Pilotierung von Leseflüssigkeits- und Leseverständnistests zur Entwicklung von Instrumenten der Lernverlaufsdiagnostik

Ergebnisse einer Längsschnittstudie in der 3ten und 4ten Jahrgangsstufe[I]

Jana Jungjohann[1], Michael Schurig[2], Markus Gebhardt[1]
[1] **Universität Regensburg**
[2] **Technische Universität Dortmund**

Zusammenfassung: Individuelle Lernverläufe im Lesen können mittels Lernverlaufsdiagnostik im Längsschnitt beobachtet werden. Die Lernverlaufstests müssen dafür auch über die Zeit reliabel und änderungssensibel messen. Dieser pilotierende Beitrag prüft erstmalig die Reliabilität sowie die Änderungssensibilität von bereits im Querschnitt evaluierten Lernverlaufstests zur Messung der Leseflüssigkeit und des basalen Leseverständnisses im Längsschnitt. Zu vier Messzeitpunkten innerhalb eines Schuljahres wurden Lernende des dritten und vierten Jahrgangs einer regulären Grundschule ($N = 90$) mit drei Leseflüssigkeitstests (Silben, Wörter, Pseudowörter) sowie einem Satzverständnistest von externen Testerinnen in der Onlineplattform www.levumi.de getestet. Die Lehrkräfte erhielten am Ende der Studie die Ergebnisse. Die Ergebnisse zeigen für beide Klassenstufen und alle Messzeitpunkte zufriedenstellende Paralleltest-Reliabilitätswerte ($r = .72 – .90$). Im Schnitt steigen die Leseleistungen in 12 Wochen zwischen 2.15 (0.55) und 5.41 (0.98) Items. 4.26 – 14.80 % der Lernenden zeigen negative Lernverläufe ($b = -0.8 – -4.0$), die anteilig durch studiendesignbedingte Deckeneffekte erklärt werden können. Die Befunde der Pilotstudie deuten auf eine grundsätzliche Einsatzmöglichkeit der Tests im Feld hin, sodass sie bei größeren Studien Verwendung finden können.

Schlüsselbegriffe: Leseentwicklung, Lernverlaufsdiagnostik, Leseflüssigkeit, Leseverständnis, Grundschule

Piloting Reading Fluency and Reading Comprehension Tests for the Development of Instruments for Curriculum-Based Measurement. Results of a Longitudinal Study in 3rd and 4th Grade

Summary: Individual learning progress in reading can be longitudinally monitored by curriculum-based measurement. For this purpose, the learning progress tests must measure reliably and change sensitively over time. For the first time, this pilot paper examines the reliability as well as the change sensitivity of curriculum-based measurement for quantifying reading fluency and reading comprehension in the second half of primary school over a school year. To this end, reading developments in the 3rd and 4th grades ($N = 90$) were measured at four measurement points in time. For each measurement point, the learners were tested with three reading fluency tests (syllables, words, pseudowords) and a sentence comprehension test from the online platform www.levumi.de. The teachers received the results at the end of the study. The results show satisfactory parallel test reliability values ($r = .72 – .90$) for both grades levels and all measurement points. On average, the reading performance increases between 2.15 (0.55) and 5.41 (0.98) items in 12 weeks. 4.26 – 14.80 % of the learners show negative slopes ($b = -0.8 – -4.0$), which can be proportionally explained by design-related ceiling effects. The findings of the pilot study indicate that the tests can basically be used in the field so that they can be applied in larger studies.

Keywords: Curriculum-based measurement, primary school, reading comprehension, reading development, reading fluency

1 Lernverlaufsdiagnostik im Bereich Lesen

Der Erwerb basaler Lesekompetenzen aller Lernenden ist ein zentrales Ziel der Grundschule, um eigenständig und sicher Informationen aus Texten konstruieren zu können (Scheerer-Neumann, 2015). Zu den bedeutendsten Schritten des Leseerwerbs in der Grundschule zählen die Entwicklung einer angemessenen schnellen Leseflüssigkeit sowie ein sicheres basales Leseverständnis (Klicpera, Schabmann, Gasteiger-Klicpera & Schmidt, 2017; Lenhard & Artelt, 2009; Rosebrock & Nix, 2017). Allerdings gelingt der Leseerwerb bei einzelnen Schülerinnen und Schülern unterschiedlich gut. Ca. 20 % verlassen die Grundschule ohne ausreichende Lesekompetenzen für das Lernen in der Sekundarstufe, wie die internationale Lesekompetenzstudie IGLU 2016 in Deutschland aufzeigt (Hußmann et al., 2017). Besorgniserregend ist, dass der Anteil der schwachen Leserinnen und Leser im Vergleich zu den Vorgängerkohorten 2006 und 2011 gewachsen ist (Bos, Valtin, Hußmann, Wendt & Goy, 2017).

Insbesondere Schülerinnen und Schüler mit einem Risiko für Leseschwierigkeiten im inklusiven Unterricht oder an der Förderschule benötigen einen Unterricht, der ihre individuellen Lernvoraussetzungen und Lernbedürfnisse berücksichtigt und fördert (Walter, 2008). Das Ziel des Leseunterrichts ist es, die Lernenden bei aufkommenden Lernschwierigkeiten frühzeitig effektiv zu fördern, sodass sich Schwierigkeiten nicht langfristig manifestieren. Damit Schwierigkeiten im Lesen möglichst frühzeitig sichtbar werden, werden diagnostische Informationen über den Lernstand und die Lernentwicklungen der Lernenden in den Unterricht einbezogen. Hierbei werden diagnostische Instrumente verwendet, welche entweder den Status quo oder die Lernentwicklung erfassen (Hartke, 2017). Während es für den deutschsprachigen Raum viele Lesetests gibt, um den Status quo des Lernstandes festzustellen (Mayer, 2016; Hesse & Latzko, 2011), stehen bisher nur wenige geprüfte Tests zur Erfassung der Lernentwicklung zur Verfügung (Jungjohann, Gegenfurtner & Gebhardt, 2018).

Der Ansatz der Lernverlaufsdiagnostik wurde unter starkem Einfluss des US-amerikanischen Konzepts *curriculum-based measurement* (CBM; Deno, 1985) in Deutschland verbreitet. Er ermöglicht durch leicht handhabbare Tests mit kurzer Durchführungsdauer das Abbilden von individuellen Lernverläufen als Kurven im zeitlichen Verlauf (Klauer, 2011). Diese Lernverläufe können Lehrkräften eine Rückmeldung geben, ob und in welchem Ausmaß ihre Schülerinnen und Schüler vom angebotenen Unterricht profitieren. Die Intention der Lernverlaufsdiagnostik ist, der Manifestierung von Schwierigkeiten präventiv vorzubeugen (Deno, Fuchs, Marston & Shin, 2001). Das kann gelingen, da Lernende mit besonderen Unterstützungsbedarfen durch langsam ansteigende oder stagnierende Lernverläufe identifizierbar sind. Die Tests der Lernverlaufsdiagnostik sind für einen häufigen Einsatz im Unterricht möglichst praktikabel konstruiert. Sie zeichnen sich dadurch aus, dass sie nur wenige Minuten dauern, zahlreiche Parallelversionen für einen wiederholenden Einsatz bereitstellen und leicht interpretierbar sind, trotzdem aber den Ansprüchen einer reliablen Messung im Längsschnitt genügen (Wilbert, 2014).

In Anlehnung an amerikanische Forschungsergebnisse wurden einige deutschsprachige Lernverlaufsdiagnostiken zur Erfassung von Lesekompetenzen in der Grundschule konstruiert (z. B. Gebhardt, Diehl & Mühling, 2016; Voß & Blumenthal, 2020; Walter, 2010 a; 2013). Lernverlaufstests für das Lesen sind üblicherweise als eindimensionale Konstrukte konstruiert. Sie messen meist robuste Indikatoren wie die Leseflüssigkeit, die mit einer umfassenden Kompetenz hoch korrelieren (Deno, Mirkin & Chiang, 1982; Fuchs, Fuchs, Hosp & Jenkins, 2001). Die Leseflüssigkeit wird häufig durch lautes Vor-

lesen von einzelnen Silben, Wörtern oder zusammenhängenden Texten erfasst (ebd.). Diese Tests dauern meist 60 Sekunden. Während der Durchführung bewertet eine Lehrkraft, ob die Schülerin oder der Schüler korrekt vorliest. Zur Erfassung des Leseverständnisses werden lückenhafte Sätze oder Texte auf Zeit bearbeitet, bei denen mithilfe von vorgegebenen Antwortmöglichkeiten die Lücken gefüllt werden (Shin, Deno & Espin, 2000). Beide Indikatoren werden meist als Speedtest ausgewertet.

