Empirische Sonderpädagogik, 2015, Nr. 3, S. 258-268
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse

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

[1] *Universität zu Köln*
[2] *Northeastern University, Boston (MA)*
[3] *Bergische Universität Wuppertal*

Zusammenfassung

Die vorliegende Studie untersucht die Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen (DVB) des Lern- und Arbeitsverhaltens in einer inklusiven Grundschule. In einer Generalisierbarkeitsstudie mit einem vollständig gekreuzten Zwei-Facetten-Design (*Rater* und *Item*) werden 10 Grundschulkinder von 6 geschulten Beurteilenden anhand von gefilmten Unterrichtsphasen beobachtet und das Lern- und Arbeitsverhalten mit einer DVB mit 5 Items eingeschätzt. Die Ergebnisse der Generalisierbarkeitsstudie zeigen erwartungskonform eine hohe Varianzaufklärung durch Unterschiede zwischen den Personen. Allerdings ist der Interaktionseffekt zwischen Ratern und Kindern trotz umfangreicher Schulung substantiell. Dennoch weisen die Ergebnisse einer Entscheidungsstudie auf eine hohe Generalisierbarkeit und Zuverlässigkeit der Daten hin. Die Befunde sprechen für einen Einsatz des Instruments zur Verlaufsdiagnostik von Schülerverhalten.

Schlagwörter: Verlaufsdiagnostik, Schülerverhalten, Generalisierbarkeitstheorie

Generalizability and Dependability of Direct Behavior Ratings of Academically Engaged Behavior in an Inclusive Classroom Setting

Abstract

This study focuses on generalizability and dependability of direct behavior ratings of academically engaged behavior in an inclusive classroom setting. In a fully-crossed 2 facet generalizability study design (*raters* and *items*) 6 trained observers rated 10 students' academically engaged behavior with a direct behavior rating multiple item scale. As expected, results of the generalizability study show that differences between persons explain most of the total variance. However, the interaction between raters and students is despite rater training relatively high. Nevertheless, decision study results suggest good generalizability and dependability. The results support the usability of direct behavior ratings for formative assessment of student behavior.

Keywords: formative assessment, students' behavior, direct behavior rating, generalizability theory

Die Erfassung von Schülerverhalten im Entwicklungsverlauf über die Zeit spielt eine entscheidende Rolle in sonderpädagogischen Handlungsfeldern (Grosche & Volpe, 2013; Hillenbrand, 2015; Huber & Grosche, 2012). Die diagnostizierten individuellen Entwicklungsverläufe liefern Daten, anhand derer entschieden werden kann, ob die pädagogische Förderung so weiter geführt werden kann oder besser auf die Lernbedürfnisse eines Kindes ausgerichtet werden muss.

Die testdiagnostischen Anforderungen, die an Instrumente zur Erfassung des Verlaufs von Schülerverhalten gestellt werden, sind allerdings enorm (Christ, Riley-Tillman & Chafouleas, 2009). Zum einen müssen sie wichtige psychometrische Testgütekriterien erfüllen (Wilbert, 2014), zum anderen müssen sie auch flexibel und ökonomisch einsetzbar sowie für häufige Messzeitpunkte geeignet sein (Grosche, 2014). Die bisherigen zur Statusdiagnostik von Verhalten eingesetzten Verfahren entsprechen diesen Gütekriterien jedoch nur unzureichend (Casale, Hennemann, Huber & Grosche, 2015a). Damit fehlt es im deutschsprachigen Raum an für die Verlaufsdiagnostik von Schülerverhalten geeigneten und wissenschaftlich überprüften Instrumenten (Casale et al., 2015a; Huber & Rietz, 2015).

Im englischsprachigen Raum hat sich das sogenannte *direct behavior rating* als eine neuartige Methode zur Verlaufsdiagnostik von Schülerverhalten entwickelt, die diese hohen Anforderungen erfüllen könnte (Christ et al., 2009). Es vereint die Vorteile der systematischen und direkten Verhaltensbeobachtung und der Verhaltensbeurteilung mittels Ratingskalen. In einem festgelegten relativ kurzen Zeitraum wird ein bestimmtes konkret operationalisierbares Zielverhalten beobachtet und direkt im Anschluss an diesen Zeitraum auf einer Ratingskala eingeschätzt. Aufgrund dieser ökonomischen Vorgehensweise kann das Rating sehr häufig – bis zu mehrmals am Tag – wiederholt werden. Die Ergebnisse lassen sich über die Zeit in einem Liniendiagramm darstellen, so dass Verläufe und Entwicklungen von Schülerverhalten sichtbar werden (Christ et al., 2009). Im deutschsprachigen Raum wird die Methode als *Direkte Verhaltensbeurteilung* (DVB) bezeichnet (Casale et al., 2015a; Casale, Hennemann & Grosche, 2015b; Huber & Rietz, 2015).

