Lernverlaufsdiagnostik Schreiben (LVD – Schreiben): Reliabilität, Validität und Sensitivität für mittelfristige Lernfortschritte im deutschsprachigen Raum

Julia Winkes
Universität Freiburg (CH)

Pascale Schaller
Pädagogische Hochschule Bern

Zusammenfassung: Um das Repertoire an schreibdiagnostischen Instrumenten zu erweitern, wird das im englischen Sprachraum etablierte Instrument Lernverlaufsdiagnostik Schreiben (LVD – Schreiben, im englischsprachigen Raum CBM Writing) vorgestellt, das sowohl als Screening als auch zur Lernverlaufsdiagnostik eingesetzt werden kann. Die Testgütekriterien Paralleltestreliabilität und Validität sowie die Sensitivität für mittelfristige Leistungsverbesserungen werden in einer Untersuchung mit N = 354 Schülerinnen und Schülern der dritten bis sechsten Klassenstufe adressiert. Zu zwei Testzeitpunkten (Herbst und Frühjahr) verfassten die Teilnehmenden jeweils zehn dreiminütige Schreibproben. Abhängig von der Klassenstufe erweisen sich verschiedene Auswertungsmethoden als reliabel und valide, wobei die Nutzung des Medians von drei Schreibproben der Verwendung einer einzelnen Schreibprobe überlegen ist. In allen Klassenstufen und mit allen Auswertungsmethoden machten die Schülerinnen und Schüler signifikante Leistungsfortschritte zwischen Herbst und Frühjahr. Das Verfahren erweist sich somit als sensibel für Leistungsveränderungen innerhalb eines Schuljahres. Ein explorativer Vergleich mit amerikanischen Normwerten weist auf Gemeinsamkeiten und Unterschiede in der Entwicklung der Schreibflüssigkeit in beiden länderspezifischen Kontexten hin. Basierend auf den gewonnenen Erkenntnissen werden Implikationen für weiterführende Forschung diskutiert.

Schlüsselbegriffe: Lernverlaufsdiagnostik, Curriculumbasiertes Messen, Schreiben, Schreibflüssigkeit, Diagnostik, Testgütekriterien

Curriculum-Based Measurement in Writing: Reliability, Validity and Sensitivity to Growth of a German Progress Monitoring Instrument

Summary: With the aim to expand the toolkit of instruments for writing assessment, Curriculum-Based Measurement Writing (CBM Writing) is presented, which is already established in the English-speaking countries. It can be used for screening purposes as well as for progress monitoring. The technical features of alternate form reliability, validity and sensitivity to medium-term writing growth are addressed in a study with N = 354 students in third through sixth grades. At two time points (fall and spring), participants wrote ten three-minute writing samples each. Depending on the grade level, different scoring procedures proved to be reliable and valid, with the use of the median of three writing samples being superior to the use of a single writing sample. At all grade levels and with all scoring procedures, students made significant gains in performance between fall and spring. CBM writing thus proves to be sensitive to changes in performance within a school year. An exploratory comparison with American norm data points to similarities and differences in the development of writing fluency in both contexts. Implications for further research are discussed.

Keywords: Curriculum Based Measurement, writing, writing fluency, assessment, technical adequacy

1 Zielsetzung

Unter Lernverlaufsdiagnostik (LVD) werden kurze, wiederholt einsetzbare Messungen verstanden, mit denen Lernentwicklungen in wichtigen schulischen Bereichen (z. B. Lesen, Mathematik) begleitend dokumentiert und evaluiert werden können. Damit gehören sie zu den Instrumenten der formativen Diagnostik (vgl. Gebhardt, Jungjohann & Schurig, 2021). Der vorliegende Beitrag stellt mit der *Lernverlaufsdiagnostik – Schreiben* (LVD – Schreiben) ein Verfahren vor, welches im englischen Sprachraum seit vielen Jahren als Screening und als Mittel der lernprozessbegleitenden Diagnostik eingesetzt wird. Ziel ist es, dieses bereits etablierte Instrument ins Deutsche zu übertragen und in einer ersten Pilotstudie zu erproben. Im Rahmen einer Untersuchung mit Schülerinnen und Schülern der Klassenstufen 3 – 6 werden zu diesem Zweck die Testgütekriterien (Paralleltest-) Reliabilität und Validität für den Einsatz im deutschsprachigen Raum adressiert, ebenso wie die Sensitivität des Verfahrens für mittelfristige Leistungsverbesserungen. Schließlich interessiert auch ein erster explorativer Vergleich des Leistungsniveaus von Schweizer Schulkindern mit vorhandenen Referenzwerten aus dem angloamerikanischen Raum.

2 Problemaufriss: Schreibschwierigkeiten in der Schule und Desiderate der Schreibdiagnostik

Im Bereich Schreiben fokussieren vorhandene deutschsprachige Verfahren der LVD derzeit einzig auf die Rechtschreibung (Lernfortschrittsdiagnostik Rechtschreibung LDO; Lernlinie.de; Levumi.de). Das Beherrschen einer korrekten Rechtschreibung, welche in der Regel auf Wortebene überprüft wird, ist aber nur einer von vielen Teilaspekten der umfassenden Sprachhandlung des Schreibens, was sich auch in den Lehrplänen der deutschen Bundesländer und im Lehrplan 21 der Schweiz niederschlägt. Die Tatsache, dass Schreibprobleme sich bei vielen Schülerinnen und Schülern nicht nur in der Rechtschreibung äußern, sondern in umfangreicheren Schwierigkeiten bei der Planung, der Verschriftung und dem Überarbeiten von adressatengerechten und sprachformal korrekten Texten, spiegelt sich unter anderem in der Kategorie der „Störung des schriftlichen Ausdrucks“, die sowohl im DSM-5 (Falkai et al., 2018) als auch in der ICD-11 (WHO, 2021) dokumentiert wird. Schwache Schreiber verfassen in der Regel kurze Texte, die sich sowohl sprachformal als auch inhaltlich/qualitativ von den Texten gleichaltriger Peers unterscheiden (Saddler & Asaro-Saddler, 2013; Chenu, 2020). Wie McCloskey und Rapp (2017) herausarbeiten, wirken sich anhaltende Schwierigkeiten beim Verschriften von Texten negativ auf das gesamte schulische Lernen aus, weil Schreibaufgaben jeder Art für diese Kinder eine größere Anstrengung bedeuten und sie ihre volle Aufmerksamkeit auf basale Kompetenzen (Handschrift, Rechtschreibung, Abruf von passenden Formulierungen) richten müssen. Neben Schülerinnen und Schülern mit spezifischen Lernstörungen im Schreiben (Disorder of Written Expression) dokumentieren Schulleistungsstudien sowohl im deutsch- als auch im englischsprachigen Raum eine nicht unerhebliche Zahl von *struggling writers,* die zwar keine Diagnose einer Lernstörung erhalten haben, aber gleichwohl ausgeprägte Schwierigkeiten im Schreiben aufweisen und die Anforderungen der Schule in diesem Bereich nicht erfüllen (Chenu, 2020; Wanzek, Gatlin, Al Otaiba & Kim, 2017). Es ist daher überraschend, dass sich die Aufmerksamkeit von Fachpersonen in der Schule bislang noch wenig auf Schülerinnen und Schüler mit Schreibschwierigkeiten richtet. Eine mögliche Ursa-

che dafür könnte die vorhandene Diskrepanz zwischen der schulischen und beruflichen Bedeutung des Schreibens und den diagnostischen Möglichkeiten der Sonderpädagogik sein, Schülerinnen und Schüler mit Schreibschwierigkeiten überhaupt zu identifizieren und ihre Entwicklung lernverlaufsbegleitend zu dokumentieren.

Es existiert zum aktuellen Zeitpunkt kein deutschsprachiges Verfahren, um Schreiben über die Rechtschreibung hinaus standardisiert zu erfassen. Es bleibt daher nur der Rückgriff auf informelle Verfahren, welche aber mit verschiedenen Schwierigkeiten behaftet sind. Beispielsweise stellt die Zuverlässigkeit von Verfahren zur Schreibdiagnostik eine Herausforderung dar. Bedingt durch das komplexe Zusammenspiel unterschiedlicher Teilkompetenzen und durch das Vorhandensein mannigfaltiger Einflussfaktoren (z. B. Aufgabe, Tagesform, Vorwissen, Rater usw.) unterliegt Schreiben einer gewissen intraindividuellen Variabilität und scheint zudem auch genrespezifisch zu sein (Graham, Harris & Hebert, 2011; Wilson, Chen, Sandbank & Hebert, 2019). Idealerweise sollte sich Schreibdiagnostik daher auf mehrere Texte unterschiedlicher Genres beziehen, was zu Lasten der Ökonomie der Durchführung und Auswertung geht (Kim, Schatschneider, Wanzek, Gatlin & Al Otaiba, 2017 a). Eine weitere Schwierigkeit ergibt sich dadurch, dass es beim Schreiben keine vordefinierte „richtige Lösung" gibt. Als Antwort auf eine Schreibaufgabe oder einen Schreibimpuls wird auch eine Gruppe von Schülerinnen und Schülern mit ähnlichen Schreibfähigkeiten ein sehr großes Spektrum an unterschiedlichen Texten produzieren, die sich nicht nur sprachformal, sondern auch in Bezug auf Textlänge und Inhalt stark unterscheiden. Entsprechend kann es keine Auswertungskriterien geben, die eine Einschätzung als richtig/falsch vorsehen, und viele Beurteilende tun sich schwer damit, die Qualität von Schreibprodukten in Bezug auf einen geforderten Entwicklungsstand von Schülerinnen und Schülern objektiv zu analysieren (Ritchey et al., 2016). Im Bereich des summativen Schreibassessments bleiben also derzeit viele Wünsche offen. Formative Schreibdiagnostik als zweite Form der Beurteilung hat vor allem die Funktion, Lernverläufe in ihrem Prozess zu dokumentieren und Feedback für die Lernenden und Lehrenden bereitzustellen, welches sich positiv auf die weitere Lernentwicklung auswirkt (Klauer, 2014). Portfolios stellen eine solche Form der Diagnostik dar. Sie bestehen aus einer Sammlung von Verschriftungen verschiedener Genres, die über die Zeit von einem Schüler erstellt werden. Somit repräsentieren sie die Schreibkompetenz auf authentische Art und Weise und sind eine wichtige (qualitative) Quelle von Informationen für die weitere Förderung (Saddler & Asaro-Saddler, 2013). Ihre Stärke liegt vor allem in ihrem Potenzial für die Selbstreflexion der Lernenden und für die Identifikation von individuellen Ressourcen und Schwächen, welche in pädagogische Ziele umgewandelt werden können. Als Screeningverfahren eignen sie sich hingegen nicht (Traga Philippakos & FitzPatrick, 2018). Wie durch diesen kurzen Problemaufriss dokumentiert wird, ist das Spektrum von zur Verfügung stehenden Verfahren für die Schreibdiagnostik stark limitiert und vorhandene Materialien sehen sich mit Schwierigkeiten konfrontiert, die durch inhärente Eigenschaften des Schreibens selber bedingt werden (z. B. Integration verschiedener Teilprozesse, keine objektiv „richtige Antwort", hohe Variabilität innerhalb von Personen und zwischen Genres und Aufgaben).