Die Anwendung der Lernverlaufsdiagnostik umfasst mehrere Schritte: (1) Auswahl passender Lernverlaufstests, (2) Durchführung mehrerer Messungen, (3) Übertragung der Messergebnisse in einen Lernverlaufsgraphen, (4) Ableitung von zukünftigen Lernzielen und (5) Planung und Adaption von Förderungen auf der Basis der Lernverlaufsdaten (data-based decision-making; Hosp, Hosp & Howell, 2007). Je nach Einsatzzweck variieren die zeitlichen Abstände zwischen den Tests sowie die Häufigkeit der Messungen (Deno, 2003). Durch wöchentliche Messungen können Lernentwicklungen einzelner Schülerinnen und Schüler kleinschrittig beobachtet werden, um Effekte von intensiven und zeitaufwendigen Förderungen zu beurteilen. Messergebnisse von ganzen Schulklassen mit größeren Messabständen von mehreren Monaten werden hingegen genutzt, um Risikokinder zu identifizieren.

2 Die Erforschung einer Lernverlaufsdiagnostik

Zur Erforschung und Erprobung einer Lernverlaufsdiagnostik schlägt Fuchs (2004) ein dreistufiges Vorgehen vor. Zuerst müssen die Reliabilität und die Objektivität im Querschnitt zu einem Messzeitpunkt anhand einer großen Stichprobe überprüft werden (Stufe 1). Dies wird häufig gemäß dem Vorgehen der klassischen Testtheorie gemacht (Jungjohann, Gegenfurtner & Gebhardt, 2018). Alternativ wird die Prüfung der Testgüte mittels der Item-Response-Theorie empfohlen (Gebhardt, Heine, Zeuch & Förster, 2015; Wilbert & Linnemann, 2011). In der zweiten Stufe wird die technische Einsetzbarkeit der Lernverlaufstests im Längsschnitt geprüft (Fuchs, 2004). Studien dieser Stufe können mit und ohne Rückmeldung der Ergebnisse an die Lehrkräfte erfolgen. Zum zweiten Forschungsstadium zählen auch Untersuchungen der Änderungssensibilität über Zuwachsraten mittels linearer Regressionsanalysen. In den Studien von Walter (2011 a) und Shin et al. (2000) wurde zunächst die Paralleltest-Reliabilität mittels Pearson-Produkt-Moment-Korrelation zwischen den einzelnen Messzeitpunkten anhand von Stichproben ohne Kontrollgruppe geprüft. Anschließend wurden die Mittelwerte mit Standardabweichungen zu allen Messzeitpunkten auf deskriptiver Ebene beschrieben und auf signifikante Mittelwertsunterschiede hin überprüft. Die Sensibilität der Messverfahren wird durch mehrere aus der Theorie und Empirie abgeleitete Hypothesen, wie Unterschiede zwischen Leistungsgruppen, Klassenstufen oder in der Entwicklung, geprüft. In beiden Studien wird nicht beschrieben, ob die Lehrkräfte Rückmeldungen über die Lernverläufe ihrer Schülerinnen und Schüler erhielten. In der dritten Stufe der Erforschung der Lernverlaufsdiagnostik nach Fuchs (2004) werden die Effektivität und der praktische Nutzen der Lernverlaufsdiagnostik bei ihrer Anwendung durch Lehrkräfte im Feld untersucht (Stufe 3). Studien der Stufen 1 und 2 sind Voraussetzung für die Belastbarkeit der Maße vor einem Praxiseinsatz, die stets mehrere Messzeitpunkte einer abhängigen, meist kleineren Stichprobe enthalten (Voß & Gebhardt, 2017). Förster, Kuhn und Souvignier (2017) ergänzen die Stufen nach Fuchs (2004) und beschreiben eine Abhängigkeit in Normierungsstudien zwischen den spezifischen Bedingungen und dem individuellen Einsatzzweck der geprüften Lernverlaufsdiagnostik. Die Autorengruppe nennt als grundlegende Voraussetzung für eine Übertragbarkeit

von Vergleichswerten eine möglichst hohe Vergleichbarkeit der schulischen Bedingungen, der Testsituation sowie der Art des praktizierten Unterrichts während der Studien. Nur wenn in Studien schulalltagsähnliche Bedingungen herrschen, werden Erkenntnisse über spezifische Lernverläufe in verschiedenen schulischen Settings ermöglicht.

2.1 Vergleichswerte im Querschnitt und Längsschnitt

Um die Ergebnisse einer Lernverlaufsdiagnostik für diagnostische Zwecke im Unterricht zu nutzen, ziehen Lehrkräfte querschnittliche Vergleichswerte zu einem Messzeitpunkt und durchschnittliche Zuwachsraten einer Kontrollgruppe im Längsschnitt zwischen mehreren Messzeitpunkten heran (Shinn, 2007). Die Rohwerte eines Kindes können zur besseren Interpretation und zur Beschreibung des Leistungsniveaus zu normativen Vergleichswerten (engl. benchmark) in Beziehung gesetzt werden (Förster et al., 2017). Häufig werden zur Feststellung des Status quo Vergleichsnormen zu definierten Zeitpunkten angeboten. Diese Vergleichsnormen bestehen aus Ergebnissen einer Querschnittsstudie, in der die Schülerinnen und Schüler keine spezifische Intervention erhalten. Häufig werden für die Instrumente zu drei bis vier Messzeitpunkten pro Schuljahr (z. B. Herbst, Winter, Frühjahr, Sommer) Normen abgeleitet und getrennt nach Klassenstufe deskriptiv beschrieben (Good, Simmons & Kame'enui, 2001; Grapin, Kranzler, Waldron, Joyce-Beaulieu & Algina, 2017).

Neben den querschnittlichen Normen gibt es auch den Vergleich der Lernentwicklung anhand von längsschnittlichen Zuwachsraten (engl. slope). In der Lernverlaufsdiagnostikforschung werden die Lernentwicklungen der Schülerinnen und Schüler über die Zeit mit festen Messabständen pro Woche (Fuchs, Fuchs, Hamlett, Walz & Germann, 1993) oder pro Monat (Shin et al., 2000) beschrieben. Für den Vergleich von Lernentwicklungen werden die durchschnittlichen Zuwachsraten aller Lernenden im regulären Unterricht ohne Intervention benötigt. Zur Modellierung von Zuwachsraten werden meist lineare Modelle mit Betrachtung der Steigung als unstandardisiertem b-Wert (Regressionskoeffizient) herangezogen (Deno et al., 2001; Graney, Missall, Martínez & Bergstrom, 2009; Silberglitt & Hintze, 2005; Walter, 2010 b; 2011 a). Alternativ können auf Aggregatebene auch diskontinuierliche Zuwachsraten mittels komplexerer Vorgehensweisen abgetragen werden, wie beispielsweise bei Nese et al. (2012) durch hierarchische Modelle. Der Nachteil der Modellierung diskontinuierlicher Zuwachsraten ist jedoch, dass die Anforderungen an die Stichprobe höher und die Vorannahmen zum erwarteten Wachstum umfänglicher sind als bei Analysen mittels linearer Modelle. Eine hohe Präzision bei der Schätzung von Wachstumskurven erreichen latente Wachstumskurvenmodelle unter expliziter Berücksichtigung der Messfehlertheorie (Yeo, Fearrington & Christ, 2012). Bei diesem Ansatz stellt jeder Messzeitpunkt einen Indikator für die latente Steigung und das Ausgangsniveau der Steigung (engl. intercept) dar. Die Anzahl der Messzeitpunkte steht aber im Zusammenhang mit der benötigten Stichprobengröße (Muthén & Curran, 1997). Für latente Wachstumskurvenmodelle werden generell umfangreichere Stichproben benötigt als bei einer linearen Regression. Auch werden nicht individuelle, sondern mittlere Wachstumswerte geschätzt. Die individuellen Werte werden als Abweichungen begriffen, deren Schätzung stark von der Reliabilität des mittleren Wachstums abhängt. Wenn das mittlere Wachstum eine hohe Varianz aufweist, werden individuelle Schätzungen ungenauer. Die Schätzung der Lernzuwächse mittels linearer Regressionen berücksichtigt die Messfehlertheorie weniger stark, muss also als Vereinfachung gesehen werden. Für die Prüfung der Änderungssensibilität können lineare Regressionen als hinreichend praktikable

und reliable Indikatoren verwendet werden (Walter, 2014). Wie jeder andere Messwert sind auch b-Werte fehlerbehaftet und die Größe des Messfehlers eignet sich zur Einschätzung der Angemessenheit der verwendeten Werte (Christ, Zopluoglu, Monaghen & Van Norman, 2013; Walter, 2014). Fuchs et al. (1993) empfehlen bei der Analyse von Zuwachsraten eine Betrachtung des individuellen Ausgangsniveaus, um falsche Erwartungen an zukünftige Lernentwicklungen zu vermeiden. Denn ohne Berücksichtigung des individuellen Ausgangsniveaus werden von Lernenden mit einem geringen Ausgangsniveau höhere Lernzuwächse im gleichen Zeitraum erwartet als von Schülerinnen und Schülern mit einem höheren Ausgangsniveau. Für differenzierte Betrachtungen von Leistungsgruppen wird oft eine Unterteilung der Zuwachsraten getrennt nach Perzentilen vorgenommen (Hasbrouck & Tindal, 2006; Walter, 2010 b). Alternativ finden sich Aufteilungen der Zuwachsraten nach den manifesten Ausgangsniveaus (Deno et al., 2001).