Die Forschungsbefunde zu DVB stammen bislang ausschließlich aus dem nordamerikanischen Raum, weisen allerdings in Ansätzen auf eine gute Reliabilität und Validität der Methode hin (deutschsprachige Übersichten bei Casale et al., 2015, sowie Huber & Rietz, 2015). Daher gilt es, auch im deutschsprachigen Raum DVB in Bezug zu den oben genannten Erfordernissen zu entwickeln und deren Testgüte zu untersuchen. Die Forderung nach der Evaluation der Testgüte ist bei DVB besonders wichtig, weil es sich bei der Direkten Verhaltensbeurteilung um Fremdeinschätzungen von beobachtbaren Verhaltensweisen durch Lehrkräfte handelt. Diese Einschätzungen werden in großem Maße von mehreren teilweise abhängigen systematischen Fehlerquellen beeinflusst (Schmidt-Atzert & Amelang, 2012), die kaum durch die üblichen testtheoretischen Ansätze evaluiert werden können.

Der methodische Ansatz der Generalisierbarkeitstheorie

Daher wird in der vorliegenden Studie der methodische Ansatz der *Generalisierbarkeitstheorie* (G-Theorie) gewählt, um die Güte des Instruments zu überprüfen und daraus Implikationen für die Verbesserung des Instruments abzuleiten. Die *G-Theorie* wurde von Cronbach, Gleser, Nanda und Rajaratnam (1972) in die Sozialwissenschaften eingeführt. Ausgangspunkt war die ewig währende Frage nach der Reliabilität und Validität von Verhaltensmessungen und dem Einfluss multipler Fehlerquellen (Beobachter, Testinstrument, Beobachtungssituation etc.) auf die Ergebnisse dieser Messungen, dem die Klassische Testtheorie (KTT) nicht gerecht werden kann (Cronbach et al.,

1972). In der KTT wird postuliert, dass sich der beobachtete Wert aus dem wahren aber unbekannten Wert und einem globalen Messfehler zusammensetzt. Die G-Theorie stellt eine Erweiterung der KTT dar, in dem von einer Zerlegung des in der KTT angenommen globalen Fehlerwerts in einzelne Facetten ausgegangen wird (Brennan, 2001). Dies geschieht durch die gleichzeitige Schätzung der Varianzkomponenten, aus denen sich der Messfehler zusammensetzt. Verdeutlicht wird dies durch einen statistischen Vergleich der Varianzkomponenten. In der KTT gilt folgendes:

$$Var(Y) = Var(T) + Var(E) \quad (1)$$

wobei *Var(Y)* die Varianz des beobachteten Werts, *Var(T)* die Varianz des wahren Werts und *Var(E)* die Varianz des globalen Messfehlers repräsentiert. Hingegen wird in der G-Theorie die Varianz des beobachteten Werts in seine Bestandteile zerlegt:

$$\begin{aligned} Var(Y) = {} & Var(i) + Var(j) + Var(k) \\ & + Var(ij) + Var(ik) + \\ & Var(jk) + Var(ijk, e) \end{aligned} \quad (2)$$

Hier werden beispielhaft die Varianzkomponenten für drei Facetten (*i, j* und *k*) und deren Interaktionen geschätzt, theoretisch können aber unendlich viele Facetten modelliert werden. Es zeigt sich, dass die Varianzkomponentenschätzung in (2) für jede einzelne Facette konzeptionell dem wahren Wert *Var(T)* in (1) entspricht.

Der wahre Wert aus der KTT wird in der G-Theorie als *universaler Wert* bezeichnet. In der KTT kann der wahre Wert durch die Berechnung eines Durchschnittswerts über die Anzahl vergleichbarer, paralleler Messungen geschätzt werden. In der G-Theorie wird der universale Wert über die Werte festgelegter Bedingungen innerhalb einer Facette geschätzt (z.B. Grundschulkinder innerhalb der Facette *Person*). Theoretisch gibt es eine unendlich große Anzahl an Bedingungen, unter denen der universale Wert ermittelt werden kann (z.B. könnte man auch noch Förderschüler, Sekundarstufenschüler und Kindergartenkinder in die Facette *Person* miteinbeziehen). Daher spricht man in der G-Theorie vom *Universum der zulässigen Bedingungen*. Die Bedingungen, aus denen sich die Facetten zusammensetzen, werden hinsichtlich der Fragestellung und des Forschungsinteresses ausgewählt. Sie stellen eine Zufallsauswahl aus dem Universum aller zulässigen Beobachtungen dar (Brennan, 2001).

Das methodische Vorgehen in der G-Theorie erlaubt es, die Varianz mehrerer Fehlerquellen sowie deren Interaktionen untereinander simultan zu schätzen. Damit bietet sie einen entscheidenden Vorteil gegenüber der KTT, wo lediglich der Einbezug einer einzigen systematischen Fehlerquelle (z.B. der Beobachtereinfluss bei der Analyse der Interrater-Reliabilität oder der Einfluss der Situation bei der Analyse der Test-Retestreliabilität) zulässig ist und deren Interaktionen gar nicht überprüfbar sind. Während die KTT also nur die Größe eines Messfehlers abschätzen kann, kann die G-Theorie aufschlüsseln, welche erwünschten und unerwünschten Quellen die Messvarianz beeinflussen und wie das Messinstrument zu verbessern wäre.