3 LVD – Schreiben – ein Instrument für den Einsatz als Screening und zur Lernverlaufsdiagnostik

Unter dem Begriff „Lernverlaufsdiagnostik" oder „Curriculum-Basiertes Messen" (CBM) werden kurze und ökonomisch durchführba-

re Verfahren mit mehreren Proben gleicher Schwierigkeit bezeichnet, die wiederholt eingesetzt werden können. Eine wichtige Eigenschaft von LVD-Tests ist das Erfüllen der Testgütekriterien. Solche kurzen und direkten Beobachtungen von akademischen Outcomes wurden für verschiedene Bereiche konzipiert, beispielsweise für Lesen, Rechtschreiben, Mathematik und Schreiben (McMaster & Espin, 2007). Ihr Einsatz findet häufig – aber nicht ausschließlich – im Rahmen von Response-to-Intervention-Ansätzen (RTI) statt. Unter diesem Begriff subsummieren sich Rahmenkonzepte mit folgenden drei Kernelementen: Mehrebenenprävention, datengeleitete Förderentscheidungen (data based decision making) und Rückgriff auf evidenzbasierte Lehr- und Fördermethoden (Blumenthal, 2017). Im RTI-Paradigma haben LVD-Verfahren zwei Funktionen inne: Zum einen werden sie als *Universal Screenings* zwei- bis dreimal im Schuljahr mit allen Schülerinnen und Schülern der Klasse/Schule durchgeführt. Ziel dieser regelmäßigen Screenings ist es, Risikokinder mit ausbleibendem Lernerfolg frühzeitig zu identifizieren und ihnen niederschwellig eine zeitlich begrenzte Gruppenförderung auf Ebene 2 (oder später auf Ebene 3) anzubieten (Hosp & Kaldenberg, 2020). Die zweite Funktion von CBM-Verfahren besteht in der Lernverlaufsdiagnostik. Engmaschig (z. B. wöchentlich) eingesetzte LVD-Proben überwachen die Entwicklung von Kindern, die eine Zusatzintervention erhalten, und zeigen somit an, ob die Förderung den gewünschten Erfolg bringt. Fachpersonen erhalten dadurch ein direktes Feedback über die Passung zwischen Förderinhalt und -intensität und den Bedürfnissen des Kindes (Fuchs & Fuchs, 2017). Diese kurze Skizze lässt erahnen, dass durch die beiden wichtigen Hauptfunktionen und die Bedeutung der damit verknüpften pädagogischen Entscheidungen hohe Ansprüche an LVD-Verfahren gerichtet werden. Neben den traditionellen Gütekriterien, die für statusdiagnostische Tests relevant sind, soll LVD Veränderung adäquat über die Zeit modellieren, was weitere Voraussetzungen impliziert. Die Lernproben sollten nicht nur eine hohe Paralleltestreliabilität, sondern auch homogene Testschwierigkeiten aufweisen, damit der Schwierigkeitsgrad über die Zeit konstant bleibt. Zudem sollte das Instrument so änderungssensibel sein, dass auch kleine Lernfortschritte in kürzeren Zeiträumen sensitiv abgebildet werden (Allen et al., 2019; Förster, Kuhn & Souvignier, 2017; Strathmann, Klauer & Greisbach, 2010; Walter, 2013).

Auch in Bezug auf den Inhalt der Lernprobe gilt es, hohen Anforderungen gerecht zu werden: Mit kurzen Aufgaben ökonomisch Risikokinder zu entdecken und mit demselben Instrument Lernverläufe über einen längeren Zeitraum (Wochen, Monate oder Jahre) abzubilden, ist nur dann möglich, wenn die gemessene Leistung ein robuster und zuverlässiger Indikator für die gesamte schulische Domäne ist. Paradebeispiel für diese Indikatorfunktion von LVD ist die 1-Minute-Leseprobe (oral reading fluency), welche sich durch hohe Korrelationen mit umfassenden Leseassessments auszeichnet (Fuchs, 2004; 2017). Idealerweise erfordern LVD-Aufgaben die simultane Integration verschiedener Subskills, die auch für die Gesamtkompetenz in dem betreffenden schulischen Bereich benötigt werden. In der Domäne Schreiben fokussieren LVD-Tests auf das Konstrukt der Schreibflüssigkeit. Dieses kann am besten in Orientierung am weitläufig bekannten Begriff der Leseflüssigkeit verstanden werden: „Deno and Marston (2006) define fluent reading as the way that ‚an individual easily processes text and that the processing of text encompasses both word recognition and comprehension' (pp. 179 – 180). Applying this definition to fluent writing, we propose that it is the way an individual easily produces written text, and that the generation of written text encompasses both text generation (trans-

lating ideas into words, sentences, paragraphs, and so on) and transcription (translating words, sentences, and higher levels of discourse into print). Thus, fluent writing comprises the ease with which an individual both generates and transcribes text" (Ritchey et al., 2016, S. 27).

Schreibflüssigkeit beinhaltet somit zum einen das Formulieren von (mündlichem) Text und zum anderen Transkriptionsfähigkeiten, welche neben der Rechtschreibung auch die schreibmotorische Umsetzung mit dem Stift oder der Tastatur umfassen (vgl. Sturm, Nänny & Wyss, 2017). Damit gehört die Schreibflüssigkeit zu den hierarchieniedrigen Schreibprozessen, die von den hierarchiehohen Prozessen des Strategiegebrauchs, Planens und Überarbeitens abgegrenzt werden. Der Erwerb von Schreibflüssigkeit und die zunehmende Automatisierung der beteiligten Prozesse ist eine wichtige Entwicklungsaufgabe für Schülerinnen und Schüler der unteren Klassenstufen (und darüber hinaus), weil dadurch kognitive Ressourcen für strategiegeleitete hierarchiehohe Schreibkomponenten freigesetzt werden (Kim, Gatlin, Al Otaiba & Wanzek, 2017 b).

LVD – Schreiben bietet Lehrpersonen und Schülerinnen und Schülern ein direktes Feedback über die Entwicklung der Geschwindigkeit und Genauigkeit der hierarchieniedrigen Komponenten. Da die Schreibflüssigkeit und ihre verschiedenen Teilprozesse insbesondere bei schwachen Schreibern wichtige Ansatzpunkte einer spezifischen Schreibförderung darstellen, erhalten Lehrpersonen und Förderlehrkräfte durch LVD – Schreiben eine quantifizierbare Rückmeldung über den Erfolg ihrer Förderbemühungen (Traga Philippakos & FitzPatrick, 2018). Diese kann für datenbasierte Entscheidungen zum weiteren Vorgehen genutzt werden. Beispielsweise wird empfohlen, für Schülerinnen und Schüler mit Schreibschwierigkeiten Förderziele im Bereich der Schreibflüssigkeit zu formulieren und das Erreichen dieser Ziele durch LVD – Schreiben zu überprüfen (Hessler & Konrad, 2008; Poch et al., 2021).

3.1 Vorgehen bei der Erhebung und Auswertung von LVD-Schreibproben

Die prototypische Aufgabenstellung für LVD – Schreiben besteht aus einem einleitenden Satz (Storystarter), der als Schreibimpuls für das Verfassen eines Textes genutzt wird. Die Schülerinnen und Schüler erhalten den Storystarter vorgedruckt auf einem Blatt, ergänzt durch Linien für den zu schreibenden Text. Sie haben eine Minute Zeit für die Planung/Ideengenerierung und im Anschluss drei oder fünf Minuten für das Verschriften der Geschichte. Beispiele für mögliche Storystarter sind „Eines Tages ging ich von der Schule nach Hause, als plötzlich …" oder „Letzte Nacht konnte ich nicht einschlafen, weil …" (Hosp, Hosp & Howell, 2016; Ritchey et al., 2016). Hosp et al. (2016) empfehlen die folgende standardisierte Instruktion (Übersetzung JW): „Heute möchte ich, dass ihr eine Geschichte schreibt. Ich werde Euch einen Satz vorlesen und dann schreibt ihr eine kurze Geschichte darüber, was passiert. Ihr werdet eine Minute zum Nachdenken haben und drei Minuten zum Schreiben. Wenn ihr nicht wisst, wie man ein Wort schreibt, dann sollt ihr raten. Gibt es Fragen? Legt eure Stifte weg und hört zu. In der nächsten Minute denkt euch eine Geschichte aus über …". Der Testleiter benötigt einen Timer, um die Zeitvorgaben für das Planen und Schreiben präzise einzuhalten. Die hier skizzierte Vorgehensweise kann sowohl mit einzelnen Schülerinnen/Schülern als auch mit einer Kleingruppe oder Klasse durchgeführt werden. Wie es bei LVD-Verfahren üblich ist, wird für den ersten Einsatz von LVD – Schreiben (als Screening oder als Baseline für Lernverlaufsdiagnostik) die zeitnahe Durchführung von drei Schreibproben empfohlen, und der Ausgangswert besteht dann aus dem Median der drei Tests (ebd.).

Um einen ökonomischen Einsatz zu gewährleisten, sollte nicht nur die Durchführung, sondern auch die Auswertung zeitsparend sein. Dafür stehen verschiedene Methoden zur Verfügung, welche in Tabelle 1 anhand von Beispielen konkretisiert werden (vgl. Hosp et al., 2016; Hosp & Kaldenberg, 2020; Ritchey et al., 2016; Saddler & Asaro-Saddler, 2013). Malecki und Jewell (2003) geben die benötigte Zeit zur Bewertung einer Schreibprobe mit 1,5 bis 2,5 Minuten an. Der Zeitaufwand kann durch den Einsatz von kleinen Hand- oder Fingerzählern reduziert werden, die das Auszählen der Wörter und Sequenzen ökonomisieren.

Tab. 1 Auswertungsmethoden für die Lernverlaufsdiagnostik Schreiben

Methode	Beschreibung	Beispiele
Anzahl der geschriebenen Wörter (Total Words Written; TWW)	Gezählt wird die Anzahl der Wörter, die innerhalb des Zeitlimits geschrieben wurden. Rechtschreibung, Grammatik oder Semantik werden nicht beachtet. Auch sinnlose Buchstabenfolgen werden als Wörter gezählt.	Die Tomaten sind grün = 4 TWW Die tckttk sint krün = 4 TWW
Anzahl der orthografisch korrekt geschriebenen Wörter (Words Spelled Correctly, WSC)	Es wird die Anzahl der orthografisch korrekt verschrifteten Wörter gezählt, unabhängig vom konkreten Satzkontext. Nomen und Satzanfänge müssen großgeschrieben werden. Inkorrekte (d. h. in dieser Schreibweise nicht existierende) Wörter werden unterstrichen. Zur Berechnung des Maßes WSC wird die Anzahl der unterstrichenen Wörter von den TWW subtrahiert.	Die Tomaten sind krün = 3 WSC Ich weiß, das die Tomaten krün sind = 6 WSC (Das Wort „das" ist zwar im Satzkontext falsch, existiert in dieser Schreibweise als Wort der deutschen Sprache. Daher wird es als korrekt gewertet.)
Anzahl der korrekten und inkorrekten Schreibsequenzen (Correct/Incorrect Writing Sequences, CWS und IWS)	Eine Schreibsequenz existiert zwischen zwei Wörtern, zwischen einem Wort und einem Satzzeichen (Punkt, Komma, Ausrufe- und Fragezeichen, Doppelpunkt) und zu Beginn eines neuen Satzes. Es muss die Rechtschreibung (inkl. Groß- und Kleinschreibung), die Grammatik und die Semantik des Wortes richtig und kontextangepasst sein. Wir empfehlen für die deutsche Sprache, Kommata (nicht aber Satzzeichen zur Markierung wörtlicher Rede) bei der Bewertung einzubeziehen. Korrekte Schreibsequenzen werden mit einem Zirkumflex (^) gekennzeichnet, inkorrekte Schreibsequenzen mit einem umgedrehten Zirkumflex ($_{v}$). Fehlende Wörter oder Satzzeichen werden mit einem doppelten $_{v}$ markiert.	^ Ich ^ weiß ^, $_{v}$ das $_{v}$ die ^ Tomaten $_{v}$ krün $_{v}$ sind ^. = 5 CWS; 4 IWS ^ Ich ^ weiß ^, $_{v}$ das $_{v}$ die ^ Tomaten $_{v\,v}$ sind ^. = 5 CWS; 4 IWS
Anzahl der geschriebenen Wörter plus Korrekte Minus Inkorrekte Schreibsequenzen (TWW + CIWS)	Campbell et al. (2013) schlagen vor, die Anzahl der geschriebenen Wörter mit dem Maß CIWS additiv zu verbinden, um demotivierende negative Werte mit der Methode CIWS zu vermeiden. Vorteil dieser Methode ist, dass die Textmenge neben der Korrektheit (Accuracy) in den Wert einbezogen wird.	^ Ich ^ weiß ^, $_{v}$ das $_{v}$ die ^ Tomaten $_{v\,v}$ sind ^. = 7 TWW + CIWS (6 TWW + 5 CWS – 4 IWS)

Die verschiedenen zur Verfügung stehenden Auswertungsmethoden finden ihren Einsatz je nach Klassenstufe und individuellem Entwicklungsziel. Jewell und Malecki (2005) teilen die Methoden in drei verschiedene Kategorien ein: Produktionsabhängige Methoden (TWW, WSC, CWS) sind abhängig von der Menge des geschriebenen Textes. Produktionsunabhängige Methoden fokussieren auf die Schreibgenauigkeit (Accuracy), da sie unabhängig von der Länge der Schreibprobe sind. Beispiele sind Prozentangaben, wie z. B. %WSC oder %CWS. McMaster und Espin (2007) raten von der Verwendung von Prozentindizes ab, da diese nicht intervallskaliert sind. Auch bei der Nutzung im Rahmen von Lernfortschrittsdiagnostik ergeben sich Schwierigkeiten (z. B. könnte ein Schüler im Herbst 10 von 20 Wörtern korrekt schreiben und im Frühjahr 20 von 40 Wörtern, was sich im Maß %WSC nicht niederschlägt). Die dritte Gruppe der sogenannten „Accurate-production indices" messen gleichzeitig Schreibmenge und Genauigkeit. CIWS und TWW + CIWS gehören zu den Auswertungsmethoden dieser Kategorie.

