	.40	.53
Sorgen	EP	0.22	.50	.54	.49	.51
Kopfschmerzen	EP	0.04	.37	.43	.42	.45
Kontakt Erw.	VPG	-0.06	.28	.37	.39	.45
Einzelgänger	VPG	-0.07	.43	.44	.43	.39
Folgsam	VP	-0.12	.56	.56	.60	.66
Nervosität	EP	-0.15	.50	.62	.49	.55
Streit	VP	-0.27	.61	.58	.60	.65
Zappelig	HA	-0.34	.64	.62	.61	.61
Beliebtheit	VPG	-0.80	.65	.65	.66	.67
Unruhig	HA	-0.80	.64	.65	.61	.64
Konzentration	HA	-1.73	.65	.71	.64	.69
Bedacht	HA	-1.98	.73	.70	.67	.70
Ablenkbarkeit	HA	-2.02	.66	.68	.66	.66

Erläuterungen: VP – Verhaltensprobleme; EP – Emotionale Probleme; VPG – Verhaltensprobleme mit Gleichaltrigen; HA – Hyperaktivität

gen der Outfit-Werte weniger bedeutsam als es die der Infit-Statistiken wären.

In Tabelle 3 sind die Itemparameter nach Schwierigkeit geordnet und die punktbiseriale Korrelation des Items mit dem WLE-Schätzer des Gesamtproblemwertes der jeweiligen Klassenstufen dargestellt. Es zeigt sich, dass die Items „Ablenkbarkeit“ oder „Bedacht“ zu den leichten, d. h. häufig von den Lehrkräften als zutreffend gewählten Items gehören, „Stehlen“ ist mit Abstand das schwierigste, d.h. am seltensten als zutreffend angegebene Item. Die punktbiseriale Korrelation kann im Raschmodell als Trennschärfe betrachtet werden. Die Kennwerte können – bis auf Ausnahme des Items „Stehlen“ (zu hohe Schwierigkeit) – als zufriedenstellend eingeschätzt werden. Auffällig ist der Mangel an Items in einem Schwierigkeitsbereich von 1.0 bis 3.7. Die Items der ursprünglichen Dimensionen des SDQ verteilen sich anhand der Itemschwierigkeit gleichmäßig. Eine Ausnahme ist die ursprüngliche Dimension Hyeraktivität (HA), welche ausschließlich sehr leichte Items beinhaltet.

Die Verteilung der Personenparameter

Die Verteilung der Personenparameter im Vergleich zu den Itemparametern kann anhand der Person-Item-Map dargestellt werden. Nachdem das Item „Kontakt zu Erwachsen“ einen schlechten Modellfit zu allen Messzeitpunkten aufwies, wurde für die Berechnung der Personenwerte ein Modell unter Ausschluss dieses Items berechnet. Um die Entwicklung über die Zeit abzubilden, wurden die Itemparameter zum Ende der ersten Klasse auch für die Schätzung der Personenparameter bis zum Ende der Klasse 4 zugrunde gelegt. In Abbildung 2 sind die Personenparameter als Histogramme dargestellt, welche den Itemparametern gegenübergestellt werden. Hierbei zeigt sich, dass der Großteil der Personen vor allem unterhalb des Nullpunktes angesiedelt ist und die Verteilung der Daten, optisch beurteilt,