Das methodische Vorgehen innerhalb der Generalisierbarkeitstheorie gliedert sich in zwei Schritte. Der erste Schritt erfolgt in Form einer sogenannten Generalisierbarkeitsstudie (G-Studie). G-Studien schätzen die Varianz der einzelnen Facetten und deren Interaktionen untereinander, um festzustellen, in welchem Ausmaß sie zur Messgenauigkeit beitragen. Dies geschieht mittels Varianzanalysen, wobei die Facetten als Faktoren und die Bedingungen der Facetten als Faktorstufen behandelt werden (Brennan, 2001). Die Ergebnisse der G-Studie werden dann als Ausgangspunkt für den zweiten Schritt, die sogenannte Entscheidungsstudie (D-Studie; *decision study*), genutzt. Ziel der D-Studie ist es, das Messinstrument mit Blick auf praktische Entscheidungen zu optimieren. In der D-Studie wird simuliert, wie sich die Varianzaufklärung

verändern würde, wenn man die Anzahl der zulässigen Bedingungen innerhalb bestimmter Facetten variiert. Außerdem werden zwei Indizes zur Bewertung der Testgüte ermittelt. Der Generalisierbarkeitskoeffizient (G-Koeffizient) p^2 (3) entspricht der Definition des Reliabilitätskoeffizienten in der KTT und berechnet sich wie folgt:

$$p^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \quad (3)$$

Er setzt sich also aus der Varianz der universalen Werte der Personen (σ_ρ^2) in Beziehung zur Summe dieser Varianz und der relativen Fehlervarianz (σ_δ^2), also der Varianz aus den gemessenen Werten von mehreren Personen, zusammen. Daher wird der G-Koeffizient p^2 auch als relativer Fehlerkoeffizient bezeichnet, der vor allem als Grundlage für normorientierte Gruppenvergleiche dient, da er auf der Rangfolge (Relation) der untersuchten Personen basiert und nur die relative Fehlervarianz in die Berechnung mit eingeht. Bei einem genügend hohen G-Koeffizienten erlaubt das Instrument die Messung von Rangfolgen von Personen (z.B. dass ein bestimmter Schüler ein besseres Lernverhalten als eine andere Schülerin zeigt). Es ist jedoch noch nicht möglich, die Größe dieses Unterschieds zu bewerten.

Das zweite Gütemaß ist der Abhängigkeitsindex (D-Koeffizient) Φ (4). Für ihn gilt

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} \quad (4),$$

wobei σ_ρ^2 die Varianz der universalen Werte und σ_Δ^2 die absolute Fehlervarianz, also die Varianz aus mehreren gemessenen Werten der gleichen Person, darstellen. Er wird als absoluter Fehlerkoeffizient bezeichnet und dient als Grundlage für intraindividuelle Vergleiche, da nur die absolute Fehlervarianz in die Berechnung eingeht. Bei einem genügend hohen D-Koeffizienten erlaubt das Instrument also die zuverlässige Bestimmung der Veränderung einer Person über die Zeit (z.B. wie sehr sich das Lernverhalten einer Schülerin oder eines Schülers während einer Förderung verbessert). Beide Testgüteindizes liefern also Informationen darüber, wie reliabel eine Messung für relative (normorientiert innerhalb einer Gruppe von Personen; z.B. Schulklassen) und absolute (intraindividuell hinsichtlich eines spezifischen Kriteriums einer Person) Vergleiche ist. Damit eignet sich die G-Theorie hervorragend für die Entwicklung verlaufsdiagnostischer Instrumente, die eine Aussage hinsichtlich der individuellen Bezugsnorm zulassen.

Fragestellung

Die vorliegende Studie untersucht die Testqualität einer Direkten Verhaltensbeurteilung des Lern- und Arbeitsverhaltens von Schülerinnen und Schülern einer jahrgangsübergreifenden inklusiven Grundschule. Im Fokus des Interesses stehen die Interrater-Reliabilität und die interne Konsistenz der Items, sowie die Generalisierbarkeit und Zuverlässigkeit dieser Ergebnisse. Ausgehend von bisherigen Forschungsbefunden werden eine hohe Interrater-Reliabilität sowie eine hohe interne Konsistenz der verwendeten Items vermutet. Es wird erwartet, dass die Ergebnisse sowohl für normorientierte als auch für intraindividuelle Entscheidungen generalisierbar und zuverlässig sind.

Methode

Studiendesign

In der vorliegenden Studie wird eine Direkte Verhaltensbeurteilung des Lern- und Arbeitsverhaltens von zehn Schülerinnen und Schülern einer inklusiven Grundschulklasse anhand von Videobeobachtungen durch sechs geschulte Rater durchgeführt, die jeweils fünf Items pro Kind beurteilen. Es handelt sich um eine Generalisierbarkeitsstudie

mit einem vollständig gekreuzten Zwei-Facetten-Design mit den Facetten *Rater* und *Item*, d.h. jeder Rater beurteilt jedes Kind mit jedem Item. Die Facette *Rater* umfasst sechs geschulte Lehramtsstudentinnen im Hauptstudium. Da diese sechs Rater prinzipiell durch andere Rater (z.B. könnte man genauso gut Lehrkräfte als Rater wählen) ersetzt werden können, handelt es sich hierbei um eine zufällige Facette (Brennan, 2001; Eisend, 2007). In der Facette *Item* werden fünf durch ein standardisiertes universelles Verhaltensscreening (s.u.) ermittelte Items zum Lern- und Arbeitsverhalten genutzt. Auch hier ließen sich prinzipiell andere Items des Screenings bzw. Items aus anderen Screenings nutzen, so dass auch diese Facette als zufällig definiert wird. Zehn Schülerinnen und Schüler werden beobachtet. So ergeben sich insgesamt 300 Datenpunkte.