3.2 Testgütekriterien

Die im Folgenden referierten Befunde zu den Testgütekriterien von LVD – Schreiben stammen alle aus dem englischsprachigen Raum und beschränken sich auf Studien, welche die Altersstufe der 3. – 6. Klasse umfassen. Untersuchungen zu LVD-Beginning Writing (Kindergarten bis 2. Klasse) greifen in der Regel auf andere Erhebungsmethoden für die Schreibproben zurück und werden daher hier ausgeklammert. Auch Studien mit älteren Schülerinnen und Schülern werden nicht in den nachfolgenden Überblick einbezogen, da es überzeugende Hinweise darauf gibt, dass Reliabilität und Validität der einzelnen Auswertungsmethoden sich zwischen den Klassenstufen deutlich unterscheiden (siehe z. B. Weissenburger & Espin, 2005).

3.2.1 Reliabilität

Wie in der Einleitung bereits ausgeführt, stellt Reliabilität in der Schreibdiagnostik ein allgegenwärtiges Problem dar, bedingt durch die vielen Facetten des Schreibens, die zu Fehlervarianz über den „wahren Wert" der Schreibkompetenz einer Person hinaus beitragen (Wilson et al., 2019). Standardisierte statusdiagnostische Schreibassessments erreichen im angloamerikanischen Raum in der Regel Werte für die Paralleltest- und die Test-Retest-Reliabilität zwischen r = .70 und .90 (Allen et al., 2019). Das *National Center for Intensive Intervention* schlägt für die Entwicklung von LVD-Verfahren unterschiedliche Standards abhängig vom Ziel des Einsatzes vor. Für die Funktion als Screening sollte ein LVD-Test eine Reliabilität von mindestens .80 aufweisen, für die Funktion im Rahmen von Lernverlaufsdiagnostik hingegen wird ein Wert über .70 angesetzt (Hosp & Kaldenberg, 2020). McMaster und Espin (2007, S. 69) diskutieren die Frage nach angemessenen Reliabilitätskoeffizienten speziell von LVD – Schreiben und bewerten in diesem Rahmen Werte von r > .80 als *relatively strong,* r = .70 bis .80 als *moderately strong,* r = .60 bis .70 als *moderate* und Werte unter r = .60 als *weak.*

Vorhandene Untersuchungen zur Reliabilität von LVD – Schreiben beziehen sich in der Regel auf Paralleltestreliabilität und Interrater-Reliabilität. Die berichteten Werte zur Interrater-Reliabilität befinden sich durchgehend in einem hohen Bereich (z. B. 91 – 100 % bei Gansle et al., 2004; 88 – 99 % bei Keller-Margulis, Payan, Jaspers und Brewton, 2016 a; > .98 bei Malecki und Jewell, 2003, und bei Jewell und Malecki, 2005). Dabei ist allerdings zu beachten, dass Rater in Forschungsprojekten häufig ein ausführliches Training durchlaufen, welches in der Schulpraxis nicht zwingend der Fall ist. Aus diesem Grund regen Allen et al. (2019) an, in zukünftigen Studien die Übereinstimmung zwischen den Ratings von Lehrpersonen zu erheben, um die Eigenschaft von LVD – Schreiben als zuverlässiges Instrument in der Schule

Tab. 2 Studien zur Paralleltestreliabilität von LVD – Schreiben in den Klassenstufen 3–6

Studie	Klassenstufe	Paralleltestreliabilität
Allen et al. (2019)	3	TWW: .77 (.50–.91) WSC: .76 (.51–.94) CWS: .73 (.43–.92) CIWS: .66 (.31–.92)
Weissenburger & Espin (2005)	4	TWW: .80 CWS: .79 CIWS: .73
Gansle, Noell, Vanderheyden, Naquin und Slider (2002)	3–4	TWW: .62 WSC: .53
Gansle, Vanderheyden, Noell, Resetar und Williams (2006)	1–5	TWW: .80 WSC: .82 CWS: .78
Campbell, Espin & McMaster (2013)	High School English Learners	TWW+CIWS: .82

Anmerkung: Die Berechnung der durchschnittlichen Reliabilität in der Untersuchung von Allen et al. 2019 erfolgte durch die Autorin JW.

besser einschätzen zu können. Die oben dargestellten Auswertungskonventionen können in der Tat beim Beurteilen von CWS und IWS schnell eine hohe Komplexität erlangen, beispielsweise bei der Nutzung von wörtlicher Rede oder dem Gebrauch von verschachtelten, aber fehlerhaften sprachlichen Strukturen.

Untersuchungen zur Paralleltestreliabilität von LVD – Schreiben in den hier fokussierten Klassenstufen (3.–6. Klasse) gibt es nur wenige. Tabelle 2 bietet einen Überblick über die vorhandenen Angaben, welche sich überwiegend im moderaten bis hohen Bereich befinden. Der kritische Wert von .70 wird in der Regel überschritten. In der Untersuchung von Allen et al. (2019) werden alle Korrelationen zwischen den durchgeführten Paralleltests einzeln berichtet. Dabei fällt die große Variabilität der Werte auf (z. B. zwischen .31 und .92 für CIWS). Für das Maß TWW + CIWS sind bislang nur Paralleltestreliabilitäten für Highschool English Learners verfügbar. Diese liegen mit .82 im hohen Bereich (Campbell, Espin & McMaster, 2013).

Wie McMaster und Espin (2007) anmerken, stellt eine geringe Reliabilität ein Problem dar, insbesondere wenn LVD – Schreiben als Screening-Verfahren genutzt wird, um schwache Schreiber zu identifizieren. Aus diesem Grund wird empfohlen, mehrere Schreibproben zu aggregieren, was zu stabileren Einschätzungen führt (Fuchs, Deno & Marston, 1982).

3.2.2 Validität

Studien zur Validität von LVD – Schreiben stützen sich überwiegend auf den Aspekt der Kriteriumsvalidität und überprüfen entsprechend den Zusammenhang zwischen LVD – Schreiben und einem umfangreicheren Schreibmaß (z. B. einem standardisierten Schreibtest, aber auch Lehrerratings) (McMaster & Espin, 2007). Typischerweise berichten Untersuchungen zu LVD – Schreiben deutlich niedrigere Validitätswerte als zu LVD – Lesen. Der Grund dafür wird allgemein in der Schwierigkeit gesehen, zu definieren, was „gutes Schreiben" überhaupt ist. Hinzu kommt das komplexe Zusammenspiel

von Ideengenerierung, Übersetzung in eine sprachliche Form, Transkribieren, Überarbeiten, Strategieanwendung, Monitoring, motivationalen Prozessen, Arbeitsgedächtnis und exekutiven Funktionen, was eine exakte und zuverlässige Erhebung von Schreibkompetenz erschwert. Entsprechend ist das Problem von geringer Validität in der Schreibdiagnostik allgemein nicht unbekannt und tritt auch bei standardisierten statusdiagnostischen Tests und somit bei den potenziellen Kriteriumsmaßen für LVD – Schreiben selber auf (Parker, McMaster & Burns, 2011). McMaster und Espin (2007) halten fest, dass die Validitätskoeffizienten von LVD – Schreiben zwar geringer sind als in anderen LVD-Bereichen, aber immer noch vergleichbar hoch oder gar besser als die von üblicherweise genutzten schreibdiagnostischen Verfahren. In Anbetracht der hier skizzierten Schwierigkeiten in der Schreibdiagnostik und der ohnehin reduzierten Anzahl von verfügbaren diagnostischen Instrumenten für die Schreibkompetenz (vgl. Kapitel 1), besteht im englischen Sprachraum ein allgemeiner Konsens über den Schwellenwert von r = .50 als minimale Voraussetzung für die Kriteriumsvalidität von LVD – Schreiben (Keller-Margulis, Mercer & Matta, 2021; McMaster et al., 2011).

Einen umfassenden Überblick über die Validität verschiedener Auswertungsmethoden von LVD – Schreiben bietet die Metaanalyse von Romig, Therrien und Lloyd (2017), welche die Resultate von 22 Studien zusammenfasst. Über alle Klassenstufen hinweg betrachtet, erreichte TWW eine Korrelation mit anderen schreibbezogenen Konstrukten von r = .37, WSC von r = .44, CWS von r = .51. Die höchsten Werte werden für die Methode CIWS (r = .60) dokumentiert. Die Autoren unterteilen die Studien anschließend noch weiter in verschiedene Alterskategorien. In den Klassenstufen 3 – 5 weist lediglich das Maß CIWS eine zufriedenstellende Korrelation mit anderen Schreibmaßen auf (r = .65), während die anderen drei Auswertungsmethoden unter r = .50 bleiben (TWW: .34, WSC: .35; CWS: .48). Auch für die älteren Schülerinnen und Schüler aus den Stufen 6 – 8 ist CIWS die Methode mit der höchsten Kriteriumsvalidität (.59) und somit der Auswertung via CWS (r = .50) vorzuziehen (TWW: r = .32 und WSC: r = .39).

Während einige Autoren von der Nutzung der Methoden TWW und WSC allgemein abraten, da sie die niedrigsten Zusammenhänge mit Außenkriterien der Schreibkompetenz zeigen, weisen andere Untersuchungen hingegen darauf hin, dass diese produktionsabhängigen Auswertungsmethoden zu Beginn der Schreibentwicklung, also bei jüngeren Schülerinnen und Schülern, durchaus eine adäquate und zeitsparende Alternative sein können (Jewell & Malecki, 2005; Weissenburger & Espin, 2005). In höheren Klassenstufen werden durchgehend die etwas aufwendigeren, aber valideren Maße CWS und CIWS empfohlen (Amato & Watkins, 2011; Campbell et al., 2013; McMaster & Campbell, 2008). Wie bereits im Abschnitt zur Reliabilität angesprochen, könnte die Verwendung mehrerer Schreibproben (z. B. Median aus drei Schreibproben) notwendig sein, um eine zuverlässige und valide Einschätzung der Schreibkompetenz von Schülerinnen und Schülern zu erhalten (Ford & Kaldenberg, 2019; Keller-Margulis et al., 2021; McMaster & Espin, 2007). Hinweise zur Validität der Auswertungsmethode TWW + CIWS liefert wiederum nur die Studie von Campbell et al. (2013) mit der speziellen Population der Highschool English Learners. Diese lag abhängig vom Kriteriumsmaß zwischen .54 und .74.

3.2.3 Änderungssensitivität

Werden LVD-Verfahren im Kontext von Lernverlaufsdiagnostik eingesetzt, so ist die Sensitivität für kurz- und mittelfristige Leistungsveränderungen eines der wichtigsten Charakteristika des Instruments. Um ihre Feedbackfunktion und ihr pädagogisches Potenzial im Rahmen von datenbasierten Förderentscheidungen

ausschöpfen zu können, sollten auch kleine Lernfortschritte sichtbar gemacht werden, insbesondere da die Hauptzielgruppe (Schülerinnen und Schüler mit Lernschwierigkeiten) häufig einen verlangsamten Lernfortschritt zeigt (Allen et al., 2019; Ritchey et al., 2016). Änderungssensitivität unterscheidet sich von den oben präsentierten Merkmalen (Reliabilität, Validität), weil sie kein Charakteristikum der einzelnen Messungen (technical features of the static score) darstellt, sondern eines von mehreren Gütekriterien des Lernverlaufs (technical features of the slope) (Fuchs, 2004).

Die Änderungssensitivität von Testverfahren kann auf verschiedene Arten überprüft werden. Im Bereich LVD – Schreiben wurden Fragen nach diesem Aspekt bislang häufig über den Vergleich mehrerer Messzeitpunkte im Schuljahr adressiert, was also die Veränderung in mittelfristigen Zeitabständen umfasst. Es zeigen sich übereinstimmend signifikante Lernzuwächse, wenn die Leistungen im Herbst und Frühling eines Schuljahres verglichen werden (Dockrell, Connelly, Walter & Critten, 2015; Malecki & Jewell, 2003; McMaster & Campbell, 2008). Auch über ein Intervall von zwei Monaten konnte in der Studie von Ritchey und Coker (2013) die Änderungssensitivität von LVD – Schreiben aufgezeigt werden. Bislang gibt es nur wenige Untersuchungen, die wöchentliche Datenerhebungen mit LVD – Schreiben umfassen. In der hier interessierenden Altersstufe (3. – 6. Klasse) ist vor allem die Studie von McMaster et al. (2017) erwähnenswert. Sie zeigte, dass – analog zu Reliabilität und Validität – auch die Änderungssensitivität in Abhängigkeit der Klassenstufe und der verwendeten Auswertungsmethode variiert. In den Stufen zwei und drei konnten signifikante Lernzuwächse mit der Methode CIWS schon nach drei Wochen entdeckt werden, mit der Methode CWS aber erst nach acht Wochen. Bei den Viert- und Fünftklässlern waren beide Methoden innerhalb von fünf Wochen änderungssensitiv.