2.2 Erwartungen über durchschnittliche Lernverläufe im Lesen

In Anlehnung an Entwicklungsmodelle des Lesens sind bei einem regulären Leseerwerb gegen Ende der Grundschule geringere Lernzuwächse in der Leseflüssigkeit und höhere Lernzuwächse im Leseverständnis zu erwarten, da durchschnittliche Leserinnen und Leser in dieser Phase bereits routiniert flüssig lesen (Müller, Križan, Hecht, Richter & Ennemoser, 2013). Gleichzeitig zeigen ältere Lernende geringere Lernzuwächse in basalen Lesekompetenzen als jüngere Lernende zu Beginn des Leseerwerbs (Richter, Isberner, Naumann & Kutzner, 2012). In nationalen und internationalen Forschungsarbeiten zeigt sich die zentrale Tendenz, dass bei regulärem Grundschulunterricht die Zuwachsraten bei den Verfahren zum Leseverständnis über alle Jahrgangsstufen hinweg geringer ausfallen als bei Leseflüssigkeitstests und dass mit steigendem Alter der Lernenden die Zuwachsraten geringer ausfallen (Fuchs et al., 1993; Graney et al., 2009; Walter, 2010 b; 2011 b). Silberglitt und Hintze (2007) untersuchen, ob sich Zuwachsraten in Abhängigkeit vom Ausgangsniveau verändern. Sie modellieren die Zuwachsraten im Lesen von 7544 Lernenden in den Klassenstufen zwei bis sechs über ein Schuljahr zu drei Messzeitpunkten mittels eines hierarchischen linearen Wachstumsmodells. Im Vergleich zur Gruppe der durchschnittlichen Schülerinnen und Schüler (50 – 59 Perzentil) waren beim untersten als auch beim obersten Perzentil signifikant geringere Zuwachsraten messbar. Dies bestätigt die Relevanz, Perzentilbereiche in der Pilotierung von Verfahren getrennt zu analysieren. Weitere Einflüsse auf die Abbildung der individuellen Lernentwicklungen können durch die erhobene Kompetenz, das konkret eingesetzte Verfahren und die Erhebungsbedingungen (z. B. Untersuchungszeitraum, Ort der Testung, Testleitung, Testsituation) entstehen (Christ et al., 2013).

3 Forschungsfragen

Jedes neue Instrument der Lernverlaufsdiagnostik benötigt eigene evaluierende Studien, da Vergleichswerte und Zuwachsraten testspezifisch sind und durch viele Faktoren beeinflusst werden. Die Onlineplattform www.levumi.de besteht seit 2015 und bietet browserbasierte digitale Lernverlaufstests zur Messung der Leseflüssigkeit und des basalen Leseverständnisses in der Grundschule an. In Anlehnung an das Vorgehen von Fuchs (2004) liegen bereits Studienergebnisse für die Stufe 1 vor (Gebhardt et al., 2016; Jungjohann, DeVries, Gebhardt & Mühling, 2018; Jungjohann, DeVries, Mühling & Gebhardt, 2018). In diesen Studien wird die psychometrische Güte der hier eingesetzten Tests mittels klassischer Testtheorie sowie der Item-Response-Theorie erfolgreich geprüft.

Beispielsweise wurden in einer Querschnittsstudie mit 132 Schülerinnen und Schülern für den Test Silbenlesen N4 eine Reliabilität von $r_{WLE} = 0.90$ sowie eine Test-Retest-Reliabilität von $r = 0.85$ beobachtet (Jungjohann, DeVries, Gebhardt et al., 2018). Zusätzlich wurde die Eignung der Tests für den Einsatz im leistungsheterogenen Grundschulunterricht mit Schülerinnen und Schülern mit unterschiedlichen Personenmerkmalen (z. B. Geschlecht, Migrationshintergrund, sonderpädagogischer Förderbedarf) untersucht. Ebenso gibt es eine Studie mit einer standardisierten Leseintervention über ein Schulhalbjahr, in der mäßige bis hohe positive Korrelationen zwischen den Levumi-Tests und einem standardisierten Leseverständnistest (ELFE II, Lenhard, Lenhard & Schneider, 2017) von $r = 0.52 – 0.81$ beobachtet wurden (Anderson, Jungjohann & Gebhardt, 2020). Ein Längsschnitt ohne Intervention über ein Schuljahr von mehreren Jahrgangsstufen fehlt für die Lesetests von Levumi und stellt einen nächsten Schritt in der Prüfung einer Lernverlaufsdiagnostik nach Fuchs (2004) dar.

In der vorliegenden Pilotstudie werden drei Lernverlaufstests zur Messung der Leseflüssigkeit und ein Test zum Leseverständnis zu vier Messzeitpunkten innerhalb eines Schuljahres an einer Partnergrundschule eingesetzt, um die generelle Einsetzbarkeit der Tests im Längsschnitt in der dritten und vierten Klassenstufe erstmalig zu überprüfen. In Anlehnung an Walter (2011 a) und Shin et al. (2000) werden die Reliabilität und die Änderungssensibilität der Levumi-Lesetests im Längsschnitt ohne Kontrollgruppe geprüft. Das Ziel ist, pilotierend Lernverläufe und Zuwachsraten in der Leseflüssigkeit und dem Leseverständnis im regulären leistungsheterogenen Unterricht aufzuzeigen, um auf dieser Basis die grundlegende Einsatzfähigkeit der Tests in der dritten und vierten Klassenstufe zu beurteilen. Dafür wird erstens die Paralleltest-Reliabilität zu allen Messzeitpunkten geprüft. Die eingesetzten Testverfahren werden in einer Onlineplattform administriert, die pro Messzeitpunkt durch eine zufällige Ziehung aus einem begrenzten Itempool für jedes Kind eine neue Parallelform erstellt (Jungjohann, Diehl, Mühling & Gebhardt, 2018). Für alle Levumi-Lesetests werden Paralleltest-Reliabilitäten von mindestens $r = .80$ (Döring & Bortz, 2016) erwartet. Zweitens werden die beobachteten Punktrohwerte deskriptiv beschrieben. Es wird vermutet, dass die Schülerinnen und Schüler der vierten Jahrgangsstufe zu jedem Messzeitpunkt mehr Aufgaben lösen als die der dritten Jahrgangsstufe. Drittens werden die Zuwachsraten im Längsschnitt untersucht. Es wird angenommen, dass sich distinkte Entwicklungen mittels Methoden der Lernverlaufsdiagnostik nachzeichnen lassen. Die dritte und vierte Klassenstufe werden dabei getrennt voneinander betrachtet, um die Unterschiede in den Lernverläufen zwischen den Jahrgängen zu beobachten. In Anlehnung an die Argumentation von Müller et al. (2013) wird in der vierten Klassenstufe eine geringere Lernentwicklung in der Leseflüssigkeit und eine höhere im Satzverständnis erwartet. Zusätzlich werden unterschiedlich starke Anstiege in den Lernverläufen auf Ebene der Schülerinnen und Schüler vermutet. Viertens folgt eine differenzierte Betrachtung der Leseentwicklung unterschiedlicher Leistungsgruppen. Für die Untersuchung unterschiedlicher Leistungsgruppen werden, angelehnt an frühere Forschungsarbeiten (Hasbrouck & Tindal, 2006; Walter, 2010 b), Perzentilgruppen gebildet. In Anlehnung an Silberglitt und Hintze (2007) werden in den unteren und oberen Perzentilen flachere Entwicklungskurven angenommen. Gleichzeitig werden vor dem Hintergrund der Ergebnisse von Walter (2010 b) bei den Schülerinnen und Schülern mit den schwächsten Leseleistungen geringere Lösungshäufigkeiten vermutet. Außerdem wird erwartet, dass die Tests über die Grenzen der Jahrgangsstufen hinweg einsatzfähig sind. Die Prüfung dieser Kriterien kann als Facette einer externalen Validierung begriffen werden (Klauer, 2011).