Stichprobe

Insgesamt wird das Lern- und Arbeitsverhalten von zehn Schülerinnen und Schülern (fünf Mädchen, fünf Jungen) einer inklusiven, jahrgangsübergreifenden Schulklasse beobachtet. Der Altersbereich liegt zwischen sieben und elf Jahren ($M = 8.3$, $SD = 1.34$, $MED = 8.5$). Die Auswahl der Schülerinnen und Schüler erfolgt mittels eines universellen Verhaltensscreenings (s.u.). Da es sich hierbei um eine Zufallsstichprobe vieler möglicher Bedingungen handelt (z.B. könnte man ebenfalls Kinder einer Förderschule oder einer nicht-inklusiven Schulklasse untersuchen) und die in dieser Studie erzielten Befunde über die vorliegende Stichprobe hinaus generalisiert werden sollen, werden sie in der Datenanalyse als zufällige Facette behandelt (Eisend, 2007).

Erhebungsinstrumente

Bei der *Integrated Teacher Rating Form* (ITRF) nach Volpe und Fabiano (2013) handelt es sich um ein universelles Verhaltensscreening, das spezifische Verhaltensprobleme in schulischen Settings fokussiert. Die ITRF gliedert sich in die zwei Subskalen „Störendes Verhalten“ und „Lern- und Arbeitsverhalten“ und umfasst 43 problemorientierte Items, die Schülerverhalten im Unterricht erfassen, wie z.B. „Does not complete classwork on time“, „Disrupt others“ oder „Moves around the room“. Die englischsprachige ITRF erfüllt die gängigen Testgütekriterien (Daniels, Volpe, Briesch & Fabiano, 2014). Die ITRF wurde ins Deutsche übersetzt und einzelne Items für den deutschen Kulturraum adaptiert. Das so überarbeitete Screening wurde dann von der Lehrerin der Klasse, in der die Videobeobachtungen durchgeführt wurden, für alle Schülerinnen und Schüler ausgefüllt. Die fünf Items aus dem Bereich „lernbezogenes Verhalten“, die in der Klasse die größten Probleme bereiteten, wurden für die vorliegende Studie ausgewählt. Die fünf Schülerinnen und Schüler mit dem problematischsten Verhalten und die fünf Schülerinnen und Schüler mit dem unproblematischsten Verhalten in diesen Items nahmen an der Studie teil.

Zur Einschätzung des Lern- und Arbeitsverhaltens wurde eine DVB mit den folgenden fünf ITRF-Items verwendet: „Arbeitet konzentriert an seinen Aufgaben“, „Befolgt Anweisungen“, „Beginnt Aufgaben selbstständig“, „Kontrolliert seine eigenen Aufgaben“ und „Beteiligt sich am Unterricht“. Zur Einschätzung wurde eine sechsstufige Skala (*0 = Verhalten tritt nie auf* bis *5 = Verhalten tritt immer auf*) genutzt. Die Items wurden aufgrund der Ergebnisse der ITRF ausgewählt.

Vorgehensweise bei den Videobeobachtungen

Es wurde eine Stillarbeitsphase im Mathematikunterricht von zehn Minuten gefilmt. Dafür wurden drei Kameras im Klassenraum so aufgestellt, dass der gesamte Raum zu beobachten war. Die Kameras wurden ca. drei Wochen vor Beginn der Aufnahmen in der Klasse positioniert, damit sich die Kin-

der an die Kameras gewöhnen konnten. Rote Lämpchen, die bei der Aufnahme leuchten, wurden abgeklebt. Die Klassenlehrerin schaltete die Kameras ein, bevor die Kinder in der Klasse waren. So wurde gewährleistet, dass sich die Schülerinnen und Schüler möglichst natürlich verhalten. Sechs geschulte Rater beobachteten dieses zehnminütige Video und schätzten das Lern- und Arbeitsverhalten der zehn Schülerinnen und Schüler im Anschluss anhand der DVB ein. Die Reihenfolge, in der die Rater die Kinder beobachten und bewerten sollten, wurde randomisiert. Die Schulung der Rater erfolgte in zwei Schritten: In einem ersten Schritt wurde ein englischsprachiges Online-Tutorial zur Anwendung von DVB absolviert. In einem zweiten Schritt wurde die Anwendung von Items aus der ITRF, die in der vorliegenden Studie nicht zum Einsatz kamen, an ausgewählten Videosequenzen, die nicht Gegenstand der vorliegenden Studie sind, geübt. Die Rater wurden dazu angehalten, die Videos pro Kind ohne Pause durchzusehen und nicht zurückzuspulen.

Datenanalyse

Auch wenn die G-Theorie keine Verteilungsannahmen voraussetzt, werden die vorliegenden Daten zunächst auf Normalverteilung geprüft, um das für die G-Studie am besten geeignete Schätzverfahren zu wählen. Die vorliegenden Daten sind laut KS-Test nicht normalverteilt ($M = 2.45$, $SD = 1.53$, $z = 2.60$, $p < .05$). Daher wird das *Minimum Norm Quadratic Unbiased*-Schätzverfahren (MINQUE), das keine Verteilungsannahmen voraussetzt, gewählt.