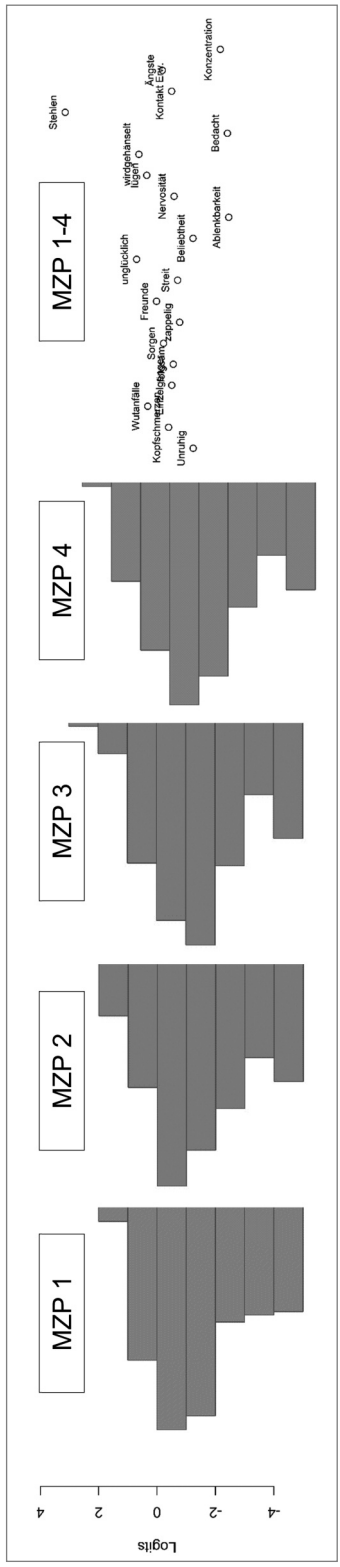

Abbildung 2: Person-Item-Map für die Messzeitpunkte 1 bis 4

recht stabil bleibt. Besonders auffällig ist, dass es insbesondere an Items mangelt, deren Messbereich sich über diejenigen Kinder erstreckt, die sich im unteren (unauffälligen) Bereich der Verteilung befinden.

Mehrebenenmodell zur Veränderung über die Zeit

Zur Analyse der Veränderung der SDQ-Daten über die Zeit wurde ein Mehrebenenmodell über die zuvor anhand des Raschmodells ermittelten Personenparameter (WLEs) spezifiziert. Die Ergebnisse der HLM weisen zum Ende der Klasse 1 einen durchschnittlichen Gesamtproblemwert von β_{00} = -1.05 ($p < .001$) für die untersuchten Jungen aus, mit einem mittleren Anstieg für ebendiese Gruppe von β_{10} = 0.16 je Schuljahr ($p < .001$, $d = 0.18 \pm 0.04$). Das Geschlecht der Kinder hat einen signifikanten Einfluss auf den Problemwert zum Ende der ersten Klasse zum Vorteil der Mädchen (β_{01} = -0.69, $p < .001$, $d = -0.77 \pm 0.18$), nicht aber auf den Anstieg über die Schuljahre. Die Zufallseffekte weisen weitere schülerspezifische Unterschiede im Niveau und Anstieg des SDQ aus, die im Rahmen des gerechneten Modells nicht erklärt werden konnten. Die Ergebnisse des Mehrebenenmodells sind in Tabelle 4 zusammengefasst.

Modelliert man die Verläufe im SDQ über die einzelnen Messzeitpunkte, wird deutlich, dass es unterschiedliche Anstiege zwischen den Klassenstufen gibt. Während vom Ende der ersten zum Ende der zweiten Klasse (β_{10} = 0.30, $p < .01$, $d = 0.38 \pm 0.13$) sowie vom Ende der dritten zum Ende der vierten Klasse (β_{10} = 0.28, $p < .01$, $d = 0.31 \pm 0.11$) Änderungen im Sinne eines signifikant von null verschiedenen Anstiegs festzuhalten sind, bleiben die SDQ-Daten vom Ende der zweiten bis zum Ende der dritten Klasse stabil (β_{10} = 0.02, $p > .05$).

Diskussion

Der SDQ ist ein weit verbreitetes Screeninginstrument, für welches im Rahmen diverser Studien die psychometrische Güte bereits nachgewiesen werden konnte (Döpfner & Petermann, 2008; Koglin et al., 2007; Lohbeck et al., 2015). Auch durch die vorliegende Studie konnte gezeigt werden, dass der SDQ Gesamtproblemwert über die vier

Tabelle 4: Ergebnisse des random-coefficient-Modells zur Analyse des SDQ Gesamtproblemwerts (auf Basis der ermittelten WLEs) über die Zeit