LVD – Schreiben differenziert zwischen verschiedenen Klassenstufen (Allen et al., 2019; Ritchey & Coker, 2013), was ebenfalls als Hinweis auf Änderungssensitivität interpretiert werden kann (McMaster & Espin, 2007). Es gibt für den englischen Sprachraum Normwerte für die Klassenstufen Kindergarten bis Klasse 8 (Messzeitpunkte Herbst – Winter – Frühjahr) für die Maße TWW, WSC und CWS (publiziert in Hosp et al., 2016). Aus diesen Referenzwerten lässt sich ablesen, wie viel Lernzuwachs in den entsprechenden Klassenstufen erwartet wird. Beispielsweise erreichen Schülerinnen und Schüler mit einem Prozentrang von 50 in Klassenstufe 3 im Herbst 18 CWS, im Winter 24 CWS und im Frühjahr 30 CWS pro dreiminütiger Schreibprobe. Für den deutschen Sprachraum existieren bislang keinerlei Anhaltspunkte zur Entwicklung der Schreibflüssigkeit.

4 Empirische Studie

4.1 Fragestellungen

1) Wie hoch ist die Paralleltestreliabilität von LVD – Schreiben in den Klassenstufen 3 – 6 unter Einsatz der Scoring-Methoden TWW, CWS, CIWS und TWW + CIWS für eine einzelne Schreibprobe und für den Median aus drei Schreibproben?
2) Wie hoch ist die Kriteriumsvalidität von LVD – Schreiben in den Klassenstufen 3 – 6 unter Einsatz der Scoring-Methoden TWW, CWS, CIWS und TWW + CIWS für eine einzelne Schreibprobe und für den Median aus drei Schreibproben?
3) Zeigen sich zwischen den Erhebungszeitpunkten im Herbst und im Frühjahr signifikante Lernfortschritte?
4) Stimmt die durchschnittliche Leistung der Schweizer Schülerinnen und Schüler mit den Orientierungswerten aus dem englischen Sprachraum von Hosp et al. überein?

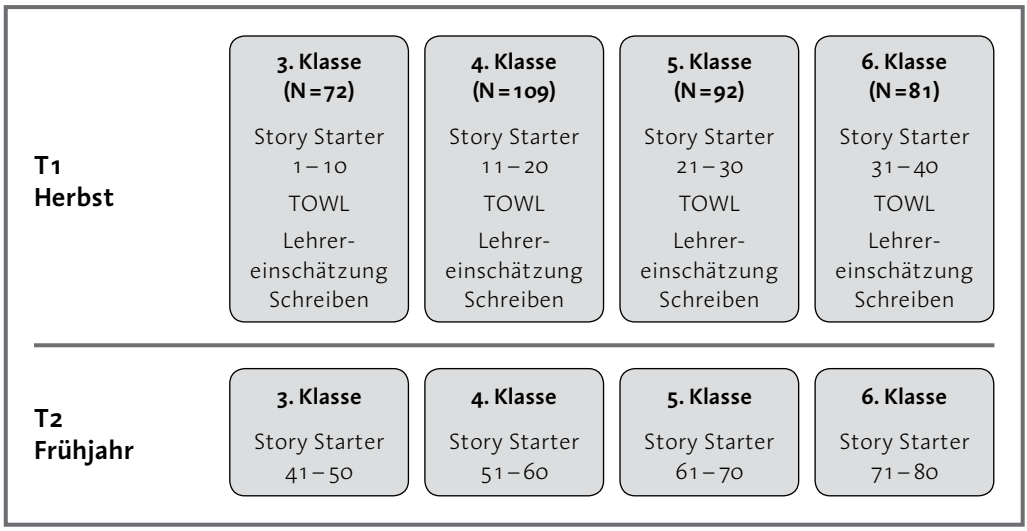

Abb. 1 Übersicht über die Testinstrumente und Datenerhebungszeitpunkte

4.2 Methode

4.2.1 Design, Stichprobenrekrutierung und -merkmale

Die Fragestellungen wurden im Rahmen einer Studie mit N = 354 Schülerinnen und Schülern der 3. bis 6. Primarklasse adressiert. Die Klassen stammen aus verschiedenen Schulen in den Kantonen Bern und Luzern. Die Rekrutierung der Klassen erfolgte durch Masterstudierende der Schulischen Heilpädagogik. Die Zustimmung der Schulleitung, der Lehrperson, der Eltern und der Schülerinnen und Schüler selber war Bedingung für die Teilnahme an der Untersuchung.

Wie in Abbildung 1 ersichtlich, umfasste die Untersuchung zwei Erhebungszeitpunkte (Herbst, Frühjahr). Zu jedem Zeitpunkt verfassten die teilnehmenden Schülerinnen und Schüler zehn LVD-Schreibproben innerhalb von maximal zehn aufeinanderfolgenden Schultagen mit maximal zwei LVD-Proben pro Tag. Innerhalb des ersten Messzeitpunktes im Herbst wurden zudem weitere Daten erfasst, wie die Schreibkompetenz (TOWL-4) und die Lehrereinschätzung der Schreibleistungen, welche als Kriterien für die Einschätzung der Validität fungieren.

Tab. 3 Stichprobe

Stufen	Schüler N	Alter in Mt. M (SD)	Geschlecht ♀ : ♂ (in %)	Nicht Deutsch als Muttersprache (in %)	Sondermaßnahme (in %)
3. Klasse	72	106.4 (6.3)	55.6 : 44.4	36.1	43.1
4. Klasse	109	119.2 (6.7)	57.8 : 42.2	32.1	18.3
5. Klasse	92	134.3 (7.6)	40.2 : 59.8	44.6	26.1
6. Klasse	81	143.9 (6.1)	51.9 : 48.1	42.0	22.3
Gesamt	354	126.16 (15.1)	51.4 : 48.6	38.4	31.9

Tabelle 3 zeigt die Merkmale der Stichprobe auf. In den vier Klassenstufen nahmen zwischen 72 (3. Klasse) und 109 (4. Klasse) Schülerinnen und Schüler an der Erhebung teil. Das Durchschnittsalter in der Gesamtstichprobe betrug 10;6 Jahre. Das Geschlechterverhältnis ist insgesamt gesehen ausgeglichen, wobei in einzelnen Klassenstufen ein Ungleichgewicht zugunsten der Mädchen (4. Klasse) bzw. zugunsten der Jungen (5. Klasse) besteht. Etwa 40 % der teilnehmenden Kinder haben eine andere Muttersprache als Deutsch. Über die Länge des Sprachkontaktes mit der Schulsprache sind keine weiteren Informationen vorhanden. Unter dem Begriff Sondermaßnahme werden zusätzliche Förderangebote im Bereich Sonderpädagogik (z. B. integrative Förderung, reduzierte individuelle Lernziele), Logopädie oder Deutsch als Zweitsprache zusammengefasst. Ein Drittel der Gesamtstichprobe erhielt eine solche Fördermaßnahme.

4.2.2 Instrumente

LVD-Schreiben: Die Schreibproben wurden in enger Anlehnung an das im englischsprachigen Raum empfohlene Vorgehen durchgeführt. Den Schülerinnen und Schülern wurde jeweils ein Story Starter als Schreibimpuls präsentiert. Nach einer Nachdenkzeit von einer Minute standen drei Minuten zum Verschriften des Textes zur Verfügung. Die Auswertung der Schreibproben erfolgte nach den oben beschriebenen Auswertungsmethoden TWW (Total Words Written), WSC (Words Spelled Correctly), CWS (Correct Writing Sequences), CIWS (Correct Minus Incorrect Writing Sequences) und TWW + CIWS (Total Words Written plus Correct Minus Incorrect Writing Sequences). Es wurden in jeder Klassenstufe und auch zu jedem Messzeitpunkt unterschiedliche Story Starter verwendet, so dass insgesamt 80 verschiedene Story Starter zum Einsatz kamen (vier Klassenstufen mit jeweils zwei Messzeitpunkten à 10 Schreibproben).

Schreibkompetenz: Subtests 6 und 7 des TOWL-4 (Hammil & Larsen, 2009): Da im deutschen Sprachraum aktuell keine formalen Testverfahren zur Erhebung der Schreibkompetenz zur Verfügung stehen, wurden die Subtests 6 und 7 aus dem Test of Written Language 4 ins Deutsche übersetzt. Dabei handelt es sich um die Untertests *Contextual Conventions* und *Story Composition*. Der Testleiter präsentiert zunächst in standardisierter Form mündlich ein Beispiel für eine gute Geschichte und kommentiert dabei auch die Elemente einer gelungenen Geschichte. Als Schreibimpuls dient anschließend ein Bild einer Situation (z. B. Kindergeburtstag, Autounfall). Die Schülerinnen und Schüler werden aufgefordert, sich während fünf Minuten Notizen zu machen und anschließend in 15 Minuten die Geschichte zu verschriften. Untertest 6 (Contextual Conventions) enthält in der für die deutsche Sprache leicht angepassten Übersetzung 15 Auswertungskriterien, die sich auf die sprachliche Form des Textes beziehen (z.B. „Benutzt mindestens ein Fragezeichen" [ja/nein]; „Anzahl der richtig geschriebenen Wörter mit mehr als 7 Buchstaben"). Der Untertest 7 (Story Composition) bezieht sich auf die gleiche Schreibprobe und umfasst elf Kriterien zur Struktur und zum Inhalt der Geschichte (z. B. „Die Charaktere zeigen Gefühle, Emotionen" [ja/nein]). Beide Untertests wurden zu einem Gesamtscore addiert. Diese Gesamteinschätzung der Schreibqualität synthetisiert entsprechend inhaltliche und sprachformale Aspekte.

Fragebogen an die Lehrpersonen: Weitere Informationen über die teilnehmenden Schülerinnen und Schüler wurden schriftlich von den Lehrpersonen erfragt. Dabei handelt es sich um Daten, die der Stichprobenbeschreibung dienen (Alter, Mehrsprachigkeit, Sondermaßnahmen). Weiterhin wurden die Lehrpersonen gebeten, die Schreibkompetenzen jedes/r ihrer Schülerinnen und Schüler auf einer Skala von

1 – 10 global einzuschätzen (1 = sehr schlecht, 10 = sehr gut). Die Korrelationen zwischen den beiden Subtests aus dem TOWL-4 und der Lehrereinschätzung der Schreibkompetenz betrugen zwischen r = .44 (3. Klasse) und r = .68** (4. Klasse).

4.2.3 Datenerhebung, Testauswertung und Datenerfassung

Die Tests fanden in allen teilnehmenden Klassen in einem vorgegebenen Zeitintervall im Herbst bzw. Frühjahr statt und wurden durch die Lehrpersonen durchgeführt. Die Lehrpersonen wurden sorgfältig in die Testinstruktionen eingeführt. Die Auswertung der Daten erfolgte durch Bachelor- und Masterstudierende (Logopädie, Schulische Heilpädagogik) und eine der Autorinnen. Die Testauswertenden wurden in mehreren Sitzungen in die Bewertung der TOWL-Geschichte und in die Anwendung der LVD-Auswertungsmethoden eingeführt, bis eine hohe Übereinstimmung zwischen allen Ratern vorlag. Aus Kapazitätsgründen konnten nicht alle TOWL-Tests ausgewertet werden, sondern es wurde für jede Klassenstufe eine Zufallsauswahl von ca. 20 Schülerinnen und Schülern getroffen.