Die Varianzkomponentenschätzung erfolgt mittels einer mehrfaktoriellen Varianzanalyse, wobei sowohl Generalisierungsfacetten (Rater und Item) als auch Differenzierungsfacette (Person) als zufällige Faktoren definiert werden. Die Varianzkomponentenschätzung wird bei vollständig gekreuzten Designs durch die Erwartungswerte der mittleren Quadratsummen angegeben und ergibt sich durch die Summe der gewichteten Varianzkomponenten (Brennan, 2001). Demnach kann sie in diesem Studiendesign in sieben Komponenten unterteilt werden (Person, Rater, Item, Person x Rater, Person x Item, Rater x Item, Person x Rater x Item x Residuum).

In der Entscheidungsstudie (D-Studie) werden die Informationen aus der G-Studie für eine Optimierung des Instruments verwendet. Dies erfolgt durch die systematische simulierte Manipulation der verschiedenen Bedingungen einer Facette. In der vorliegenden Untersuchung wird die Anzahl der Items sowie die Anzahl der Rater systematisch manipuliert, um zu untersuchen, welche Bedingungen für eine Erhöhung der Generalisierbarkeit der Ergebnisse notwendig wären. Um die Generalisierbarkeit und Zuverlässigkeit der Ergebnisse zu beurteilen, werden sowohl der relative (p^2) als auch der absolute Fehlerkoeffizient (Φ) berechnet. Der kritische Wert, um von einer hohen Generalisierbarkeit und Zuverlässigkeit zu sprechen, wird bei .8 angelegt (Salvia, Ysseldyke & Bolt, 2010).

Ergebnisse

Ergebnisse der G-Studie

Die Ergebnisse der Varianzkomponentenschätzung zeigen, dass der Großteil der Varianz durch die Unterschiede zwischen den Kindern (Facette *Person* 49.6%) aufgeklärt wird (siehe Tabelle 1). Die Werte in der DVB werden also maßgeblich durch das situative Verhalten der Kinder erklärt. Die Varianzaufklärung durch die Items (2.5%) ist gering. Die unterschiedlichen Items repräsentierten das Zielkonstrukt demnach sehr ähnlich. Die Rater klären 4.7% der Varianz auf. Es gibt also kleine, wenn auch beachtenswerte Unterschiede zwischen den Beurteilungen der verschiedenen Rater über alle Schülerinnen und Schüler. Zu beachten ist der Interaktionseffekt zwischen Personen und Rater, der mit 17.0 % einen beträchtlichen Teil der Varianz aufklärt. Die Rater be-

Tabelle 1: Ergebnisse der Varianzanalyse sowie der prozentuale Anteil der Varianzkomponenten an der Gesamtvarianz

Variationsquelle	df	MS	Schätzung der Varianzkomponente	%	SE
Person p	9	40.06	1.24	49.6	0.57
Rater r	5	9.66	0.12	4.7	0.10
Item i	4	5.47	0.06	2.5	0.05
p x r	45	2.65	0.43	17.0	0.11
p x i	36	0.66	0.02	0.9	0.03
r x i	20	1.63	0.11	4.4	0.05
p x r x i, res	180	0.52	0.52	20.8	0.05
Total	**299**			**100**	

Anmerkungen: df = Freiheitsgrade, MS = Mittlere Quadratsummen, SE = Standardfehler

urteilen also das Verhalten bestimmter Kinder unterschiedlich. Die Interaktionen zwischen Person und Item (0.9%) sowie zwischen Rater und Item (4.4%) sind hingegen wieder gering, d.h. die Items funktionieren für alle Kinder und alle Rater ähnlich gut. Die durch unser Design nicht aufzuklärende Residualvarianz beträgt 20.8%. Der G-Koeffizient liegt bei $p^2 = .93$, der D-Koeffizient bei $\Phi = .91$.

Ergebnisse der D-Studie

Aufgrund der Ergebnisse aus der G-Studie, wo eine hohe Varianzaufklärung durch die Differenzierungsfacette *Person* sowie eine geringere Varianzaufklärung durch die Generalisierbarkeitsfacetten feststellbar ist, konzentriert sich die D-Studie auf eine Optimierung der Bedingungen der beiden Facetten *Item* und *Rater*, um zu überprüfen, ob ähnlich positive Ergebnisse bei einem ökonomischeren Einsatz von weniger Ratern und Items erzielt würden. Die Ergebnis-