Feste Effekte	**β (*SE*)**	***t***	***df***	***d***
Modell für den intercept β_{0i}				
SDQ Niveau, β_{00}	-1.05 (0.13)***	-7.94	427	-1.16 ± 0.14
weiblich, β_{01}	-0.69 (0.17)***	-3.99	427	-0.77 ± 0.18
Modell für den slope β_{1i}				
SDQ Anstieg, β_{10}	0.16 (0.04)***	3.67	427	0.18 ± 0.04
weiblich, β_{11}	-0.11 (0.06)	-1.80	427	
Zufallseffekte	**Varianz**	**χ^2**	**df**	
Niveau u_{0i}	2.11***	1471.78	382	
Anstieg u_{1i}	0.13***	598.12	382	
Level-1-Fehler e_{ij}	0.81			
Devianzstatistik = 4725.03, *df* = 4				

Erläuterung: *** $p < .001$

Messzeitpunkte jeweils zum Ende der Klassenstufen 1 bis 4 reliable Werte liefert. Für die Skala Prosoziales Verhalten mit ihren fünf Items ergeben sich erwartungsgemäß geringere Werte hinsichtlich der internen Konsistenz. Die Kennwerte zur internen Konsistenz, Normierungswerte und Korrelationen fielen über die Jahrgangsstufen ähnlich hoch aus wie in der Forschungsliteratur beschrieben (z. B. Lohbeck et al., 2015; Rothenberger, Becker, Erhart, Wille, Ravens-Sieberer & die BELLA-Arbeitsgruppe, 2008).

Im Gegensatz zu seiner statusdiagnostischen Eignung fehlt es an Forschungsbeiträgen, die die Güte des SDQ zum Einsatz als Monitoring untersuchen. In der Analyse der verlaufsdiagnostischen Eignung des SDQ lag ein weiteres Ziel des vorliegenden Beitrags. Da die Skala Prosoziales Verhalten nur wenige Items aufweist und im SDQ nicht mit den anderen Dimensionen verrechnet wird, wurde bei den weiteren Analysen auf die Betrachtung dieser Skala Verhalten verzichtet.

Nach Zusammenlegung der Kategorien „teilweise zutreffend“ und „eindeutig zutreffend“ erfüllte der SDQ Gesamtproblemwert die Anforderungen des Raschmodells weitgehend und fiel über die vier Messzeitpunkte grundlegend messinvariant aus. Geringere Abweichungen im grafischen Modelltest sind vor allem zwischen den Zeitpunkten Ende Klasse 1 zu Ende Klasse 2 festzustellen. Dies kann dahingehend interpretiert werden, dass die Lehrkräfte ihre Schülerinnen und Schüler hinsichtlich einzelner Items offenbar mit leicht veränderten Bezugsrahmen sehen und bewerten. Aus entwicklungspsychologischer Sichtweise ist eine Veränderung des Bezugsrahmens dahingehend nachvollziehbar, da von einem Kind einer höheren Klassenstufe ein anderes Verhalten als von einer Erstklässlerin bzw. einem Erstklässler erwartet wird. Insbesondere werden in der ersten Klasse Arbeitstechniken und Unterrichtsverhalten grundgelegt und eingeübt, die für die höheren Klassenstufen vorausgesetzt werden. Ein veränderter Bezugsrahmen für das von der Lehrperson wahrgenommene Verhalten von Klasse 3 zu 4 ist vor dem Hintergrund einleuchtend, dass die Grundschulzeit in Mecklenburg-Vorpommern mit der vierten Klasse endet und die Schulkarriere mit Klasse 5 in der Regionalen Schule fortgesetzt wird. In diesem Zusammenhang könnte eine „verschärfte“ Verhaltensbeurteilung durch die Lehrperson, wie sie sich hier abbildet, im Sinne einer Eignungsprüfung für die „neue Schule“ aufgefasst werden.