4 Methode

4.1 Stichprobenbeschreibung

An der Studie nahm eine inklusive zweizügige Partnergrundschule in städtischer Lage zur Erprobung der kostenlosen Onlineplattform Levumi (Gebhardt et al., 2016) teil. Die Stichprobe umfasst N = 90 Schülerinnen und Schüler (43 Mädchen und 47 Jungen). Zum ersten Messzeitpunkt (Oktober 2017) waren die teilnehmenden Schülerinnen und Schüler im Schnitt 9.2 (*SD* = 0.69) Jahre alt. Elf Schülerinnen und Schüler hatten einen Migrationshintergrund (Kind oder mindestens ein Elternteil im Ausland geboren). Zwei Schülerinnen und Schüler hatten einen diagnostizierten sonderpädagogischen Förderbedarf (1 × Lernen, 1 × Sprache), zwei andere hatten einen Fluchthintergrund und befanden sich zu Beginn der Studie im Deutscherwerb. Für alle Teilnehmenden lag das Einverständnis der Erziehungsberechtigten vor und die Teilnahme war freiwillig. Zwei Drittklässler fehlten bei mehr als der Hälfte der Erhebungen und wurden aus der Studie genommen. Die beiden Jahrgänge ($n_{Stufe\,3} = 47$, $n_{Stufe\,4} = 41$) waren in jeweils zwei Klassenverbände aufgeteilt. Um den Einfluss der Klasse auf die einzelnen Ergebnisse nachzuvollziehen und damit mögliche Effekte auf der Ebene der Klassenzusammensetzung sowie der Lehrkräfte zu kontrollieren, wurden die Intra-Klassenkorrelationen (ICC) für alle verwendeten Tests bestimmt. Die ICC quantifizieren die Bedeutsamkeit einer Gruppenzugehörigkeit, also hier der Zugehörigkeit zu den Klassen (Zurbriggen, 2016). Es zeigt sich, dass im Mittel etwa 20 % ($ICC_{Silbenlesen} = 14{,}5\,\%$, $ICC_{Wörterlesen} = 28{,}7\,\%$, $ICC_{Pseudowörterlesen} = 16{,}4\,\%$, $ICC_{Satzlesen} = 17{,}6\,\%$) der individuellen Variabilität der Testergebnisse auf die Klassenzugehörigkeit zurückgeführt werden kann, was für relativ starke Klasseneffekte bei den Leseleistungen spricht.

4.2 Studiendesign

Zu Beginn der Studie war die Plattform allen Lehrkräften und Lernenden unbekannt. Um eine mögliche formative Wirkung der Lernverlaufsdiagnostik zu verhindern, bekamen die teilnehmenden Lehrkräfte während der Testphase keinen Einblick in die Ergebnisse. Nach der Studie erhielten die Lehrkräfte die Ergebnisse ihrer Schülerinnen und Schüler sowie Fortbildungen. In Absprache mit der Schulleitung wurden jeweils die Wochen vor den Ferien als Messwochen definiert, um rückläufige Effekte durch die Ferien zu vermeiden. Die Messungen fanden in den Kalenderwochen KW 41 (Herbst), KW 50 (Winter), KW 11 (Frühling) und KW 24 (Sommer) vormittags während individualisierter Lernzeiten statt. Jeder Klasse wurde ein fester Testtag zugeteilt, sodass die Messabstände bei allen Kindern gleich waren. Aus methodischer Sicht wäre ein gleichmäßiger Abstand zwischen den Messungen wünschenswert gewesen. Dies war aber aufgrund der Planung der Schule sowie der Ferienzeiten nicht möglich und spiegelt die schulischen Einsatzbedingungen der Lernverlaufsdiagnostik wider. In den größeren Messabständen lagen mehr einzelne Ferientage (Feiertage, Karneval). Im Schnitt lagen 12 Wochen zwischen den Messzeitpunkten. Pro Klassenverband standen in den Messwochen jeweils drei Schulstunden und eine geschulte Testerin zur Verfügung. Zu jedem Messzeitpunkt bearbeiteten die Teilnehmenden drei aufeinanderfolgende Leseflüssigkeitstests als Einzeltests und einen Gruppentest zum Leseverständnis im schuleigenen Computerraum. Die Leseflüssigkeitstests wurden stets in derselben Reihenfolge und in der vermutlichen theoretischen Schwierigkeit steigend (Silben, Wörter, Pseudowörter) administriert.

4.3 Erhebungsinstrumente

Alle eingesetzten Lernverlaufstests sind in der Onlineplattform Levumi kostenlos eingebettet und auf mehreren Niveaustufen verfügbar. Die Niveaustufen der Lesetests sind nach linguistischen Regeln definiert und generieren die Schwierigkeit der Aufgaben. Um die Vergleichbarkeit der vier Tests zu sichern, wurden in der

vorliegenden Studie alle auf derselben Niveaustufe (N4) administriert. Nach der Registrierung werden die Tests digital im Browser durchgeführt. Sie sind auch als Papierversion frei erhältlich (Jungjohann, Diehl & Gebhardt, 2019; Jungjohann & Gebhardt, 2019). Die digitalen Parallelformen werden über eine zufällige Ziehung aus fest definierten Itempools generiert. Dies erlaubt eine hohe Anzahl von Testwiederholungen ohne Erinnerungseffekte (Klauer, 2011).

Die theoriegeleiteten Konstruktionen der Leseflüssigkeitstests berücksichtigen je nach Testart (Silbenlesen, Wörterlesen, Pseudowörterlesen) unterschiedliche Leseteilkompetenzen (Jungjohann & Gebhardt, 2018). Die Tests dauern 60 Sekunden und benötigen eine lesekompetente Person zur Beurteilung von Lesefehlern. In den unabhängigen Itempools existieren 134 Items für den Silbentest, 61 Items für den Wörtertest und 189 Items für den Pseudowörtertest. Die Items werden pro Testdurchführung zufällig (ohne Zurücklegen) gezogen, sodass hypothetisch alle Items bearbeitet werden können. Die durchschnittlich Lesenden erreichen am Ende der Grundschule eine schwierigkeitsabhängige Bearbeitungszeit von 1,5 bis 3 Sekunden pro Item und bearbeiten somit in 60 Sekunden 20 bis 40 Items in der Leseflüssigkeit.

Der Test zum Leseverständnis präsentiert einzelne Sätze hintereinander, in denen jeweils ein Wort fehlt (Jungjohann, DeVries, Mühling et al., 2018). Der Itempool enthält 60 Items. Mithilfe von vier Auswahlwörtern (1 × richtiges Zielwort, 3 × Ablenker) konstruieren die Lesenden den Sinn der Sätze in Stillarbeit. Berücksichtigt werden unterschiedliche Argument-Prädikat-Strukturen, um theoriegeleitet syntaktische Schwierigkeitsstufen der deutschen Sprache zu repräsentieren. Zum Zeitpunkt der Studie dauerte der Test acht Minuten. Nach Prüfung der Reliabilität wurde der Test unter gleichbleibender Reliabilität aus ökonomischen Gründen auf fünf Minuten verkürzt. Die übliche Bearbeitungszeit pro Item liegt bei etwa 10 bis 15 Sekunden.

5 Ergebnisse

5.1 Reliabilität der eingesetzten Lernverlaufstests für die vorliegende Stichprobe

Modellanpassung und Reliabilität für die vorliegende Stichprobe werden basierend auf einer Item-Response-Theorie-Skalierung mit dem R-Paket *pairwise* (Heine, 2019) geschätzt. Dabei

Tab. 1 Korrelationen der Lesetests zwischen benachbarten Messzeitpunkten

	Stufe 3			Stufe 4		
	T1/T2	T2/T3	T3/4	T1/T2	T2/T3	T3/4
	$r\ (KI_{low}/KI_{high})$	$r\ (KI_{low}/KI_{high})$	$r\ (KI_{low}/KI_{high})$	$r\ (KI_{low}/KI_{high})$	$r\ (KI_{low}/KI_{high})$	$r\ (KI_{low}/KI_{high})$
Silbenlesen	.85*** (0.74/0.94)	.80*** (.66/.89)	.79*** (0.61/0.89)	.85*** (0.69/0.94)	.72*** (0.51/0.98)	.82*** (0.66/0.90)
Wörterlesen	.87*** (0.77/0.92)	.87*** (0.77/0.93)	.86*** (0.76/0.92)	.85*** (0.72/0.92)	.77*** (0.60/0.88)	.87*** (0.75/0.92)
Pseudowörterlesen	.87*** (0.77/0.93)	.84*** (0.73/0.91)	.85*** (0.73/0.91)	.76*** (0.58/0.87)	.89*** (0.80/0.94)	.87*** (0.76/0.93)
Satzlesen	.83*** (0.71/0.91)	.89*** (0.80/0.94)	.90*** (0.81/0.95)	.81*** (0.65/0.91)	.81*** (0.64/0.90)	.86*** (0.75/0.92)

Anmerkung: T1 = Herbst 2017, T2 = Winter 2017, T3 = Frühling 2018, T4 = Sommer 2018, KI = Konfidenzintervall, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

wurde ein konditionaler Maximum Likelihood Ansatz zur Parameterschätzung verwendet. Zur robusten Schätzung von Parametern am Rand der Verteilung wurden gewichtete Schätzer (weighted likelihood estimators; WLE) verwendet. In Anlehnung an Adams (2005) weisen die Tests zum ersten Messzeitpunkt sehr geringe Messfehleranteile auf ($WLE\text{-}Rel._{\text{Silben}} = .96$, $WLE\text{-}Rel._{\text{Wörter}} = .96$, $WLE\text{-}Rel._{\text{Pseudowörter}} = .92$, $WLE\text{-}Rel._{\text{Satzlesen}} = .96$).