4.2.4 Statistische Analysen

Zur Schätzung der Zuverlässigkeit des Verfahrens LVD-Schreiben dient die Paralleltestreliabilität. Da zu beiden Testzeitpunkten jeweils 10 Schreibproben verfasst wurden, wird im Folgenden zur besseren Übersichtlichkeit die durchschnittliche Korrelation der 10 Schreibproben zu beiden Messzeitpunkten und jeweils die geringste und höchste Korrelation für die Maße TWW, WSC, CWS, CIWS und TWW+CIWS angegeben, gesondert nach Klassenstufen. In der praktischen Anwendung von LVD wird für den Einsatz als Universal Screening und für die Baseline der Lernverlaufsdiagnostik die Verwendung des Medians aus drei Schreibproben empfohlen (Hosp et al., 2016), weshalb es von Relevanz ist, wie Reliabilität und Validität dieses Wertes eingeschätzt werden können. Angegeben werden in Bezug auf die Paralleltestreliabilität des Medians die Mittelwerte der Korrelationen von drei Triplets zu jedem Messzeitpunkt (z. B. in Klassenstufe 3 die mittlere Korrelation zwischen den Medianen von Story Starter 2 – 4, 5 – 7 und 8 – 10 für T1 und von den Story Startern 42 – 44; 45 – 47 und 48 – 50 zu T2). Der erste Story Starter des Erhebungszeitpunkts wurde nicht verwendet, da die Schülerinnen und Schüler zu Beginn erst das Testformat kennenlernten.

Für Angaben zur Kriteriumsvalidität wird auf Korrelationen zwischen den Kennwerten des LVD-Verfahrens zu T1 mit dem Lehrerurteil und mit den Subtests 6 und 7 aus der Übersetzung des TOWL-4 zurückgegriffen. Dabei ist zu beachten, dass für den TOWL-4 nicht für alle Teilnehmenden Werte zur Verfügung stehen, sodass hier der Stichprobenumfang reduziert ist. Auch hier interessiert sowohl die durchschnittliche Korrelation einer einzelnen Schreibprobe als auch des Medians aus drei Schreibproben mit den beiden Kriteriumsmaßen (TOWL-4 und Lehrerbeurteilung).

Für die Beantwortung von Fragestellung vier zur Sensitivität für mittelfristige Lernfortschritte werden die Mittelwerte der Leistungen zu T1 für alle Auswertungsmethoden und Klassenstufen mit den Mittelwerten zu T2 mittels t-Tests für verbundene Stichproben verglichen. Einer Alphafehlerkumulation wird durch eine Bonferroni-Korrektur innerhalb der Klassenstufen entgegengewirkt. Um die Höhe und somit die praktische Relevanz der Effekte einzuschätzen, werden die Effektstärken dieser Vergleiche und die Korrelationen der Mittelwerte zu T1 und T2 präsentiert.

Zum Zweck einer ersten explorativen Einschätzung, wie sich die Werte der Schweizer Kinder zu den Normwerten aus dem anglo-

amerikanischen Raum verhalten, wurden die Mittelwerte der 10 Schreibproben der vorliegenden Stichprobe zu T1 und T2 den Angaben von aimsweb (publiziert in Hosp et al., 2016) deskriptiv gegenübergestellt. Im amerikanischen Schulsystem beginnt die Schreibinstruktion im Kindergarten und es werden auch bereits Normen für LVD – Schreiben auf Kindergartenstufe angegeben. Daher erfolgt der Vergleich der Schweizer Schülerinnen und Schüler jeweils mit der darunterliegenden amerikanischen Klassenstufe (Schweiz Stufe 3 mit den Normwerten von aimsweb von Stufe 2 usw.).

5 Ergebnisse

5.1 Paralleltestreliabilität von LVD – Schreiben in verschiedenen Klassenstufen zu T1 und T2

Tabelle 4 zeigt die durchschnittlichen Korrelationen zwischen den 10 Story Startern innerhalb der beiden Messzeitpunkte getrennt nach Klassenstufen für die verschiedenen Auswertungsmethoden auf. In Klammern ist jeweils die geringste und die höchste Korrelation der 10 Schreibproben angezeigt. Dabei offenbart sich durchgängig eine große Spannweite. Der Vergleich der durchschnittlichen Korrelationen zu T1 und T2 zeigt auf, dass diese in der Regel in einem sehr ähnlichen Bereich ausfallen. Alle durchschnittlichen Korrelationen befinden sich im Bereich zwischen .60 und .70 – mit Ausnahme der komplexeren Werte CWS, CIWS und TWW + CIWS in Klassenstufe 3. Es lässt sich eine Tendenz zu höheren Paralleltestreliabilitäten der komplexeren Maße gegenüber den einfachen Auswertungsmethoden TWW und WSC ab Klassenstufe 4 feststellen.

Verwendet man statt eines einzelnen Starters den Median von drei Schreibproben, so ergeben sich in der vorliegenden Studie drei Triplets pro Messzeitpunkt und Klassenstufe. In Tabelle 5 werden die durchschnittlichen Paralleltestreliabilitäten zwischen diesen Triplets präsentiert. Erwartungsgemäß führt das Aggregieren mehrerer Schreibproben zu einer deutlich erhöhten Paralleltestreliabilität, die ab Klassenstufe 4 mit wenigen Ausnahmen den Wert von .80 übertrifft.

Tab. 4 Durchschnittliche Korrelation und Minimum und Maximum der Korrelationen der 10 Story Starter zu T1 und T2 in den Klassenstufen 3–6 mit verschiedenen Auswertungsmethoden

		3. Klasse M (Min.–Max.)	4. Klasse M (Min.–Max.)	5. Klasse M (Min.–Max.)	6. Klasse M (Min.–Max.)
TWW	T1	.60 (.41–.77)	.65 (.40–.81)	.62 (.38–.80)	.65 (.44–.83)
	T2	.63 (.45–.81)	.69 (.57–.81)	.66 (.47–.75)	.69 (.43–.83)
WSC	T1	.59 (.36–.73)	.68 (.52–.81)	.65 (.50–.79)	.65 (.46–.81)
	T2	.65 (.46–.78)	.68 (.57–.78)	.66 (.48–.75)	.69 (.46–.82)
CWS	T1	.54 (.28–.69)	.70 (.59–.80)	.70 (.60–.78)	.69 (.56–.81)
	T2	.66 (.48–.76)	.68 (.57–.79)	.70 (.57–.82)	.74 (.60–.85)
CIWS	T1	.36 (.09–.56)	.67 (.50–.78)	.69 (.59–.79)	.73 (.62–.84)
	T2	.51 (.26–.67)	.66 (.53–.75)	.72 (.61–.80)	.73 (.60–.80)
TWW+CIWS	T1	.52 (.30–.66)	.70 (.62–.79)	.69 (.59–.78)	.70 (.56–.80)
	T2	.62 (.47– 74)	.67 (.58–.77)	.70 (.59–.82)	.74 (.60–.84)

Tab. 5 Durchschnittliche Paralleltestreliabilitäten zu T1 und T2 zwischen den Medianen von drei Schreibproben

		3. Klasse	4. Klasse	5. Klasse	6. Klasse
TWW	T1 T2	.77 .80	.79 .84	.78 .81	.79 .83
WSC	T1 T2	.76 .81	.82 .83	.81 .80	.79 .84
CWS	T1 T2	.66 .83	.84 .85	.83 .83	.82 .89
CIWS	T1 T2	.47 .74	.81 .82	.83 .85	.85 .88
TWW + CIWS	T1 T2	.65 .81	.84 .86	.84 .84	.82 .89

Tab. 6 Durchschnittliche Korrelation einer einzelnen Schreibprobe mit zwei verschiedenen Kriteriumsmaßen zu T1

Klassenstufe	Auswertungs-methode	Globale Lehrerbeurteilung	TOWL-4
3 (Lehrerurteil N = 72; TOWL N = 19)	TWW WSC CWS CIWS TWW + CIWS	.28 (.28) .35 (.37) .39 (.42) .20 (.23) .40 (.43)	.25 (.21) .28 (.25) .32 (.34) .07 (.12) .30 (.32)
4 (Lehrerurteil N = 109; TOWL N = 20)	TWW WSC CWS CIWS TWW + CIWS	.10 (.13) .33 (.38) .55 (.60) .61 (.64) .57 (.61)	.47 (.49) .62 (.66) .72 (.76) .59 (.67) .73 (.78)
5 (Lehrerurteil N = 92; TOWL N = 20)	TWW WSC CWS CIWS TWW + CIWS	.26 (.31) .39 (.43) .58 (.64) .59 (.65) .59 (.64)	-.10 (-.16) .07 (.04) .46 (.55) .68 (.76) .48 (.56)
6 (Lehrerurteil N = 81; TOWL N = 20)	TWW WSC CWS CIWS TWW + CIWS	.21 (.21) .32 (.35) .48 (.53) .53 (.56) .49 (.53)	.03 (.03) .17 (.19) .43 (.45) .56 (.58) .45 (.47)

Anmerkung: In Klammern: Durchschnittliche Korrelation des Medians aus drei Schreibproben mit den Kriteriumsmaßen zu T1.

5.2 Validität von LVD – Schreiben in verschiedenen Klassenstufen

Zur Bestimmung der Validität von LVD – Schreiben werden mit dem globalen Lehrerurteil und den beiden übersetzten Subtests aus dem TOWL-4 zwei verschiedene Kriteriumsmaße herangezogen. Da beide nur zu T1 erhoben wurden, bezieht sich Tabelle 6 auch nur auf den ersten Messzeitpunkt. Dargestellt ist die durchschnittliche Korrelation mit einer einzelnen Schreibprobe und die durchschnittliche Korrelation mit dem Median aus drei Schreibproben (in Klammern).

In Klassenstufe 3 sind alle beobachteten Zusammenhänge mit den beiden Außenkriterien gering, was sowohl für die Nutzung einer einzelnen Schreibprobe als auch für den Median aus drei Schreibproben gilt. Ab Klasse 4 hingegen sind die Korrelationen zwischen den auch im angloamerikanischen Raum empfohlenen Methoden CWS, CIWS und TWW + CIWS und dem TOWL-4 bzw. den Lehrereinschätzungen im moderaten Bereich anzusiedeln. Die beiden produktionsabhängigen Methoden TWW und WSC zeigen in allen Klassenstufen sehr niedrige Zusammenhänge mit den Kriteriumsmaßen.

5.3 Sensitivität für mittelfristige Lernfortschritte von LVD – Schreiben

Die Berechnung der mittelfristigen Änderungssensitivität erfolgte analog zur Untersuchung von McMaster und Campbell (2008), indem die Mittelwerte der 10 Schreibproben in der Herbst- und der Frühjahrserhebung mittels t-Tests für abhängige Stichproben verglichen wurden. Trotz Bonferroni-Korrektur innerhalb der Klassenstufen sind die Durchschnitts-

Tab. 7 Vergleiche der Mittelwerte zu T1 und T2 pro Klassenstufe und Auswertungsmethode

Klassenstufe	Auswertungsmethode	T1 M (SD)	T2 M (SD)	*t*	Cohens *d*	Korrelation T1 – T2
3 (df = 68)	TWW	22.10 (7.16)	30.83 (9.51)	-11.01**	1.32	.72**
	WSC	15.25 (5.78)	23.40 (8.27)	-13.05**	1.57	.78**
	CWS	7.91 (4.24)	14.15 (7.19)	-11.78**	1.41	.82**
	CIWS	-8.44 (5.85)	-5.65 (9.29)	-2.94*	.35	.54**
	TWW + CIWS	13.66 (8.00)	25.17 (13.15)	-11.66**	1.40	.80**
4 (df = 106)	TWW	33.70 (10.13)	39.50 (11.10)	-8.70**	.84	.79**
	WSC	26.91 (9.20)	33.06 (10.00)	-11.39**	1.10	.83**
	CWS	18.52 (8.46)	24.42 (10.01)	-12.81**	1.24	.88**
	CIWS	-.19 (13.13)	5.05 (14.95)	-7.88**	.76	.88**
	TWW + CIWS	33.51 (16.28)	44.56 (19.16)	-12.97**	1.25	.88**
5 (df = 91)	TWW	35.50 (10.03)	40.72 (11.52)	-7.43**	.75	.81**
	WSC	29.87 (9.64)	35.51 (11.00)	-9.03**	.94	.84**
	CWS	20.49 (9.05)	26.23 (10.88)	-11.27**	1.17	.89**
	CIWS	1.98 (12.77)	7.86 (15.10)	-8.55**	.89	.90**
	TWW + CIWS	37.49 (17.43)	48.59 (20.99)	-11.36**	1.18	.89**
6 (df = 78)	TWW	40.97 (11.73)	46.35 (12.70)	-7.03**	.79	.84**
	WSC	36.44 (11.19)	42.11 (12.53)	-8.32**	.93	.87**
	CWS	28.29 (11.70)	34.64 (13.33)	-11.32**	1.27	.92**
	CIWS	11.69 (16.02)	18.86 (16.49)	-8.47**	.95	.89**
	TWW + CIWS	52.67 (22.59)	65.22 (25.62)	-11.55**	1.30	.92**

werte in allen vier Stufen und mit jeweils allen fünf Auswertungsmethoden im Frühjahr hochsignifikant höher als im Herbst. Einzige Ausnahme stellt die Methode CIWS in der dritten Klasse dar, bei welcher der Vergleich nur auf dem 5 %-Niveau Signifikanz erreicht. Die durchwegs hohen Effektstärken unterstreichen das Bild, dass mit LVD – Schreiben ein deutlicher Lernfortschritt von Schülerinnen und Schülern der vier untersuchten Klassenstufen zwischen September und März dokumentiert werden kann.