Abbildung 1

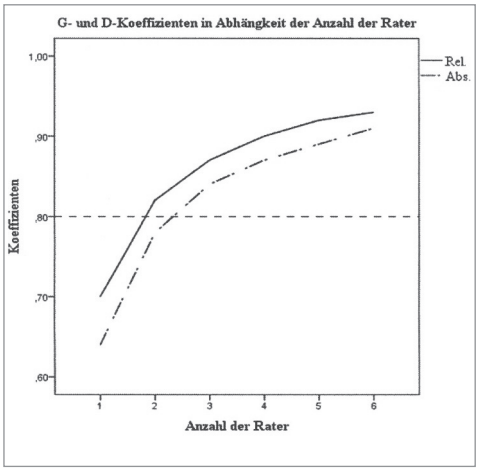
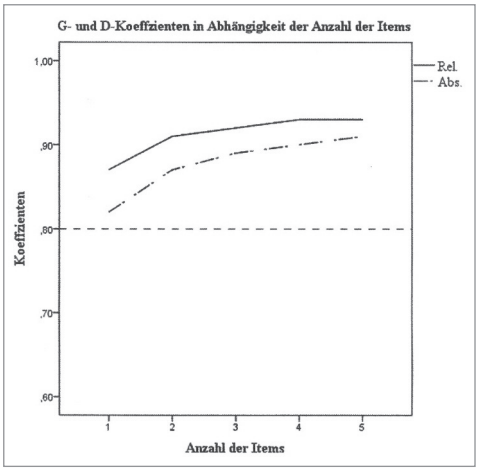


Abbildung 2



se hinsichtlich einer Variation innerhalb der Facette *Rater* zeigen, dass die Generalisierbarkeit und Zuverlässigkeit der Daten mit der Anzahl der Rater steigt (Abbildung 1). Der G-Koeffizient bei nur einem Rater beträgt $p^2 = .70$, der D-Koeffizient liegt bei $\Phi = .64$. Ein für Einzelfallentscheidungen genügendes Kriterium (.80) wird für relative Entscheidungen bei einem Einsatz von zwei Ratern ($p^2 = .82$) sowie für absolute Entscheidungen bei einem Einsatz von drei Ratern ($\Phi = .84$) erreicht. Die Ergebnisse der D-Studie hinsichtlich der Facette *Item* zeigen, dass die Generalisierbarkeit und Zuverlässigkeit der Daten bei weniger Items nur minimal geringer wird (Abbildung 2). Selbst bei der Verwendung eines einzigen Items ($p^2 = .87$, $\Phi = .82$) werden ähnlich hohe Werte wie bei der Verwendung von fünf Items ($p^2 = .93$, $\Phi = .91$) errechnet.

Diskussion

Die Ergebnisse zeigen, dass der Großteil der Varianz in den Messwerten durch die Unterschiede zwischen den Schülerinnen und Schülern aufgeklärt wird. Die Rater und die Items klären hingegen nur einen geringen Teil der Varianz auf. Unter Berücksichtigung der Ergebnisse einer Meta-Analyse von Hoyt und Kernes (1999), wonach selbst geschulte Rater im Durchschnitt eine Varianz von 10.0 % aufklären, liefern diese Befunde Evidenz für eine mehr als akzeptable Interrater-Reliabilität im Vergleich zu anderen Beurteilungsverfahren. In diesem Zusammenhang ist allerdings auch der hohe Interaktionseffekt zwischen Kindern und Ratern zu beachten, der 17.0% der Varianzaufklärung ausmacht. Dieser Effekt ist dahingehend zu bewerten, dass bestimmte Rater bestimmte Schülerinnen und Schüler unterschiedlich einschätzen. Wie das Verhalten eines Kindes beurteilt wird, hängt also davon ab, welche Person die Beurteilung vornimmt, obwohl die Rater intensiv geschult und die Beurteilung des Verhaltens der Personen randomisiert vorgenommen wurden. Mit Blick auf einen möglichen Halo-Bias bei Verhaltensbeurteilungen könnte jedoch gerade diese Randomisierung eine Erklärung für die unterschiedlichen Einschätzungen sein (Schmidt-Atzert & Amelang, 2012). So kann die Beurteilung eines Kindes mit stark problematischem Verhalten einen Einfluss auf die Bewertung des nächsten Kindes dahingehend haben, dass das Verhalten negativer eingeschätzt wird als es tatsächlich ist (und umgekehrt). Eine weitere Erklärung für die unterschiedlichen Ratings könnte auch die nicht immer optimale Qualität der Videos sein, so dass das Verhalten bestimmter Kinder nicht eindeutig zu beobachten war.

Insgesamt scheint es also, dass sich die Subjektivität von Verhaltensbeurteilungen als systematische Fehlerquelle (Schmidt-Atzert & Amelang, 2012) bei DVB trotz intensiver Schulung nicht ausschließen lässt. Dieser Befund deutet darauf hin, dass der Einsatz von DVB als prozessbegleitende Diagnostik zwar unbedenklich ist, solange konsistent die gleiche Person (z.B. eine Lehrkraft) beurteilt und die Veränderung des Verhaltens reliabel abgebildet wird (strukturelle Invarianz). Dies muss allerdings im deutschsprachigen Raum für die DVB noch nachgewiesen werden (z.B. Huber & Rietz, 2015). Im Rahmen einer G-Studie sollte daher unbedingt die Facette *Messzeitpunkt* mit mehreren engmaschigen Messungen mittels DVB berücksichtigt und deren Interaktion mit Ratern und Personen analysiert werden.