Mittels Pearson-Produkt-Moment-Korrelation wird die Paralleltest-Reliabilität zwischen den benachbarten Messzeitpunkten geprüft. In Tabelle 1 sind die Korrelationen inklusive des oberen und unteren Konfidenzintervalls getrennt nach Testart und Klassenstufen abgebildet. Alle Korrelationen liegen zwischen $r = .72 - .90$ und sind mit $p < .001$ signifikant.

5.2 Punktrohwerte in der dritten und vierten Jahrgangsstufe

Um die Lernentwicklung der Schülerinnen und Schüler über das Schuljahr hinweg zu beschreiben, werden die Punktrohwerte der Stichproben abgebildet. Dafür wurde die Anzahl der richtig gelösten Items (Summenwerte) separat für die einzelnen Tests analysiert. In Tabelle 2 sind die arithmetischen Mittel ($\bar{x}$) und die Standardabweichung der Summenwerte aller vier Lesetests zu den vier Messzeitpunkten getrennt nach Klassenstufe dargestellt. Zusätzlich werden die Effektstärken (Cohen's *d*) zwischen dem ersten und vierten Messzeitpunkt angeführt, um die Größen der Mittelwertsunterschiede zu quantifizieren.

Tabelle 2 zeigt, dass das arithmetische Mittel der Summenwerte bis auf eine Ausnahme kontinuierlich über alle Messzeitpunkte in beiden Schulstufen hinweg ansteigt. Die Ausnahme stellt der Test Pseudowörterlesen zu den Messzeitpunkten Frühling und Sommer 2018 in Klassenstufe 3 dar ($\bar{x}_{Frühling} = 22.93$ ($SD = 8.77$); $\bar{x}_{Sommer} = 22.85$ ($SD = 8.87$)). Hier kann kein Lernanstieg beobachtet werden. Die Effekte für die Tests Silbenlesen, Wörterlesen und Satzlesen sind in Klassenstufe 3 höher ausgeprägt als im vierten Jahrgang. Für das Pseudowörterlesen liegt ein höherer Effekt für die Klassenstufe 4 vor.

Tab. 2 Arithmetisches Mittel, Standardabweichung und Effektstärken zu allen vier Messzeitpunkten getrennt nach Klassenstufe

	Stufe 3					Stufe 4				
	Herbst 2017	Winter 2017	Frühling 2018	Sommer 2018	Effektstärke[1]	Herbst 2017	Winter 2017	Frühling 2018	Sommer 2018	Effektstärke
	$\bar{x}$ (SD)	$\bar{x}$ (SD)	$\bar{x}$ (SD)	$\bar{x}$ (SD)	d_{Cohen}	$\bar{x}$ (SD)	$\bar{x}$ (SD)	$\bar{x}$ (SD)	$\bar{x}$ (SD)	d_{Cohen}
Silbenlesen	34.48 (12.52)	42.46 (14.52)	45.64 (14.02)	47.60 (12.8)	1.56	42.92 (14.55)	48.71 (11.58)	55.00 (11.74)	58.95 (12.03)	1.36
Wörterlesen	27.30 (12.09)	31.80 (13.45)	35.48 (13.23)	38.0 (12.38)	1.54	40.67 (11.80)	41.82 (10.80)	46.12 (10.50)	48.66 (11.66)	1.11
Pseudowörterlesen	16.72 (7.34)	19.72 (8.74)	22.93 (8.77)	22.85 (8.87)	1.05	22.88 (7.20)	24.80 (7.78)	27.93 (8.65)	30.79 (9.03)	1.37
Satzlesen	27.11 (13.56)	33.55 (15.17)	36.89 (15.63)	41.24 (14.64)	1.93	39.00 (14.60)	44.71 (11.90)	51.27 (10.37)	53.85 (9.25)	1.31

Anmerkung: [1] Die Effektstärke wurde von Messzeitpunkt 1 (Herbst 2017) auf Messzeitpunkt 4 (Sommer 2018) berechnet.

Abbildung 1 stellt die Lernverläufe der beiden Jahrgänge gemessen an den durchschnittlichen Summenwerten (= arithmetisches Mittel) dar. Als Fehlerbalken wurde das 95 %-Konfidenzintervall verwendet. Beide Gruppen zeigen über das Schuljahr einen signifikanten Lernzuwachs zwischen dem ersten (Herbst 2017) und letzten (Sommer 2018) Messzeitpunkt. Der Unterschied zwischen den Messzeitpunkten wurde mittels robust gepaarter t-Tests berechnet, um die paarweise Änderung darzustellen (Stufe 3: Silben $t(35) = -9.39$, $p < .001$; Wörter $t(44) = -10.35$, $p < .001$; Pseudowörter $t(44) = -7.07$, $p < .001$; Satzlesen $t(38) = -12.03$, $p < .001$; Stufe 4: Silben $t(21) = -6.38$, $p < .001$; Wörter $t(36) = -6.72$, $p < .001$; Pseudowörter $t(36) = -8.34$, $p < .001$; Satzlesen $t(38) = -8.17$, $p < .001$).

Im Test Satzlesen sind die Unterschiede in den Leistungen deutlich und die Konfidenzintervalle überlappen sich zu keinem Messzeitpunkt. Die Leistungen beim Pseudowörterlesen liegen hingegen nahe beieinander. Abbildung 1 zeigt, dass die Lernverläufe überwiegend parallel verlaufen und einen kontinuierlichen Anstieg vorweisen.

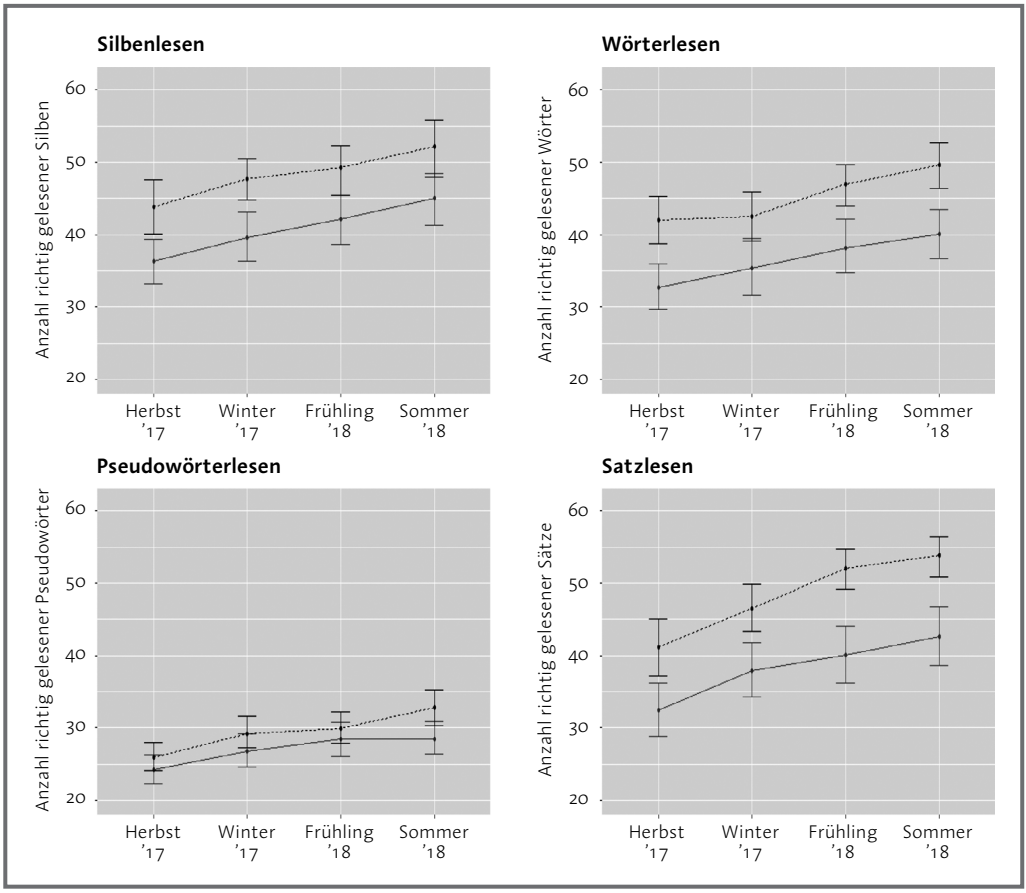

Abb. 1 Entwicklung der durchschnittlichen Summenscores mit Fehlerbalken über das Schuljahr 2017/18 getrennt nach Klassenstufen

Anmerkungen: Grauer Lernverlauf als Linie = Klassenstufe 3; Schwarzer Lernverlauf als Punktlinie = Klassenstufe 4

5.3 Zuwachsraten innerhalb der Klassenstufen

Um die Lernverläufe beider Jahrgangsstufen miteinander zu vergleichen, wurde eine lineare Regression über die vier Messzeitpunkte pro Test berechnet. Diese erlaubt die Schätzung des mittleren substanziellen Lernzuwachses. Die Ergebnisse sind in Tabelle 3 dargestellt.