5.4 Vergleich mit Orientierungswerten aus dem amerikanischen Sprachraum

Fragestellung vier sieht einen explorativen, deskriptiven Vergleich von bestehenden Normwerten für LVD – Schreiben aus dem amerikanischen Raum mit den Werten der Deutschschweizer Schülerinnen und Schüler vor. Wie in Abschnitt 4.2.4 dargelegt, wird die jeweils um ein Jahr niedrigere amerikanische Klassenstufe als Referenz herangezogen, da die Schreibinstruktion im dortigen Schulsystem bereits auf

Tab. 8 Vergleich der Normwerte (50. Perzentil) von aimsweb (Hosp et al., 2016) und der gerundeten Mittelwerte der vorliegenden Stichprobe

Auswertungsmethode	Klassenstufe	Herbst	Winter	Frühjahr
TWW	2 (aimsweb)	15	25	32
	3 (Schweiz)	22	31	
	3 (aimsweb)	26	34	39
	4 (Schweiz)	33	39	
	4 (aimsweb)	35	41	45
	5 (Schweiz)	36	40	
	5 (aimsweb)	41	48	51
	6 (Schweiz)	41	46	
WSC	2 (aimsweb)	10	21	24
	3 (Schweiz)	15	23	
	3 (aimsweb)	21	28	33
	4 (Schweiz)	27	33	
	4 (aimsweb)	30	35	35
	5 (Schweiz)	30	36	
	5 (aimsweb)	36	40	49
	6 (Schweiz)	36	42	
CWS	2 (aimsweb)	9	16	21
	3 (Schweiz)	8	14	
	3 (aimsweb)	18	24	30
	4 (Schweiz)	19	24	
	4 (aimsweb)	28	34	38
	5 (Schweiz)	20	26	
	5 (aimsweb)	34	39	46
	6 (Schweiz)	28	35	

der Kindergartenstufe beginnt. Die Messzeitpunkte bei aimsweb, auf welche sich die in Hosp et al. (2016) publizierten Daten beziehen, sind im September/Oktober (Herbst), Januar (Winter) und Mai (Frühling) vorgesehen (Clark, 2017). Die Daten der vorliegenden Schweizer Studie wurden im September und im März erhoben, womit der zweite Erhebungszeitpunkt zwischen den Winter- und Frühlingsdaten von aimsweb lag. Tabelle 8 zeigt, dass mit den Auswertungsmethoden TWW und WSC die Referenzwerte für den Herbst in Klassenstufe 5 und 6 gut übereinstimmen, weniger aber in den Klassenstufen 3 und 4, wo die Schweizer Schülerinnen und Schüler besser abschneiden als die amerikanischen. Für die Methode CWS verhält es sich umgekehrt: In den Klassen 2 und 3 sind die amerikanischen Normen fast äquivalent zu den Mittelwerten der Schweizer Kinder, in den Stufen 5 und 6 erreicht die englischsprachige Stichprobe höhere Werte.

Für den Erhebungszeitpunkt im März (Schweiz) gibt es für alle Auswertungsmethoden und Klassenstufen in den Winter- oder Frühjahrsnormen von aimsweb einen Vergleichswert, welcher sich um maximal zwei Einheiten unterscheidet und somit als nahezu äquivalent angesehen werden kann. Einzige Ausnahme stellt wiederum die Methode CWS in den Stufen 5 und 6 dar, wo der Abstand zwischen den etwas besser abschneidenden amerikanischen Kindern und der Schweizer Stichprobe etwas größer ist.

6 Diskussion

LVD-Verfahren dienen als Grundlage für pädagogische Entscheidungen von Lehrpersonen und Speziallehrkräften (Sonderpädagogik, Logopädie). Daher werden zu Recht hohe Anforderungen an ihre psychometrischen Eigenschaften gestellt, die sowohl den einzelnen Messzeitpunkt als auch die Messung des Lernverlaufs betreffen (Förster et al., 2017; Gebhardt et al., 2021). Ziel der vorliegenden Studie ist es, das Instrument LVD – Schreiben im deutschsprachigen Raum explorativ zu erproben und erste Resultate zu den Kriterien Paralleltestreliabilität, Validität und mittelfristige Änderungssensitivität für den Einsatz in den Klassenstufen 3 – 6 zu präsentieren.

Wie in Abschnitt 3.2.1 ausgeführt, sollte eine Paralleltestreliabilität > .70 angestrebt werden, damit ein LVD-Verfahren in der Funktion von Lernverlaufsdiagnostik eingesetzt werden kann, während eine Verwendung als Screening-Verfahren eine Reliabilität > .80 voraussetzt (Hosp & Kaldenberg, 2020). In der vorliegenden Studie variiert die durchschnittliche Paralleltestreliabilität für einzelne Schreibproben in den Klassenstufen 4 – 6 zwischen r = .65 und r = .75 und kann somit als moderat bezeichnet werden. In der 3. Klasse hingegen ist die durchschnittliche Paralleltestreliabilität bei allen Auswertungsmethoden schwach bis maximal moderat ausgeprägt, wenn eine einzelne Schreibprobe verwendet wird. Präsentiert wurde nicht nur die durchschnittliche Korrelation zwischen den zehn Schreibproben eines Messzeitpunktes, sondern auch der geringste und höchste Zusammenhang. Analog zur Studie von Allen et al. (2019) fällt dabei die große Spannbreite an Korrelationen auf. Das Design der vorliegenden Studie mit jeweils zehn Schreibproben pro Messzeitpunkt erlaubt es, auch die durchschnittliche Paralleltestreliabilität des Medians von drei Schreibproben zu berechnen. Unter Verwendung des Medians fällt die Paralleltestreliabilität deutlich aus und befindet sich ab der 4. Klasse für alle Auswertungsmethoden in einem zufriedenstellenden Bereich um den Wert von r = .80 oder gar höher. In Klassenstufe 3 trifft dies nicht auf alle Auswertungsmethoden zu und unterscheidet sich auch teilweise zwischen T1 und T2 deutlich. Die Reliabilität von LVD – Schreiben kann somit nicht pauschal bewertet werden. Es gibt weder eine Klassenstufe noch eine Auswertungsmethode, welche sich durch durchgängig hohe Werte der Paralleltestreliabilität auszeichnen. In Kapitel 2 wurde

begründet, weshalb die Zuverlässigkeit in der Schreibdiagnostik generell als problematisch beschrieben wird. Insbesondere die Komplexität des Konstrukts der Schreibkompetenz, welche in starkem Maße auch situativen Einflüssen unterliegt, kommt dabei zum Tragen. Eine Optimierung der Testbedingungen, also die Frage, wie häufig LVD – Schreiben in welcher Klassenstufe mit welcher Auswertungsmethode angewendet werden sollte, kommt an seine Grenzen, wo die beobachtete Variabilität ein Merkmal des zu messenden Konstrukts an sich darstellt. Gleichwohl zeigen sich über die internationale Forschungsliteratur hinweg auch stabile Muster (z. B., dass komplexere Methoden insbesondere in den höheren Klassen den simpleren Maßen überlegen sind), die spezifisch für den deutschen Sprach- und Unterrichtskontext genauer untersucht werden und schließlich in der Praxis Beachtung finden sollten.

Die Untersuchung der Validität von CBM – Schreiben erweist sich im deutschsprachigen Raum als schwierig, weil evaluierte Kriteriumsmaße gänzlich fehlen. In der vorliegenden Studie wurde die Schreibkompetenz der teilnehmenden Schülerinnen und Schüler sowohl durch ein ganzheitliches Lehrerurteil als auch durch eine informelle Übersetzung der Subtests 6 und 7 des im englischen Sprachraum sehr populären *Test of Written Language* (TOWL-4; Hammil & Larsen, 2009) erhoben. Da beide Außenkriterien selber nicht evaluiert sind, können diese Auswertungen nur als sehr explorativ angesehen werden. Weiterhin erhebt LVD – Schreiben nicht Schreibkompetenz insgesamt, sondern die Schreibflüssigkeit, welcher eine Indikatorfunktion für die Schreibkompetenz zugesprochen wird. Entsprechend wurde bereits dokumentiert, dass LVD – Schreiben zwar mit der Schreibqualität eng verknüpft ist, jedoch ein unterschiedliches Konstrukt erfasst (Kim et al., 2017 b). Beide Aspekte, also zum einen das gänzliche Fehlen von sauber operationalisierten Instrumenten zur Messung von Schreibkompetenz im Deutschen und zum anderen die konzeptionelle Unterscheidung zwischen Schreibflüssigkeit und Schreibkompetenz, müssen bei der Interpretation der Resultate berücksichtigt werden. Während im englischen Sprachraum $r = .50$ als akzeptabler Schwellenwert explizit benannt wird (z. B. Keller-Margulis et al., 2021; Allen et al., 2019), sollte im Deutschen vorsichtiger agiert werden. Im zu erwartenden Rahmen liegen moderate Korrelationen in einem Bereich zwischen $r = .40$ und $r = .70$.

In Klassenstufe 3 zeigt keine Auswertungsmethode einen akzeptablen Zusammenhang mit den Außenkriterien, und zwar unabhängig davon, ob eine einzelne Schreibprobe verwendet wird oder der Median aus drei Schreibproben. Ab Klassenstufe 4 weisen die drei Methoden CWS, CIWS und TWW+CIWS einen moderaten Zusammenhang zu den beiden Maßen für die Schreibqualität auf, häufig auch schon unter Verwendung einer Schreibprobe. Die Nutzung des Medians aus drei Schreibproben erhöht die Validität jeweils geringfügig. Die Scoringmethoden TWW und WSC hingegen sind in keiner Klassenstufe genügend valide, um als Indikatoren für die Schreibkompetenz von Schülerinnen und Schülern eingesetzt zu werden. Tabelle 9 fasst zusammen, welche Auswertungsmethoden in welcher Klassenstufe unter Verwendung von einer Schreibprobe oder des Medians von drei Schreibproben die in der angloamerikanischen Literatur dokumentierten kritischen Schwellenwerte von .70 (Paralleltestreliabilität) und .50 (Kriteriumsvalidität) überschreiten. Aufgrund des Pilotcharakters der Untersuchung und der fehlenden Repräsentativität der Stichprobe sind die Angaben aus Tabelle 9 nicht als abschließende Orientierung für die Praxis zu verstehen, sondern vielmehr als Vergleichs- und Diskussionsgrundlage für weiterführende Forschungsarbeiten. Zudem sollte beachtet werden, dass die Angaben zur Validität sich auf selbst nicht evaluierte Kriteriumsmaße stützen und dass im deutschen Sprachraum eine Diskussion und eine empirische Untermauerung des Richtwertes von .50 noch aussteht.

Wie bereits aus den vorhergehenden Beschreibungen deutlich wird, kann in Klasse 3 zwar eine ausreichende Paralleltestreliabilität durch das Aggregieren von drei Schreibproben erreicht werden, allerdings sind die Zusammenhänge aller Scoringmethoden mit den verwendeten Außenkriterien zur Schreibkompetenz (zu) niedrig ausgeprägt. Sollte sich dieses Bild durch weitere Studien verfestigen, wären für die Klassenstufe 3 andere Erhebungs- und/oder Auswertungsmethoden in Betracht zu ziehen. Beispielsweise könnte als Schreibimpuls auch ein Bild herangezogen oder die Erhebung der Schreibflüssigkeit auf Satzebene intendiert werden (z. B. Picture-Word-Aufgaben; McMaster et al., 2011; Ritchey et al., 2016).