Das Item „Kontakt zu Erwachsen“ wurde aufgrund ungünstiger Fit-Werte im Rahmen der Raschmodellierung des Gesamtproblemwertes als ungeeignet eingestuft und aus weiteren Analysen ausgeschlossen. Dies ist auch theoretisch nachvollziehbar, da der SDQ Gesamtproblemwert in der Lehrkraftversion internalisierende oder externalisierende Verhaltensauffälligkeiten im Klassenzimmer messen soll. Offenbar fällt es Lehrkräften schwer, dieses Item zu bewerten, zumal es in diesem Setting weniger von Bedeutung ist als im außerschulischen Bereich. Dies spiegelt sich in der vorhandenen Datenlage, so gibt es deutliche Unterschiede zwischen dem Item und dem Gesamtwert, das Item scheint eine andere Dimension abweichenden Verhaltens zu erfassen. Dass einzelne Items zu einer ungünstigen Modellpassung führen und entsprechend eine Überarbeitung des SDQ angezeigt scheint, konstatieren bereits andere Forschergruppen (u. a. Lohbeck et al., 2015). Weitere Fit-Statistiken einzelner Items in der vorliegenden Untersuchung („Einzelgänger“ und „Kopfschmerzen“) fielen lediglich zum ersten Messzeitpunkt eher ungünstig aus. Diese Modellverletzungen erscheinen nicht so erheblich, als dass ein Vergleich der Entwicklung über die Zeit nicht gerechtfertigt wäre. Eine weitere Verwendung dieser Items unter verlaufsdiagnostischer Perspektive ist somit unbedenklich. Daher wurden in einem gemeinsamen Modell die Personenparameter des jeweiligen Messzeitpunktes aus den gemeinsamen Itemparametern über alle Messzeitpunkte gebildet (Rost, 2004).

Zur Analyse der Eignung des SDQ als Verlaufsinstrument ist zudem die Auswertung der Person-Item–Map der Daten von Relevanz. Hier erkennt man, dass der SDQ eine Diskrepanz zwischen dem Messbereich der Items und den ermittelten Personenparametern aufweist. Dieses ungünstige „Targeting" liegt darin begründet, dass der SDQ ein Screeninginstrument darstellt, was vor allem einen differenzierteren Blick im Randbereich einer Verteilung erzielen soll. Da es sich bei der verwendeten Untersuchungsgruppe nicht um eine klinische Stichprobe handelt, ist davon auszugehen, dass viele Personen sich auch nicht in diesem Randbereich bewegen und deren Merkmale somit nicht hinreichend durch die Itemparameter abgedeckt sind. Verfolgt man das Ziel, mit dem SDQ Verläufe zu ermitteln, sollten zusätzliche Items entwickelt werden, die eine Beurteilung von Personen im unteren Bereich (leichte Verhaltensauffälligkeiten) ermöglichen. Die Ergänzung von weiteren Items ist durch die hierarchische Stufung der Items durch die Itemparameter dank des Raschmodells möglich. Eine Erweiterung um Items sollte insbesondere auch im Schwierigkeitsbereich zwischen *WLE* = 1.17 („unglücklich") und *WLE* = 3.71 („stehlen") erfolgen, um einen differenzierte Einordnung der Kinder im oberen Randbereich zu ermöglichen.

Im Rahmen von Mehrebenenmodellierungen wurden die Veränderungen der Personenparameter der Schülerinnen und Schüler analysiert. Die untersuchten Kinder wurden von ihren Lehrkräften über die Messzeitpunkte durchaus unterschiedlich hinsichtlich ihres Verhaltens eingeschätzt. So gibt es signifikante Anstiege in den anhand des Raschmodells geschätzten Personenparametern über die Zeit, die vor allem im Übergang von Ende Klasse 1 zu 2 sowie Klasse 3 zu 4 zu begründen sind. Insgesamt erscheint eine klassen- bzw. altersstufenabhängige Normierung des SDQ angezeigt.

Deutliche Geschlechtsunterschiede im Verhalten zeigen sich zum Ende der ersten Klasse. Hier weisen Jungen deutlich ungünstigere Verhaltenstendenzen auf. Dieser Befund ist konform zu Angaben aus anerkannten Prävalenzstudien zum Thema Verhaltensauffälligkeiten im Kindes- und Jugendalter (z. B. Ihle & Esser, 2008; Petermann, 2005). Die hier dargelegte Befundlage zeigt jedoch keine geschlechtsspezifischen Unterschiede im Anstieg an, d. h. die Mädchen und Jungen der untersuchten Stichprobe entwickelten sich hier gleich.