Tab. 3 Lineare Regressionen der Lernverläufe nach Klassenstufe

	Stufe 3			Stufe 4		
	Intercept	Steigung (in 12 Wochen)	Bestimmtheitsmaß	Intercept	Steigung (in 12 Wochen)	Bestimmtheitsmaß
	$a\ (SE_a)$	$b\ (SE_b)$	R^2	$a\ (SE_a)$	$b\ (SE_b)$	R^2
Silbenlesen	31.76 (2.49)	4.33*** (0.94)	0.1111	37.93 (2.78)	5.41*** (0.98)	0.1797
Wörterlesen	24.18 (3.59)	3.59*** (0.84)	0.0918	37.28 (2.82)	2.18*** (0.80)	0.0746
Pseudowörterlesen	15.15 (1.55)	2.15*** (0.55)	0.0766	19.89 (1.59)	2.68*** (0.58)	0.1203
Satzlesen	23.30 (2.71)	4.57*** (1.00)	0.1064	34.45 (2.29)	5.11*** (0.82)	0.2004

Anmerkungen: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

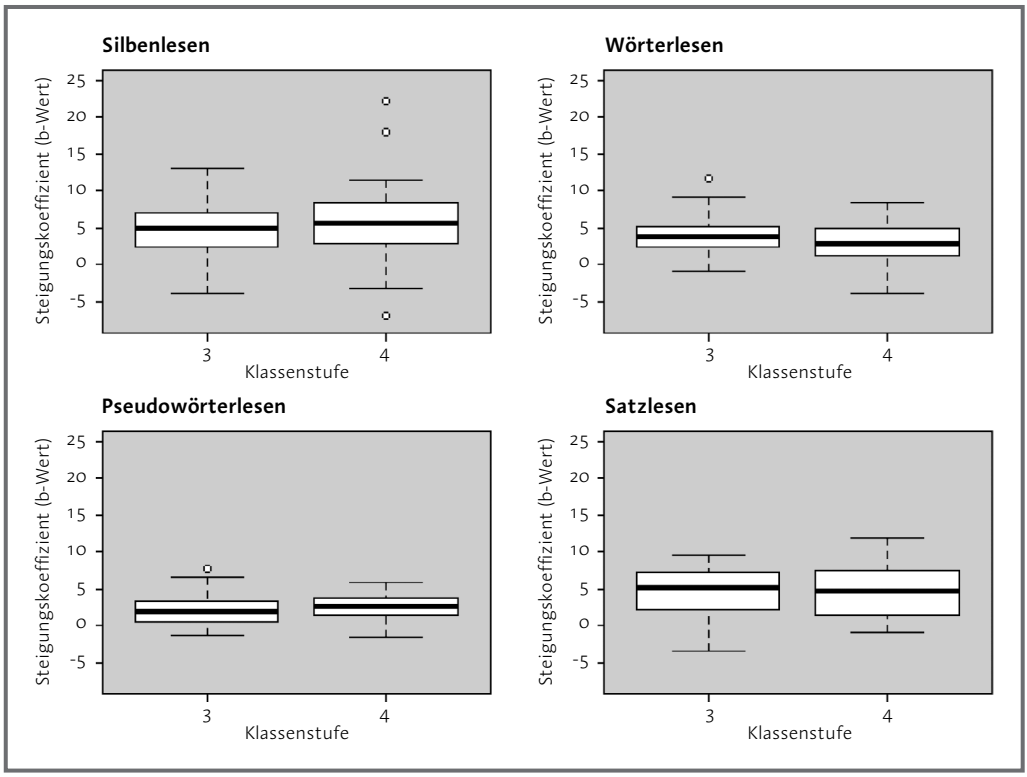

Abb. 2 Verteilung der Steigungskoeffizienten (b-Werte) auf Ebene der Schülerinnen und Schüler getrennt nach Klassenstufe und Testart

In allen Tests zeigt die dritte Jahrgangsstufe einen signifikant geringeren Ausgangswert als die vierte Jahrgangsstufe. Der Unterschied zwischen den Klassenstufen wurde mittels ungepaarter t-Tests berechnet (Silben: $t(43.45) = -2.45$, $p < .001$; Wörter: $t(82.87) = -5.18$, $p < .01$; Pseudowörter $t(82.75) = -3.92$, $p < .001$; Satzlesen $t(78.01) = -3.83$, $p < .001$). Die Steigungen der Lernverläufe sind bei beiden Klassen im Mittel ähnlich hoch und liegen zwischen 2.23 Pseudowörtern Wachstum in 12 Wochen in der dritten Klasse und 5.49 Silben in 12 Wochen in der vierten Klasse. Die Varianzaufklärung R^2 deutet aber auf ein uneinheitliches Entwicklungsverlaufsmuster hin.

Auf der Ebene der Schülerinnen und Schüler zeigen sich in beiden Klassenstufen unterschiedlich starke Steigungen in den Lernverläufen, deren Verteilungen in Abbildung 2 als Boxplot dargestellt sind. Die Boxplots in Abbildung 2 zeigen die Verteilungen der b-Werte getrennt nach Testart und Klassenstufen. Gemessen an den individuellen b-Werten zeigen zwischen zwei und sieben Schülerinnen und Schüler pro Klassenstufe (4.26 – 14.80 %) negative Lernverläufe mit $b = -0.8 - -4.0$ (Silben: $n_{Klasse\,3}$: 4 (8.51 %), $n_{Klasse\,4}$: 5 (12.20 %); Wörter: $n_{Klasse\,3}$: 2 (4.26 %), $n_{Klasse\,4}$: 4 (9.76 %); Pseudowörter: $n_{Klasse\,3}$: 7 (14.80 %), $n_{Klasse\,4}$: 4 (9.76 %); Satzlesen: $n_{Klasse\,3}$: 6 (12.77 %), $n_{Klasse\,4}$: 4 (9.76 %)).

5.4 Entwicklung der Lesefähigkeit nach Leistungsgruppen

Für eine differenzierte Betrachtung der Lernentwicklungen von unterschiedlich starken Leserinnen und Lesern sind die Zuwachsraten nach Leistungsperzentilen in Anlehnung an die Studie von Hasbrouck und Tindal (2006) aufgeteilt. Dadurch wird zusätzlich das individuelle Ausgangsniveau der Schülerinnen und Schüler berücksichtigt. Exemplarisch werden die Ergebnisse für den Leseflüssigkeitstest Wör-

Tab. 4 Lineare Regressionen im Test Wörterlesen nach Perzentilgruppen[1] getrennt nach Klassenstufe

Perzentil	Stufe 3				Stufe 4			
	$N_{Perzentil}$	Intercept	Range Intercept	Steigung (in 12 Wochen)	$N_{Perzentil}$	Intercept	Range Intercept	Steigung (in 12 Wochen)
		a (SE_a)	*a*	*b* (SE_b)		*a* (SE_a)	*a*	*b* (SE_b)
.10	6	7.08 (2.68)	2 – 14	3.47** (0.98)	4	16.55 (4.10)	12 – 23	2.44 (1.47)
.25	7	12.57 (2.11)	16 – 18	3.94*** (0.78)	6	24.58 (4.33)	24 – 35	4.60** (1.58)
.50	10	18.80 (2.20)	20 – 26	4.75*** (0.81)	10	34.25 (1.64)	36 – 42	3.18*** (0.62)
.75	11	28.20 (2.67)	27 – 32	2.61* (0.97)	10	39.58 (2.01)	43 – 47	3.97*** (0.73)
.90	7	33.57 (2.96)	34 – 41	4.24*** (1.08)	6	49.65 (2.51)	48 – 56	1.13 (0.93)
1.00	5	48.01 (3.06)	42 – 60	2.74* (1.15)	4	59.25 (2.65)	57 – 61	-0.85 (0.97)

Anmerkungen: [1] Die Perzentilgruppen wurden anhand des Intercepts zum ersten Messzeitpunkt Herbst 2017 gebildet, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

terlesen in Tabelle 4 und für den Leseverständnistest Satzlesen in Tabelle 5 abgebildet. Die Regressionen in den Tests Silben- und Pseudowörterlesen sind ähnlich wie die des Leseflüssigkeitstests Wörterlesen, sodass an dieser Stelle auf eine detaillierte Abbildung verzichtet wird. Die Leistungsperzentile werden anhand des Intercepts zum ersten Messzeitpunkt gebildet. Schülerinnen und Schüler ohne Messwert zum ersten Messzeitpunkt wurden nicht mit in die Verteilung nach Perzentilen eingeschlossen.