In Klassenstufe 4 kann vorläufig auf der Basis der Werte zur Reliabilität und Validität die Nutzung der Auswertungsmethoden CWS und CIWS empfohlen werden. In der 5. und 6. Klasse erweist sich die Bewertung mittels CIWS der Nutzung von CWS überlegen. Die beiden Maße CWS und TWW + CIWS erreichen in allen Klassenstufen fast deckungsgleiche Werte, und zwar sowohl zur Reliabilität als auch zur Validität. Unterstrichen wird diese Feststellung durch die Korrelationen der Mittelwerte der zehn Schreibproben von CWS und TWW + CIWS, welche beispielsweise zu T1 zwischen r = .994** (3. Klasse) und r = .999** (6. Klasse) liegen. Damit sind die Korrelationen deutlich hö-

Tab. 9 Übersicht über das Erfüllen der Minimalkriterien zur Paralleltestreliabilität und zur Kriteriumsvalidität der Auswertungsmethoden je Klassenstufe

Klassenstufe	Auswertungsmethode	Paralleltestreliabilität > .70		Kriteriumsvalidität > .50	
		1 Schreibprobe	Median von 3 Schreibproben	1 Schreibprobe	Median von 3 Schreibproben
3	TWW	✘✘	✓✓	✘✘	✘✘
	WSC	✘✘	✓✓	✘✘	✘✘
	CWS	✘✘	✘✓	✘✘	✘✘
	CIWS	✘✘	✘✓	✘✘	✘✘
	TWW+CIWS	✘✘	✘✓	✘✘	✘✘
4	TWW	✘✘	✓✓	✘✘	✘✘
	WSC	✘✘	✓✓	✘✓	✘✓
	CWS	✘✓	✓✓	✓✓	✓✓
	CIWS	✘✘	✓✓	✓✓	✓✓
	TWW+CIWS	✘✓	✓✓	✓✓	✓✓
5	TWW	✘✘	✓✓	✘✘	✘✘
	WSC	✘✘	✓✓	✘✘	✘✘
	CWS	✓✓	✓✓	✘✓	✓✓
	CIWS	✘✓	✓✓	✓✓	✓✓
	TWW+CIWS	✘✓	✓✓	✓✘	✓✓
6	TWW	✘✘	✓✓	✘✘	✘✘
	WSC	✘✘	✓✓	✘✘	✘✘
	CWS	✘✓	✓✓	✘✘	✘✓
	CIWS	✓✓	✓✓	✓✓	✓✓
	TWW+CIWS	✓✓	✓✓	✘✘	✘✓

Anmerkung: Bei der Paralleltestreliabilität stehen die zwei Symbole für die Testzeitpunkte T1 und T2 (z. B.: ✓✓ = Kriterium zu beiden Testzeitpunkten erfüllt). Bei der Kriteriumsvalidität stehen die zwei Symbole für die beiden unterschiedlichen Kriteriumsvariablen.

her als die Zusammenhänge zwischen CIWS und TWW + CIWS (zwischen r=.479** und r = .872**), obgleich CIWS direkt in den Wert TWW + CIWS eingeht und CWS nicht. Die Methode TWW + CIWS wurde von Campbell et al. (2013) in die Diskussion um geeignete Scoringmethoden eingebracht und bislang noch kaum untersucht. Die Ergebnisse der vorliegenden Studie weisen darauf hin, dass die Verwendung der komplexer zu berechnenden und für Lehrpersonen weniger intuitiv zu verstehenden Methode TWW + CIWS keinen Mehrwert gegenüber der Nutzung von CWS erbringt. Die hier präsentierten Ergebnisse sprechen weiterhin dafür, von der Verwendung von TWW und von WSC in den Klassenstufen 3 – 6 grundsätzlich abzusehen. Zwar sind beide einfach und zeitsparend in der Umsetzung, aber der erhöhte Aufwand für die Auswertung mittels CWS oder CIWS sollte in Kauf genommen werden, da die simpleren Methoden (TWW, WSC) in Bezug auf die zentralen Testgütekriterien Reliabilität und Validität in keiner Klassenstufe überzeugende Werte erzielen. Zusammenfassend spiegeln die Ergebnisse das aus dem englischen Sprachraum bereits bekannte Bild wider, dass die Wahl der Auswertungsmethode vom Alter der Schülerinnen und Schüler abhängig zu machen ist und dass komplexere Scoringverfahren den einfacheren überlegen sind, insbesondere in den höheren Klassenstufen (McMaster & Espin, 2007). Für die Praxis kann weiterhin das Verwenden des Medians aus drei Schreibproben empfohlen werden, insbesondere wenn LVD – Schreiben als Screening verwendet wird. Zwar führt dieses Vorgehen zu Abstrichen im Bereich der Ökonomie, es erhöht auf der anderen Seite aber die Reliabilität des Verfahrens wesentlich (s. auch McMaster & Espin, 2007).

Fragen zur Änderungssensitivität von LVD-Verfahren werden idealerweise durch Studien mit wöchentlichen CBM-Messungen adressiert. Erste Hinweise lassen sich allerdings auch aus Untersuchungen mit zwei Messzeitpunkten im Abstand weniger Monate ableiten (Dockrell et al., 2015; Fearrington et al., 2014; Malecki & Jewell, 2003; McMaster & Campbell, 2008). In der vorliegenden Studie verbesserten sich die Schülerinnen und Schüler aller Klassenstufen unter Verwendung aller Auswertungsmethoden signifikant zwischen Herbst und Frühjahr, was durch fast ausschließlich hohe Effektstärken unterstrichen wird. Die hohen bis sehr hohen Korrelationen der Mittelwerte von T1 und T2 sprechen zudem dafür, dass die Rangfolge der Teilnehmenden zwischen beiden Messzeitpunkten stabil geblieben ist. Innerhalb eines Schuljahres können Entwicklungsfortschritte im Schreiben mittels der hier präsentierten Methoden also zuverlässig dokumentiert werden, was angesichts des Mangels an Alternativen im Bereich der Schreibdiagnostik bereits einen hohen Gewinn darstellt. Viele Lehrpersonen würden es vermutlich schätzen, den Lernverlauf ihrer Schülerinnen und Schüler im Bereich Schreiben nicht nur durch stark subjektiv geprägte holistische Einschätzungen zu evaluieren, sondern diese durch konkret messbare Indikatoren ergänzen zu können.

Zur Entwicklung der Schreibflüssigkeit, gemessen mit LVD – Schreiben, existieren im deutschsprachigen Raum bislang keinerlei Anhaltspunkte oder gar Normwerte. Der Nutzung von LVD-Schreiben in der Praxis steht dieses Manko nicht grundsätzlich entgegen, da auch der Rückgriff auf eine individuelle Bezugsnorm der Lehrperson bereits interessante Informationen für datengeleitete Entscheidungen zur Verfügung stellt. Weiterhin können Praktiker selber lokale Normen für die Klasse, Schule oder den Schulbezirk etablieren, um Hinweise auf den typischen Entwicklungsstand und Lernfortschritt von Schülerinnen und Schülern zu erhalten (zum konkreten Vorgehen siehe Hessler & Konrad, 2008). Die hier präsentierten Mittelwerte der Schweizer Schülerinnen und Schüler der Klassenstufen 3 – 6 stimmen in vielerlei Hinsicht gut mit den amerikanischen Normwerten der jeweils darunter

liegenden Stufe überein. Allerdings trifft dies nicht immer auf beide Messzeitpunkte im Jahr zu. Angesichts unterschiedlicher sprachlicher Strukturen (z. B. Tiefe der Orthografie; Häufigkeit von Komposita), Lehrpläne und didaktischer Methoden im englisch- und deutschsprachigen Raum erscheint es also sinnvoll, das Konstrukt der Schreibflüssigkeit und ihre Entwicklung im Schreiberwerb im Deutschen durch Studien besser zu dokumentieren und zukünftig zuverlässige Referenzwerte für den deutschsprachigen Raum zu erheben. Es sei an dieser Stelle explizit darauf verwiesen, dass aus den oben genannten methodischen Gründen aktuell weder die amerikanischen Angaben noch die hier publizierten Schweizer Werte als fixe Richtwerte für Forschung oder Praxis verwendet werden sollten.

Es sollte darüber hinaus nicht verschwiegen werden, dass der Einsatz von LVD – Schreiben als „Universal Screening" in ganzen Klassen oder Schulen mit einem hohen zeitlichen Auswertungsaufwand verbunden ist, was ihrer flächendeckenden Implementation entgegensteht (Payan et al., 2019). Aus diesem Grund liegen aktuell Hoffnungen auf der Entwicklung von automatisierten Auswertungssystemen, welche diesen Zeitaufwand beträchtlich reduzieren könnten (Keller-Margulis et al., 2021; Mercer et al., 2019). Payan et al. (2019) schlagen bis dahin für den Einsatz in der Praxis vor, LVD – Schreiben gezielt bei den Schülerinnen und Schülern einzusetzen, die von der Lehrperson aufgrund von Unterrichtsbeobachtungen bereits als Risikokinder identifiziert wurden. Mit Berücksichtigung des Umstands, dass auch holistische Einschätzungen von Kindertexten oder speziell das Beurteilen von Texten anhand von Beurteilungsrastern, wie sie von Lehrpersonen eingesetzt werden, mit einem hohen Aufwand verbunden sind – dass also die Einschätzung der Schreibkompetenz grundsätzlich aufwendiger ist –, erscheint aber die Verwendung von LVD – Schreiben auch für größere Kindergruppen empfehlenswert.

7 Forschungsausblick

Die vorliegende Studie präsentiert erste Resultate zu den Testgütekriterien von LVD – Schreiben im deutschsprachigen Raum. Sie weist ein eher ungewöhnliches Design auf, indem zu beiden Messzeitpunkten jeweils zehn LVD-Schreibproben verschriftet wurden. Diese hohe Menge an Schreibproben führt dazu, dass die Paralleltestreliabilitäten der Mediane von drei Schreibproben dokumentiert werden können, was auch in der englischsprachigen Literatur bislang noch nicht geschehen ist. Die Aussagekraft der Ergebnisse kann aber angesichts des explorativen Charakters und des geringen Stichprobenumfangs pro Klassenstufe nur als sehr vorläufig betrachtet werden. Insbesondere in den Klassenstufen 4 und 5 herrscht ein unausgeglichenes Geschlechterverhältnis vor, und in den Stufen 5 und 6 ist der Anteil der mehrsprachigen Schülerinnen und Schüler hoch.

Da LVD – Schreiben im deutschsprachigen Raum noch nicht etabliert wurde, ist die Liste der offenen Fragen zu diesem Messverfahren lang und kann an dieser Stelle nicht vollständig sein. Einige Aspekte wurden bereits in Abschnitt 6 angesprochen. Nachfolgende Untersuchungen könnten die Methodik beim Erheben und Auswerten der Schreibproben differenzieren. Eine Möglichkeit besteht beispielsweise darin, die Länge der Schreibprobe von drei auf fünf Minuten auszuweiten unter der Fragestellung, ob der erhöhte Aufwand zu einer reliableren Einschätzung der Schreibleistung führt (Espin et al., 2008; Furey, Marcotte, Hintze & Shackett, 2016). Insgesamt verdient die Frage der Reliabilität, welche hier nur im Rahmen der Klassischen Testtheorie analysiert wurde, erhöhte Aufmerksamkeit. Die Generalisierbarkeitstheorie ist eine methodische Alternative, die verschiedene Facetten von Fehlervarianz gleichzeitig zu analysieren vermag und somit geeignet ist,

die Rahmenbedingungen (Anzahl und Länge der Schreibproben, Einfluss von Rater und Story Starter) zugunsten einer möglichst zuverlässigen Messung zu optimieren (Keller-Margulis, Mercer & Thomas, 2016 b; Kim et al., 2017 a).

In der vorliegenden Studie wurden die Testgütekriterien getrennt nach Klassenstufen analysiert. Neben der Dimension des Alters wird im englischen Sprachraum die Eignung von LVD – Schreiben zunehmend häufig für unterschiedliche Zielgruppen untersucht, beispielsweise für Jungen vs. Mädchen (Fearrington et al., 2014; Jewell & Malecki, 2005), für mehrsprachige Schülerinnen und Schüler (Campbell, 2010; Campbell et al., 2013; Keller-Margulis et al., 2016a) oder für Kinder mit Lernstörungen (Dockrell et al., 2015; Ford & Kaldenberg, 2019) oder Hörbeeinträchtigungen (Cheng & Rose, 2009). Zukünftige Forschung sollte solch differenzierte Auswertungen auch für CBM – Schreiben im deutschsprachigen Raum vorsehen, um Testfairness zu gewährleisten (Gebhardt et al., 2021) und die adäquaten Auswertungsmethoden für jede Gruppe zu identifizieren. Eine weitere Forschungslücke ergibt sich im Hinblick auf die Funktion von LVD – Schreiben als Instrument für die engmaschigere Fortschrittsdiagnostik, wie sie z. B. im Rahmen individueller Förderung umgesetzt werden kann. Die Gütekriterien müssen für diesen Zweck nicht nur für die einzelne Messung (Static Score) erfüllt sein, sondern auch für den Lernverlauf (Slope). Diese zweite Stufe der Evaluation von LVD-Materialien nach Fuchs (2004) wurde bislang auch im englischen Raum selten in Angriff genommen. Schließlich wären auch Normen oder zumindest Orientierungswerte zur Entwicklung der Schreibflüssigkeit im deutschsprachigen Gebiet von hoher Relevanz, gemeinsam mit Anhaltspunkten zum typischerweise zu erwartenden Lernzuwachs in verschiedenen Klassenstufen und für unterschiedliche Gruppen von Lernenden.