Die Fortentwicklung eines Verlaufsdiagnostikums aus dem SDQ erscheint vielversprechend. Um dies zu erreichen, sollte das Verfahren um „leichtere" Items ergänzt werden, die vor allem eine Erfassung von Verhaltensausprägungen im Grenzbereich zur Verhaltensauffälligkeit erlauben. Diese Items erlauben somit ein besseres Targeting in unausgelesenen Stichproben. Gleichzeitig sollte sichergestellt sein, dass diese Items nur einen geringen Aufwand bei der Bewertung besitzen (Casale et al., 2015). Dies sind vor allem Items, die klar umschrieben und operationalisiert sind und sich auf weniger komplexe Sachverhalte beziehen, z. B. „Redet oft dazwischen" oder „Meldet sich häufig im Unterricht". Damit würde der SDQ zudem um den Bereich des Arbeitsverhaltens erweitert, was für einen Einsatz im schulischen Setting sinnvoll erscheint. Ebenso sollten Items im oberen Messbereich ergänzt werden, die eine differenziertere Beurteilung von Kindern mit deutlichen Anzeichen für Verhaltensauffälligkeiten ermöglichen, z. B. „Beleidigt Mitschülerinnen und Mitschüler" oder „Stört den Unterricht".

Mit den vorgeschlagenen Überarbeitungshinweisen ist der SDQ grundsätzlich als ein Instrument für wiederholendes Messen geeignet, allerdings in größeren Zeitspannen, z. B. viertel- bzw. halbjährlich. Für einen hochfrequenten Einsatz zur täglichen oder wöchentlichen Abschätzung der Verhaltensentwicklung wie es im Bereich des Lernens durch Curriculum-based Measurements (u. a. Deno, 1985; Voß & Hartke, 2014) erreicht werden soll, eignet sich der SDQ nicht. Dafür hat der SDQ einen zu

breiten Bezugsrahmen und zu wenige Items. Für hochfrequente Messungen eignen sich sog. direkte Verhaltensbeurteilungen (Direct Behavior Ratings; u. a. Christ, Riley-Tillman & Chafouleas, 2009; Volpe & Fabiano, 2013). In diesen bewerten die Lehrkräfte ein Item eines Verhaltensaspektes mit 10 Kategorien wiederkehrend und prüfen so kurzfristige Verhaltensänderungen in diesem Bereich. Daher ist eine Kombination aus einer gröberen, aber breiteren Messung mit dem SDQ und einer feineren, aber spezifischen Messung mittels Direkt Behavior Ratings im Unterricht empfehlenswert.

Die Analysen der Studie konzentrieren sich auf die Auswertung als Verlaufsmessung. Die Dimensionalität des SDQ, die Mehrebenenstruktur in der Skalierung und die Ratingqualität der Lehrkräfte wurde nicht berücksichtigt bzw. nicht überprüft. Daher kann der Einfluss des Raters und dessen Bias nicht festgestellt werden, hierfür benötigt man weitere Studien. Aufgrund des longitudinalen Designs ist die Stichprobe auch nur auf eine kleinere Stadt beschränkt. Daher kann es auch leichter zu Stichprobeneffekten kommen im Vergleich zu großen Normierungsstudien.

Literaturverzeichnis

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Amelang, M. & Zielinski, W. (2004). *Psychologische Diagnostik und Intervention*. Berlin: Springer.

Beelmann, A. & Lösel, F. (2007). Entwicklungsbezogene Prävention dissozialer Verhaltensprobleme: Eine Meta-Analyse zur Effektivität sozialer Kompetenztrainings. In W. von Suchodoletz (Hrsg.), *Prävention von Entwicklungsstörungen* (S. 235 – 258). Göttingen: Hogrefe.

Beelmann, A. & Raabe, T. (2007). *Dissoziales Verhalten von Kindern und Jugendlichen*. Göttingen: Hogrefe.

Beelmann, A. (2008). Förderung sozialer Kompetenzen im Kindergartenalter: Programme, Methoden, Evaluation. *Empirische Pädagogik, 22*, 160 – 177.