Der Test Wörterlesen misst in allen Perzentilgruppen der Klassenstufe 3 im Mittel signifikant steigende Lernzuwächse. Die geringsten Lernzuwächse wurden in dieser Klassenstufe in der .75-Perzentilgruppe sowie in der 1.0-Perzentilgruppe gemessen ($b_{.75\text{-Perzentil}} = 2.61$, $b_{1.0\text{-Perzentil}} = 2.74$). In Klassenstufe 4 wurden keine signifikanten Lernzuwächse in der unteren (.10-Perzentil) und in den beiden oberen Perzentilgruppen (.90- und 1.0-Perzentil) beobachtet. Im .25-, .50- und .75-Perzentil steigen die Lernzuwächse dieser Klassenstufe signifikant. In Klassenstufe 4 werden die geringsten Lernzuwächse in den Randgruppen beobachtet ($b_{.10\text{-Perzentil}} = 2.44$, $b_{.90\text{-Perzentil}} = 1.13$, $b_{1.0\text{-Perzentil}} = -0.85$).

Der Leseverständnistest Satzlesen misst wie der Test Wörterlesen in der dritten Klassenstufe in allen Subgruppen signifikante Lernzuwächse. In der vierten Klassenstufe zeigen sich signifikante Lernentwicklungen im .10- bis zum .75-Perzentil. In den oberen Perzentilgruppen mit den höchsten Ausgangsniveaus wurden nicht signifikante bzw. keine Lernentwicklungen beobachtet ($b_{.90\text{-Perzentil}} = 0.93$, $b_{1.0\text{-Perzentil}} = 0.01$). Die mittleren Ausgangswerte zeigen, dass die Lernenden bereits zum ersten Messzeitpunkt fast alle Testitems in der Bearbeitungszeit korrekt lösen. In beiden Klassenstufen werden bei den Lesenden mit dem höchsten Ausgangsniveau die geringsten Lernanstiege gemessen ($b_{Klasse\,3} = 2.28$, $b_{Klasse\,4} = -0.85$).

Tab. 5 Lineare Regressionen im Test Satzlesen nach Perzentilgruppen[1] getrennt nach Klassenstufe

Perzentil	Stufe 3				Stufe 4			
	$N_{Perzentil}$	Intercept	Range Intercept	Steigung (in 12 Wochen)	$N_{Perzentil}$	Intercept	Range Intercept	Steigung (in 12 Wochen)
		a (SE_a)	*a*	*b* (SE_b)		*a* (SE_a)	*a*	*b* (SE_b)
.10	5	4.02 (2.96)	3–10	3.19* (1.11)	4	10.63 (5.00)	5–20	6.03** (1.83)
.25	6	9.58 (3.43)	13–19	5.28*** (0.90)	6	17.98 (3.47)	22–26	8.40*** (1.26)
.50	11	18.36 (2.47)	20–26	5.26*** (0.90)	12	28.17 (2.19)	29–40	7.44*** (0.80)
.75	11	22.01 (3.11)	26–34	6.73*** (1.17)	8	44.67 (2.95)	42–54	3.12** (1.05)
.90	6	36.58 (2.35)	35–44	5.92*** (0.86)	6	54.35 (1.50)	55–58	0.93 (0.54)
1.00	5	50.48 (2.52)	47–61	2.28* (1.00)	3	59.80 (0.58)	59–61	0.01 (0.21)

Anmerkungen: [1] Die Perzentilgruppen wurden anhand des Intercepts zum ersten Messzeitpunkt Herbst 2017 gebildet, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

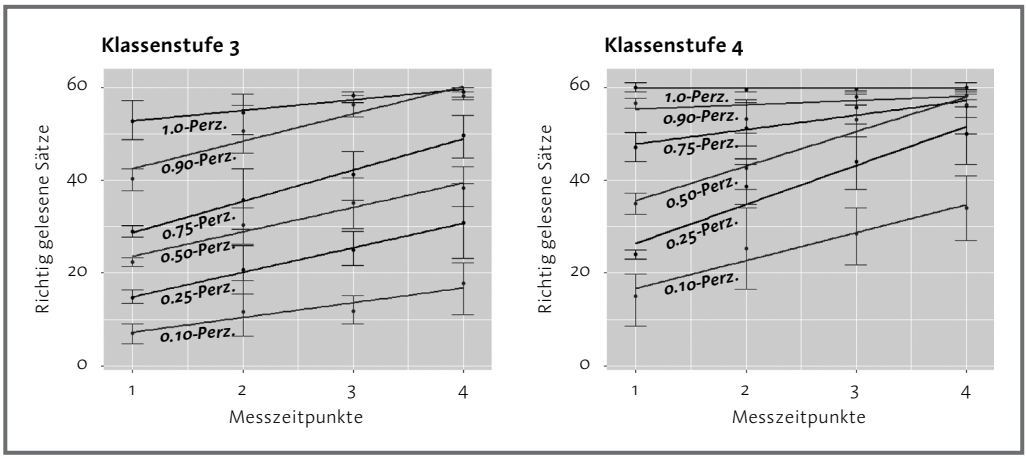

Abb. 3 Lineare Regressionen im Test Satzlesen nach Perzentilgruppen mit 95 % Konfidenzintervallen pro Messzeitpunkt

Der visuelle Vergleich der Lernverläufe zwischen den Jahrgangsstufen zeigt, dass sich die Mittelwerte der vierten Jahrgangsstufe zum ersten Messzeitpunkt mit denen der dritten Jahrgangsstufe zum vierten Messzeitpunkt auf dem gleichen Niveau befinden. Dies verdeutlicht die Abbildung 3 exemplarisch anhand der Regressionslinien im Test Satzlesen getrennt nach Perzentilgruppen. Bei allen Perzentilgruppen wurde in diesem Test kein signifikanter Unterschied zwischen den Mittelwerten der vierten Jahrgangsstufe zum ersten Messzeitpunkt und den Mittelwerten der dritten Jahrgangsstufe zum vierten Messzeitpunkt gefunden (.10-Perzentil: $t(6) = 0.57$, $p > .05$; .25-Perzentil: $t(10) = 1.51$, $p > .05$; .50-Perzentil: $t(21) = 1.337$, $p > .05$; .75-Perzentil: $t(15) = 0.83$, $p > .05$; .90-Perzentil: $t(10) = 1.93$, $p > .05$; 1.0-Perzentil: $t(4) = -1.23$, $p > .05$). Es kann somit von einer guten jahrgangsübergreifenden Anschlussfähigkeit ausgegangen werden, wenn der ausgewählte Test das jeweilige Leistungsspektrum der Lesenden abdeckt.

6 Diskussion

Die vorliegende Pilotstudie zeigt, dass standardisierte Lernverlaufstests zur Messung der Leseflüssigkeit und des basalen Leseverständnisses in der dritten und vierten Jahrgangsstufe im Längsschnitt eingesetzt werden können. In den Lernverläufen wurden erwartbare Deckeneffekte beobachtet, welche sich aber durch die Beschränkung auf eine Niveaustufe erklären lassen. Eine weitere Limitation stellt der pilotierende Charakter der Studie dar, da nur eine Schule gemessen wurde. In den Klassen können zwar ähnliche Bedingungen vermutet werden, aber die Studie erlaubt keine Kontrolle schulspezifischer Einflussfaktoren auf die Lernentwicklungen. Daher muss festgehalten werden, dass die dargestellten Ergebnisse nicht generalisierbar sind, da der Stichprobenumfang und die Untersuchungsbedingungen dies nicht zulassen. Die Pilotstudie bekräftigt aber die generelle Konstruktionsweise der vorgelegten Lernverlaufsdiagnostiktests und zeigt, dass diese stabil sind und auch geringe Lernfortschritte im unteren Leistungsbereich messen. Daher können die Tests zukünftig bei größeren Studien eingesetzt und geprüft werden, um grundsätzliche Fragen zur Lernentwicklung über die Zeit zu klären.

Die Studienergebnisse sprechen für eine längsschnittliche Einsetzbarkeit der Tests in der zweiten Hälfte der Grundschule. Die Paral-

leltest-Reliabilität zeigt akzeptable bis hohe Reliabilitäten ($r = .72 – .90$). Diese Ergebnisse ähneln den Befunden von Walter (2011 a), der Korrelationswerte zwischen $r = .74 – .84$ einer papierbasierten Lernverlaufsdiagnostik in der dritten und vierten Klassenstufe berichtet. Im Einklang mit den Ergebnissen von Müller et al. (2013) und Landerl und Wimmer (2008) wurden in dieser Studie in der dritten Klassenstufe im Mittel geringere Summenwerte bei allen Messzeitpunkten und Testarten erwartet und beobachtet als bei den Lernenden der vierten Klassenstufe. Bei den Lernenden der vierten Klassenstufe wurden zudem größere Lernzuwächse im Leseverständnis als bei den jüngeren Lernenden gefunden, wie es ebenfalls Walter (2011 a) beobachtet. Die mittleren Lernzuwächse fielen in beiden Klassenstufen im Test Satzlesen größer aus als beim Wort- und Pseudoworttest. Auch dies wurde in nationalen und internationalen Studien von Zuwachsraten im Leseverständnis und der Leseflüssigkeit zuvor beobachtet (Fuchs et al., 1993; Graney et al., 2009; Walter, 2010 b; 2011 a). Die Ähnlichkeit zwischen den Studienergebnissen sowie die zuvor geprüfte Homogenität der Items (Jungjohann, DeVries, Mühling et al., 2018; Mühling, Jungjohann & Gebhardt, 2019) deuten darauf hin, dass die Levumi-Testverfahren für die Beobachtung von Lernverläufen in der Zielgruppe reliabel und sensibel messen können.