Dank

Ein besonderer Dank gebührt den ehemaligen MA-Studentinnen Lena Müller, Pierangela Gilgen, Maura Tschanz und Milena Romano für ihre engagierte Mitarbeit bei der Datenerhebung und Auswertung.

Literatur

Allen, A. A., Jung, P.-G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S. & McMaster, K. L. (2019). Technical adequacy of curriculum-based measures in writing in grades 1 – 3. *Reading & Writing Quarterly, 33,* 1 – 25. https://doi.org/10.1080/10573569.2019.1689211

Amato, J. M. & Watkins, M. W. (2011). The predictive validity of CBM writing indices for eighth-grade students. *The Journal of Special Education, 44* (4), 195 – 204. https://doi.org/10.1177/0022466909333516

Blumenthal, Y. (2017). Ein Rahmenkonzept mit mehreren Förderebenen – Response to Intervention (RTI). In B. Hartke (Hrsg.), *Handlungsmöglichkeiten Schulische Inklusion. Das Rügener Modell kompakt,* 20 – 32. Stuttgart: Kohlhammer.

Campbell, H. (2010). The technical adequacy of curriculum-based measurement passage copying with secondary school English language learners. *Reading and Writing Quarterly, 26* (4), 289 – 307. https://doi.org/10.1080/10573569.2010.500253

Campbell, H., Espin, C. A. & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing, 26* (3), 431 – 452. https://doi.org/10.1007/s11145-012-9375-6

Cheng, S.-F. & Rose, S. (2009). Investigating the technical adequacy of curriculum-based measurement in written expression for students who are deaf or hard of hearing. *Journal of Deaf Studies and Deaf Education, 14* (4), 503 – 515. https://doi.org/10.1093/deafed/enp013

Chenu, F. (2020). Struggling writers and students with a learning disability in writing: Similarities and differences. In M. Dunn (Ed.), *Writing Instruction and Intervention for Struggling Writers. Multi-Tiered Systems of Support,* 61 – 69. Newcastle-upon-Tyne: Cambridge Scholars Publisher.

Clark, M. R. (2017). *aimswebPlus Introductory Guide.* Abgerufen am 12. 7. 2021 von https://www.marshfieldschools.org/cms/lib/WI01919828/Centricity/Domain/82/aimsplus%20introductory%20guide.pdf

Dockrell, J. E., Connelly, V., Walter, K. & Critten, S. (2015). Assessing children's writing products: the role of curriculum based measures. *British Educational Research Journal, 41* (4), 575 – 595. https://doi.org/10.1002/berj.3162

Espin, C., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D. & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children, 74* (2), 174 – 193. https://doi.org/10.1177/001440290807400203

Falkai, P., Wittchen, H.-U., Döpfner, M., Gaebel, W., Maier, W., Rief, W. et al. (Hrsg.) (2018). *Diagnostisches und statistisches Manual psychischer Störungen DSM-5.* 2., korrigierte Auflage. Göttingen: Hogrefe.

Fearrington, J. Y., Parker, P. D., Kidder-Ashley, P., Gagnon, S. G., McCane-Bowling, S. & Sorrell, C. A. (2014). Gender differences in written expression curriculum-based measurement in third- through eighth-grade students. *Psychology in the Schools, 51* (1), 85 – 96. https://doi.org/10.1002/pits.21733

Ford, J. W. & Kaldenberg, E. R. (2019). Curriculum-based measurement for written expression with postsecondary students with intellectual and developmental disabilities. *Journal of Inclusive Postsecondary Education, 1* (2), 1 – 22. https://doi.org/10.13021/JIPE.2019.2473

Förster, N., Kuhn, J.-T. & Souvignier, E. (2017). Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik, 9* (2), 116 – 122. https://doi.org/10.25656/01:14998

Fuchs, D. & Fuchs, L. S. (2017). Critique of the national evaluation of Response to Intervention: A case for simpler frameworks. *Exceptional Children, 83* (3), 255 – 268. https://doi.org/10.1177/0014402917693580

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33* (2), 188 – 193. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research & Practice, 32* (1), 5 – 7. https://doi.org/10.1111/ldrp.12127

Fuchs, L. S., Deno, S. L. & Marston, D. (1982). *Use of Aggregation to Improve the Reliability of Simple Measures of Academic Performance.* Minneapolis: Institute for Research on Learning Disabilities, University of Minnesota. Abgerufen am 12. 7. 2021 von https://files.eric.ed.gov/fulltext/ED227128.pdf

Furey, W. M., Marcotte, A. M., Hintze, J. M. & Shackett, C. M. (2016). Concurrent validity and classification accuracy of curriculum-based measurement for written expression. *School Psychology Quarterly, 31* (3), 369 – 382. https://doi.org/10.1037/spq0000138

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Naquiun, G. M. & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31* (4), 477 – 497. https://doi.org/10.1080/02796015.2002.12086169

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L. et al. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41* (3), 291 – 300. https://doi.org/10.1002/pits.10166

Gansle, K. A., Vanderheyden, A. M., Noell, G. H., Resetar, J. L. & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35* (3), 435 – 450. https://doi.org/10.1080/02796015.2006.12087977

Gebhardt, M., Jungjohann, J. & Schurig, M. (2021). *Lernverlaufsdiagnostik im förderorientierten Unterricht. Testkonstruktionen, Instrumente, Praxis.* München: Ernst Reinhardt Verlag.

Graham, S., Harris, K. & Hebert, M. (2011). *Informing Writing: The Benefits of Formative Assessment. A Carnegie Corporation Time to Act Report.* Washington DC: Alliance for Excellent Education.

Hammil, D. D. & Larsen, S. C. (2009). *Test of Written Language TOWL-4.* 4th ed. Austin: Pro-Ed.

Hessler, T. & Konrad, M. (2008). Using curriculum-based measurement to drive IEPs and instruction in written expression. *Teaching Exceptional Children, 41* (2), 28 – 37. https://doi.org/10.1177/004005990804100204

Hosp, J. L. & Kaldenberg, E. (2020). What is writing assessment for tiered decision making? In M. Dunn (Ed.), *Writing Instruction and Intervention for Struggling Writers. Multi-Tiered Systems of Support,* 70 – 85. Newcastle-upon-Tyne: Cambridge Scholars Publisher.

Hosp, M. K., Hosp, J. L. & Howell, K. W. (2016). *The ABC's of CBM. A Practical Guide to Curriculum-Based Measurement*. 2nd ed. New York, London: The Guilford Press.

Jewell, J. & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34* (1), 27 – 44.

Keller-Margulis, M., Payan, A., Jaspers, K. E. & Brewton, C. (2016 a). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Reading & Writing Quarterly, 32* (2), 174 – 198. https://doi.org/10.1080/10573569.2014.964352

Keller-Margulis, M. A., Mercer, S. H. & Thomas, E. L. (2016 b). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly, 31* (3), 383 – 392. https://doi.org/10.1037/spq0000126

Keller-Margulis, M. A., Mercer, S. H. & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: a comparison study. *Reading and Writing, 34,* 2461 – 2480. https://doi.org/10.1007/s11145-021-10153-6

Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B. & Al Otaiba, S. (2017 a). Writing evaluation: rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and Writing, 30* (6), 1287 – 1310. https://doi.org/10.1007/s11145-017-9724-6

Kim, Y.-S. G., Gatlin, B., Al Otaiba, S. & Wanzek, J. (2017 b). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities, 51* (4), 1 – 16. https://doi.org/10.1177/0022219417712016

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik,* 1 – 17. Göttingen: Hogrefe.

Malecki, C. K. & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40* (4), 379 – 390. https://doi.org/10.1002/pits.10096

McCloskey, M. & Rapp, B. (2017). Developmental dysgraphia: An overview and framework for research. *Cognitive Neuropsychology, 34* (3 – 4), 65 – 82. https://doi.org/10.1080/02643294.2017.1369016

McMaster, K. & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *The Journal of Special Education, 41* (2), 68 – 84. https://doi.org/10.1177/00224669070410020301

McMaster, K. L. & Campbell, H. (2008). New and existing curriculum based writing measures technical features within and across grades. *School Psychology Review, 37* (4), 550 – 566. https://doi.org/10.1080/02796015.2008.12087867

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D. & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77* (2), 185 – 206. https://doi.org/10.1177/001440291107700203

McMaster, K. L., Shin, J., Espin, C. A., Jung, P.-G., Wayman, M. M. & Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: grade and gender differences. *Reading and Writing, 30* (9), 2069 – 2091. https://doi.org/10.1007/s11145-017-9766-9

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K. & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42* (2), 117 – 128. https://doi.org/10.1177/0731948718803296

Parker, D. C., McMaster, K. L. & Burns, M. K. (2011). Determining an instructional level for early writing skills. *School Psychology Review, 40* (1), 158 – 167. https://doi.org/10.1080/02796015.2011.12087735

Payan, A. M., Keller-Margulis, M., Burridge, A. B., McQuillin, S. D. & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention, 45* (1), 51 – 64. https://doi.org/10.1177/1534508418781007

Poch, A. L., Allen, A. A., Jung, P.-G., Lembke, E. S. & McMaster, K. L. (2021). Using data-based instruction to support struggling elementary writers. *Intervention in School and Clinic, 57* (3), 3 – 11. https://doi.org/10.1177/10534512211014835

Ritchey, K. D. & Coker, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly, 29*(1), 89–119. https://doi.org/10.1080/10573569.2013.741957

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y.-S. G., Parker, D. C. et al. (2016). Indicators of fluent writing in beginning writers. In K. D. Cummings & Y. Petscher (Eds.), *The Fluency Construct. Curriculum-Based Measurement Concepts and Applications,* 21–66. New York, NY: Springer New York.

Romig, J. E., Therrien, W. J. & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education, 51*(2), 72–82. https://doi.org/10.1177/0022466916670637

Saddler, B. & Asaro-Saddler, K. (2013). Response to intervention in writing: A suggested framework for screening, intervention, and progress monitoring. *Reading & Writing Quarterly, 29*(1), 20–43. https://doi.org/10.1080/10573569.2013.741945

Strathmann, A., Klauer, K. J. & Greisbach, M. (2010). Lernverlaufsdiagnostik – Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule. *Empirische Sonderpädagogik, 2*(1), 64–77. http://dx.doi.org/10.25656/01:9338

Sturm, A., Nänny, R. & Wyss, S. (2017). Entwicklung hierarchieniedriger Schreibprozesse. In M. Philipp (Hrsg.), *Handbuch Schriftspracherwerb und weiterführendes Lesen und Schreiben,* 84–104. Weinheim: Beltz Juventa.

Traga Philippakos, Z. A. & FitzPatrick, E. (2018). A proposed tiered model of assessment in writing instruction: Supporting all student-writers. *Insights into Learning Disabilities, 15*(2), 149–173.

Walter, J. (2013). *VSL. Verlaufsdiagnostik sinnerfassenden Lesens.* Göttingen, Bern: Hogrefe.

Wanzek, J., Gatlin, B., Al Otaiba, S. & Kim, Y.-S. G. (2017). The impact of transcription writing interventions for first-grade students. *Reading & Writing Quarterly, 33*(5), 484–499. https://doi.org/10.1080/10573569.2016.1250142

Weissenburger, J. W. & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*(2), 153–169. https://doi.org/10.1016/j.jsp.2005.03.002

Wilson, J., Chen, D., Sandbank, M. P. & Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *Journal of Educational Psychology, 111*(4), 619–640. https://doi.org/10.1037/edu0000311

WHO/World Health Organization (2021). *International Classification of Diseases 11th Revision.* Abgerufen am 18. 11. 2021 von https://icd.who.int/en/

Anschriften der Autorinnen

Dr. Julia Winkes
Universität Freiburg
Departement für Sonderpädagogik
Petrus-Kanisius-Gasse 21
CH-1700 Freiburg
E-Mail: julia.winkes@unifr.ch

Dr. Pascale Schaller
Pädagogische Hochschule Bern
Fabrikstr. 8
CH-3012 Bern
E-Mail: pascale.schaller@phbern.ch