Bell, B. & Cowie, B. (2001). *Formative Assessment and Science Education* (Vol. 12). Dordrecht, Boston: Kluwer Academic.

Brezinka, V. (2003). Zur Evaluation von Präventivinterventionen für Kinder mit Verhaltensstörungen. *Kindheit und Entwicklung, 12*, 71-83.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Vol. 1). Newbury Park: Sage.

Casale, G., Hennemann, T., Huber, C. & Grosche, M. (2015). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung, 41*, 37-54.

Choppin, B. (1968). Item Bank using Sample-free Calibration. *Nature, 219*, 870-872.

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.

Costello, E. J., Egger, H. & Angold, A. (2005). 10-Year Research Update Review: The Epidemiology of Child and Adolescent Psychiatric Disorders: I. Methods and Public Health Burden. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 972–986.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G. & Angold, A. (2003). Prevalence and Development of Psychiatric Disorders in Childhood and Adolescence. *Archives of General Psychiatry, 60*, 837–844.

Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children, 52*, 219–232.

Döpfner, M. & Petermann, F. (2008). *Diagnostik psychischer Störungen*. Göttingen: Hogrefe.

Frostad, P. & Pijl, S. J. (2007). Does Being Friendly Help in Making Friends? *European Journal of Special Needs Education, 22*, 15-30.

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

Garner, P. W. (2010). Emotional Competence and its Influences on Teaching and Learning. *Educational Psychology Review, 22*, 297–321

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

Goodman, A., Lamping, D. L. & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ). *Journal of Abnormal Child Psychology, 38*, 1179-1191.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry, 38*, 581–586.

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337–1345.

Goodman, R., Iervolino, A.C., Collishaw, S., Pickles, A. & Maughan, B. (2007). Seemingly Minor Changes to a Questionnaire Can Make a Big Difference to the Mean Scores: A Cautionary Tale. *Social Psychiatry and Psychiatric Epidemiology, 42*, 322–327.

Hartke, B. (2005). Schulische Prävention – welche Maßnahmen haben sich bewährt? In S. Ellinger & M. Wittrock (Hrsg.), *Sonderpädagogik in der Regelschule. Konzepte, Forschung, Praxis* (S. 11-37). Stuttgart: Kohlhammer.

Hasselhorn, M., Schneider, W. & Trautwein, U. (Hrsg.). (2014). *Lernverlaufsdiagnostik*. Göttingen: Hogrefe.

Heine, J.-H. & Tarnai, C. (2015). Pairwise Rasch Model Item Parameter Recovery under Sparse Data Condtions. *Psychological Test and Assessment Modeling 57*(1), 3-36.

Heine, J.-H. (2014). *pairwise: Rasch Model Parameters by Pairwise Algorithm* [Computer software]. Munich. Zugriff am 01.02.2016. Verfügbar unter http://cran.r-project.org/web/packages/pairwise/index.html (R package version 0.2.5).

Huber, C. (2006). *Soziale Integration in der Schule?!* Marburg: Tectum.

Ihle, W. & Esser, G. (2008). Epidemiologie psychischer Störungen des Kindes- und Jugendalters. In: B. Gasteiger-Klicpera, H. Julius & C. Klicpera (Hrsg.), *Sonderpädagogik der sozialen und emotionalen Entwicklung* (Band 3 des Handbuchs Sonderpädagogik, S. 49-62). Göttingen: Hogrefe.

Klauer, K. J. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forschung, 32*(1), 16–26.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein, U. (Hrsg.), *Lernverlaufsdiagnostik* (Tests & Trends, NF Bd. 12., S. 1-17). Göttingen: Hogrefe.

Kóbor, A., Takács, Á., & Urbán, R. (2013). The Bifactor Model of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment, 29*, 299-307.

Koglin, U., Barquero, B., Mayer, H., Scheithauer, H. & Petermann, F. (2007). Deutsche Version des Strenghts and Difficulties Questionnaire (SDQ-Deu): Psychometrische Qualität der Lehrer-/Erzieherinnenversion für Kindergartenkinder. *Diagnostica, 53*, 175-183.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized Mean? *Rasch Measurement Transactions, 16*, 878.