Die Ergebnisse aus der D-Studie weisen auf eine hohe Generalisierbarkeit ($p^2 = .93$) und eine hohe Zuverlässigkeit ($\Phi = .91$) der Befunde hinsichtlich der Facetten *Rater* und *Item* hin. Das Instrument scheint in der hier angewendeten Form also als Grundlage sowohl für relative und absolute Vergleiche geeignet, wie es bereits in der Untersuchung von Kilgus, Riley-Tillman, Chafouleas, Christ und Welsh (2014) berichtet wurde. Erfreulich ist, dass sowohl der G- ($p^2 = .82$) als auch der D-Koeffizient ($\Phi = .78$) bereits bei einer Anzahl von zwei

Ratern akzeptable Werte erreichen, was in Hinblick auf Co-Teaching-Modelle und deren Profit für inklusive Schulen realisierbar und wünschenswert scheint (Scruggs, Mastropieri & McDuffie, 2007). Eine Verringerung der Bedingungen der Facette *Item* zeigt, dass die Anzahl der Items nur einen sehr geringen Einfluss auf die Testgüte hat. Für den praktischen Einsatz ist dies ein positiver Befund, da die Testlänge sogar auf bis zu ein einziges Item ökonomisch reduziert und die Items flexibel auf die individuellen Bedürfnisse der Kinder und Jugendlichen abgestimmt werden können (Christ et al., 2009). In diesem Zusammenhang stellt sich allerdings die Frage, wie valide Messungen mit nur wenigen Items sind. Sicherlich könnte die Nutzung mit wenigen Items mit höherer Ungenauigkeit (z.B. bei einem global formulierten Item) bzw. Informationsverlust (z.B. bei wenigen spezifischen Items) einhergehen. Wenn das Ziel der Messung jedoch eine ökonomische und dennoch reliable Einschätzung beobachtbarer Verhaltensweisen in der Schule ist, kann die DVB genau dies leisten (Volpe, Briesch & Chafouleas, 2010). Dennoch stellt die psychometrische Überprüfung der Items – wie bei den hier genutzten Items in der US-amerikanischen Version des Instruments geschehen (Daniels et al., 2014) – eine wichtige Aufgabe in der Zukunft dar.

Insgesamt bleibt festzuhalten, dass die Direkte Verhaltensbeurteilung zur Erfassung von Lern- und Arbeitsverhalten aufgrund der akzeptablen Testgüte gut einsetzbar ist. Damit liefert die vorliegende Studie die erste deutschsprachige Replikation der positiven Forschungsbefunde aus dem nordamerikanischen Raum sowie die erste Studie im inklusiven Setting überhaupt in Bezug auf die praktische Eignung des Instruments als prozessdiagnostische Methode zur Erfassung von Entwicklungsverläufen. Vor allem der positive D-Koeffizient, der die Zuverlässigkeit der Messung im Rahmen intraindividueller Entscheidung angibt, deutet auf die Eignung für den verlaufsdiagnostischen Einsatz hin. Wenn also der Beobachtungszeitraum klar definiert und die Items gründlich operationalisiert sind, wie in der vorliegenden Untersuchung der Fall, können die interessierenden Verhaltensweisen zuverlässig erfasst werden und die Ergebnisse der Messung sowohl für normorientierte als auch für intraindividuelle Vergleiche genutzt werden. Damit kann angenommen werden, dass das in dieser Studie überprüfte Instrument für die regelmäßige Erfassung von Schülerverhalten und die Überprüfung des Erfolgs pädagogischer Handlungsmöglichkeiten eingesetzt werden kann. Es leistet damit einen entscheidenden Beitrag zur sonderpädagogischen Diagnostik und evidenzbasierten Handlungskonzepten (Bundschuh, 2010; Casale et al., 2015b; Hillenbrand, 2015). Grundsätzlich scheinen sich DVB zur Überprüfung von *Evidenzbasierung im Einzelfall* zu eignen, wo die Passung von Förderung zu den Lernbedürfnissen jedes Individuums einzeln überprüft wird. Trotzdem muss die Überprüfung der Veränderungssensitivität über häufige und engmaschige Messungen (z.B. durch den Einbezug der Facette *Messzeitpunkt* im Rahmen weiterer G-Studien) noch erfolgen (Chafouleas, Sanetti, Kilgus & Maggin, 2012).

Bei allen Vorteilen der Generalisierbarkeitstheorie in Bezug auf die Entwicklung von verlaufsdiagnostischen Instrumenten (v.a. simultane Berücksichtigung mehrerer relevanter Fehlerquellen, Überprüfung relativer und absoluter Entscheidungen, Simulation von Entscheidungsstudien zur Verbesserung des Instruments etc.) sind die Einschränkungen dieser methodischen Vorgehensweise zu berücksichtigen. Zum einen stellt die Schätzung der Varianzkomponenten und die Gefahr negativer Schätzungen ein Problem dar (Eisend, 2007). Diese Schätzfehler resultieren in der Regel aus für die Varianzanalyse zu kleinen Stichproben. Zur angemessenen Stichprobengröße bei der Anwendung der Generalisierbarkeitstheorie herrscht im wissenschaftlichen Diskurs jedoch noch Unklarheit und weitere Forschung hierzu ist unabdingbar (Briesch,

Swaminathan, Welsh & Chafouleas, 2014). Auf Grundlage einer Empfehlung von Webb, Rowley und Shavelson (1988) scheinen die in der vorliegenden Studie erreichten 300 Datenpunkte jedoch als hinreichend. Darüber hinaus wird auch immer wieder die Frage nach der Zulässigkeit der Aussagen über die untersuchten Facetten und deren Bedingungen hinaus aufgeworfen. Hier ist anzumerken, dass die Aussagekraft von Ergebnissen aus generalisierbarkeitstheoretischen Analysen – ähnlich denen aus Einzelfallanalysen – über die Konsistenz der Befunde steigt. Es liegt also auf der Hand, dass die Generalisierbarkeitstheorie keinen Ersatz sondern eine Erweiterung der KTT darstellt (Brennan, 2001). So gesehen, stellt sie einen entscheidenden und wichtigen Mehrwert für die Entwicklung von Instrumenten zur Verlaufsdiagnostik von Schülerverhalten dar.

Literaturverzeichnis

Brennan, R.L. (2001). *Generalizability Theory*. New York: Springer.

Briesch, A. M., Swaminathan, H., Welsh, M. & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology 52(1),* 13-35.

Bundschuh, K. (2010). *Einführung in die sonderpädagogische Diagnostik*. München: Reinhardt UTB.

Casale, G., Hennemann, T. & Grosche, M. (2015b). Zum Beitrag der Verlaufsdiagnostik für eine evidenzbasierte sonderpädagogische Praxis am Beispiel des Förderschwerpunktes der emotionalen und sozialen Entwicklung. *Zeitschrift für Heilpädagogik, 7*, 325-334.

Casale, G., Hennemann, T., Huber, C. & Grosche, M. (2015a). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung, 41*(1), 37-54.

Chafouleas, S. M., Sanetti, L. M. H., Kilgus, S. P. & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change across consultation cases using Direct Behavior Rating Single-Item Scales (DBR-SIS). *Exceptional Children, 78*, 491-505.

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 Intervention34 (1)*, S. 201-213.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons.

Daniels, B., Volpe, R. J., Briesch, A. M. & Fabiano, G. A. (2014). Development of a problem-focused behavioral screener linked to evidence-based intervention. *School Psychology Quarterly*.

Eisend, M. (2007). *Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung.* Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin, N0. 2007/4, ISBN 3938369523.

Grosche, M. (2014). Fördermaßnahmen im Prozess überprüfen. Das Konzept der Lernverlaufsdiagnostik. In T. Bohl, A. Feindt, B. Lütje-Klose, M. Trautmann & B. Wischer (Hrsg.), *Friedrich Jahresheft 2014 Fördern* [Themenheft].

Grosche, M. & Volpe, R. J. (2013). Response-to-intervention (RTI) as a model to facilitate inclusion for students with learning and behaviour problems. *European Journal of Special Needs Education 28 (3),* S. 254-269.

Hillenbrand, C. (2015). Evidenzbasierte Praxis im Förderschwerpunkt emotionale-soziale Entwicklung. In R. Stein & T. Müller (Hrsg.), *Inklusion im Förderschwerpunkt emotionale und soziale Entwicklung*. Stuttgart: Kohlhammer, S. 170-215.

Hoyt, W. T. & Kerns, M. D. (1999). Magnitude and Moderators of Bias in Observer Ra-

tings: A Meta-Analysis. In *Psychological Methods* (4), S. 403-424.

Huber, C. & Grosche, M. (2012). Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik, (08)*, 312-322.

Huber, C. & Rietz, C. (2015). Behavior Assessment Using Direct Behavior Rating (DBR) - A Study on the Criterion Validity of DBR Single-Item-Scales. *Insights into Learning Disabilities, 12*(1), 73-90.

Huber, C. & Rietz, C. 2015). Direct Behavior Rating (DBR) als Methode zur Verhaltensverlaufsdiagnostik in der Schule: Ein systematisches Review von Methodenstudien. *Empirische Sonderpädagogik, 7(*2), 75-98.

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*, 63-82.

Salvia, J., Ysseldyke, J. E. & Bolt, S. (2010). *Assessment in special and inclusive education, 11th Edition*. Boston, MA: Houghton Mifflin.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik*. Heidelberg: Springer.

Scruggs, T. E., Mastropieri, M. A. & McDuffie, K. A. (2007). Co-Teaching in Inclusive Classrooms: A Metasynthesis of Qualitative Research. *Exceptional Children 73(4)*, 392-416.

Volpe, R. J., Briesch, A. M. & Chafouleas, S. M. (2010). Linking Screening for Emotional and Behavioral Problems to Problem-Solving Efforts: An Adaptive Model of Behavioral Assessment. *Assessment for Effective Intervention, 35(4)*, 240-244.

Volpe, R. J. & Fabiano, G. A. (2013). *Daily behavior report cards: An evidence-based system of assessment and intervention*. New York: Guilford Press.

Webb, N. N., Rowley, G. L. & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development 21*, 81-90.

Wilbert, J. (2014). Instrumente zur Lernverlaufsdiagnostik: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 281-308). Göttingen: Hogrefe.

Gino Casale
Erziehungshilfe und sozial-emotionale Entwicklungsförderung
Department Heilpädagogik
Klosterstraße 79c
50931 Köln
gino.casale@uni-koeln.de