Linderkamp, F. & Grünke, M. (2007). *Lern- und Verhaltensstörungen – Genese, Diagnostik & Intervention*. Weinheim: Psychologie Verlags Union.

Lohbeck, A., Schultheiß, J., Petermann, F. & Petermann, U. (2015). Die deutsche Selbstbeurteilungsversion des Strengths and Difficulties Questionnaire (SDQ-Deu-S): Psychometrische Eigenschaften, Faktorenstruktur und Grenzwerte. *Diagnostica, 62*, 3-33.

Petermann, F. (2005). Zur Epidemiologie psychischer Störungen im Kindes- und Jugendalter. Eine Bestandsaufnahme. *Kindheit und Entwicklung, 14*, 48-57.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M. R. & Rahman, A. (2007) No Health without Mental Health. *The Lancet, 370*, 859–877.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [Computer software]. Vienna, Austria. Retrieved from http://www.R-project.org.

Reef, J., Diamantopoulou, S., van Meurs, I., Verhulst, F. C. & van der Ende, J. (2011). Developmental Trajectories of Child to Adolescent Externalizing Behavior and Adult DSM-IV Disorder: Results of a 24-year Longitudinal Study. *Social Psychiatry Psychiatric Epidemiology, 46*, 1233–1241.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), Leistungsmessung in Schulen (S. 59-71). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

Rothenberger, A., Becker, A., Erhart, M., Wille, N., Ravens-Sieberer, U. & die BELLA-Arbeitsgruppe (2008). Psychometric properties of the parent strengths and difficulties questionnaire in the general population of German children and adolescents: results of the BELLA study. *European Child & Adolescent Psychiatry, 17*, 99-105.

Steinhausen, H.-C. (2010). *Psychische Störungen bei Kindern und Jugendlichen*. München: Elsevier.

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A. & Janssens, J. M. A. M. (2010). Psychometric Properties of the Parent and Teacher Versions of the Strengths and Difficulties Questionnaire for 4-to 12-year-olds: A Review. *Clinical Child and Family Psychology Review, 13*, 254–274.

Volpe; R. J. & Fabiano, G. A. (2013). Daily Behavior Report Cards. An Evidence-Based System of Assessment and Intervention. New York, NY: Guilford Press.

Voß, S. & Hartke, B. (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein, U. (Hrsg.), *Lernverlaufsdiagnostik* (Tests & Trends, NF Bd. 12., S. 83-99). Göttingen: Hogrefe.

Voß, S. (2014). *Curriculumbasierte Messverfahren im mathematischen Erstunterricht – Zur Güte und Anwendbarkeit einer Adaption US-amerikanischer Verfahren im deutschen Schulsystem*. Saarbrücken: SVH.

Voß, S., Sikora, S. & Hartke, B. (2017). Lernverlaufsdiagnostik als zentrales Element der Prävention von Rechenschwierigkeiten. In A. Fritz-Stratmann, G. Ricken & S. Schmidt (Hrsg.), *Handbuch Rechenschwäche* (3. Überarb. Aufl., S. 339-355). Weinheim: Beltz.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54*, 427–450.

Wiedebusch, S. & Petermann, F. (2011). Förderung sozial-emotionaler Kompetenz in der frühen Kindheit. *Kindheit und Entwicklung, 20*, 209-218.

Wilbert, J. & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik, 3*, 225-242.

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung: Gütekriterien und Auswer-

tungsanforderungen In M. Hasselhorn, W. Schneider & U. Trautwein, U. (Hrsg.), *Lernverlaufsdiagnostik* (Tests & Trends, NF Bd. 12., S. 281-308). Göttingen: Hogrefe.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Dr. Stefan Voß
Universität Rostock
Institut für Sonderpädagogische Entwicklungsförderung und Rehabilitation
August-Bebel-Str. 28
18051 Rostock
stefan.voss3@uni-rostock.de

Erstmalig eingereicht: 16.12.2016
Überarbeitung eingereicht: 02.04.2017
Angenommen: 11.05.2017