Abweichende Studienergebnisse wurden bei den mittleren Zuwachsraten im Test Silbenlesen beobachtet. In beiden Klassenstufen erreichen die Lernenden ähnlich hohe Zuwachsraten wie beim Satzlesen (Klassenstufe 3: $b_{Silbenlesen} = 4.27$, $b_{Satzlesen} = 4.29$; Klassenstufe 4: $b_{Silbenlesen} = 5.49$, $b_{Satzlesen} = 5.06$). Die hohen Zuwachsraten beim Silbentest können über die Itemkonstruktion der Levumi-Tests begründet werden. Der Test Silbenlesen misst das phonologische Rekodieren anhand von einzelnen Silben (Jungjohann et al., 2019). Der Test Wörterlesen schließt zusätzlich den Abruf aus dem mentalen Lexikon mit ein und beim Pseudowörterlesen wird die Lesesynthese mit Silbensegmentierung beansprucht (Jungjohann & Gebhardt, 2018). Beim Wörter- und Pseudowörterlesen bestehen die Items überwiegend aus Wörtern mit zwei Silben und sind somit länger als die Items des Tests Silbenlesen. Für alle drei Leseflüssigkeitstests ergeben sich daher unterschiedliche Bearbeitungszeiten pro Item, was zu größeren Lernanstiegen bei gleicher Bearbeitungszeit führen kann. Diese Beobachtungen stärken die Argumente von Christ et al. (2013), dass die Testkonstruktion bei der Interpretation der Zuwachsraten berücksichtigt werden muss.

Bemerkenswert ist, dass in allen Klassen bei bis zu 15 % der Schülerinnen und Schüler neutrale oder negative Lernverläufe beobachtet wurden. Neutrale und negative Lernverläufe werden häufig in äußeren Leistungsperzentilen berichtet (Fuchs et al., 1993; Walter, 2011 a). Sowohl Silberglitt und Hintze (2007) als auch Walter (2010 b) berichten in ihren Studien, dass in den äußeren Leistungsperzentilen geringere Lernzuwächse gemessen wurden als in den mittleren Perzentilen. Das ist in dieser Studie in der dritten Klassenstufe erkennbar, wo es höhere Lernanstiege in den mittleren Perzentilen gibt als in den äußeren. Trotzdem waren alle Anstiege signifikant. In der vierten Klassenstufe wurden nur in den oberen beiden Perzentilen Deckeneffekte beobachtet. Die Erklärungsansätze sind je nach Test unterschiedlich. Im Satzlesen lösten die oberen beiden Perzentilgruppen der vierten Klasse bereits zum ersten Messzeitpunkt fast alle zur Verfügung stehenden Items. Für diese Lernenden entstehen die Deckeneffekte vermutlich testbedingt aufgrund der Testkonstruktion als Speedtest. Im Wörterlesen erreichen die Lernenden zum ersten Messzeitpunkt der Studie mit 50 bis 60 Items eine hohe und kaum verbesserungsfähige Vorlesegeschwindigkeit.

In Klassen mit einer breiten Leistungsheterogenität sind bei einem engen Lernverlaufsdiagnostiktest Deckeneffekte wahrscheinlich. Im Testsystem Levumi existieren zum einen Tests verschiedener Niveaustufen und zum anderen alternative Testbereiche mit komplexeren Kompetenzansprüchen. Für diese Untersuchung wurde zur besseren Vergleichbarkeit nur eine feste Niveaustufe gewählt. Im Feld wählen zumeist die Lehrkräfte eine Lernverlaufsdiagnostik für die einzelnen Lernenden aus. Bei auftretenden Deckeneffekten wären Hinweise auf die Anschlussfähigkeit zwischen den Niveaustufen in der Praxis hilfreich. Entsprechende Studien und Analysen stehen bisher noch aus. Während eine händische Auswahl den Lehrkräften mehr Freiheiten lässt, wäre auch eine adaptive Verbindung der Niveaustufen durch einen breiteren Test möglich. Anstelle einzelner Lesetests mit engen Itempools würde der adaptive Test dann Zugriff auf mehrere Itempools in unterschiedlichen Niveaustufen haben, aus denen adaptiv gezogen wird. Alle Levumi-Lesetests bereiten ein adaptives Testen vor, da pro Testart mehrere Itempools in unterschiedlichen Niveaustufen existieren. Im Fall von Deckeneffekten könnten zukünftig schwierigere Items aus höheren Niveaustufen automatisch ausgewählt und somit der Test an die Kompetenzen der Lernenden angepasst werden. Ein solches adaptives Testen könnte insbesondere im inklusiven Unterricht mit großen Leistungsunterschieden (Gebhardt, 2015) zwischen den Lernenden den Einsatz einer Lernverlaufsdiagnostik für die Lehrkräfte erleichtern.

Anmerkung

[I] Zu dem vorliegenden Beitrag sind der verwendete Datensatz sowie Syntax für R unter folgender Zitierung veröffentlicht: Jungjohann, J., Schurig, M. & Gebhardt, M. (2021). Pilotierung von Leseflüssigkeits- und Leseverständnistests zur Entwicklung von Instrumenten der Lernverlaufsdiagnostik. Ergebnisse einer Längsschnittstudie in der 3ten und 4ten Jahrgangsstufe/ Data & Syntax. https://doi.org/10.17605/OSF.IO/4SZY5

Literatur

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2–3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Anderson, S., Jungjohann, J. & Gebhardt, M. (2020). Effects of using curriculum-based measurement (CBM) for progress monitoring in reading and an additive reading instruction in second classes. *Zeitschrift für Grundschulforschung, 13,* 151–166. https://doi.org/10.1007/s42278-019-00072-5

Bos, W., Valtin, R., Hußmann, A., Wendt, H. & Goy, M. (2017). IGLU 2016: Wichtige Ergebnisse im Überblick. In A. Hußmann et al. (Hrsg.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich,* 13–28. Münster: Waxmann Verlag.

Christ, T. J., Zopluoglu, C., Monaghen, B. D. & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology, 51*(1), 19–57. https://doi.org/10.1016/j.jsp.2012.11.001

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. https://doi.org/10.1177/00224669030370030801

Deno, S. L., Mirkin, P. K. & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*(1), 36–45.

Deno, S. L., Fuchs, L. S., Marston, D. & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(4), 507–524.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin: Springer. https://doi.org/10.1007/978-3-642-41089-5

Förster, N., Kuhn, J.-T. & Souvignier, E. (2017). Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik, 9*(2), 116–122.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L. & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27–48.

Fuchs, L. S., Fuchs, D., Hosp, M. K. & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Gebhardt, M. (2015). Gemeinsamer Unterricht von Schülerinnen und Schülern mit und ohne sonderpädagogischen Förderbedarf. Ein empirischer Überblick. In E. Kiel (Hrsg.), *Inklusion im Sekundarbereich,* 39–52. Stuttgart: Kohlhammer.

Gebhardt, M., Heine, J.-H., Zeuch, N. & Förster, N. (2015). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen. *Empirische Sonderpädagogik, 7*(3), 206–222.

Gebhardt, M., Diehl, K. & Mühling, A. (2016). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www.LEVUMI.de. *Zeitschrift für Heilpädagogik, 67*(10), 444–453.

Good, R. H., Simmons, D. C. & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257–288. https://doi.org/10.1207/S1532799XSSR0503_4

Graney, S. B., Missall, K. N., Martínez, R. S. & Bergstrom, M. K. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology, 47*(2), 121–142. https://doi.org/10.1016/j.jsp.2008.12.001

Grapin, S. L., Kranzler, J. H., Waldron, N., Joyce-Beaulieu, D. & Algina, J. (2017). Developing local oral reading fluency cut scores for predicting high-stakes test performance. *Psychology in the Schools, 54*(9), 932–946. https://doi.org/10.1002/pits.22035

Hartke, B. (2017). Gelingende Inklusion – das Rügener Inklusionsmodell (RIM). In B. Hartke (Hrsg.), *Handlungsmöglichkeiten Inklusion,* 11–19. Stuttgart: Kohlhammer.

Hasbrouck, J. & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644. https://doi.org/10.1598/RT.59.7.3

Heine, J.-H. (2019). *Pairwise: Rasch Model Parameters by Pairwise Algorithm. R package version 0.4.4-5.1.* Abgerufen am 20. 10. 2019 von https://CRAN.R-project.org/package=pairwise

Hesse, I. & Latzko, B. (2011). *Diagnostik für Lehrkräfte.* 2. Aufl. Opladen: Budrich.

Hosp, M. K., Hosp, J. L. & Howell, K. W. (2007). *The ABCs of CBM. A practical guide to curriculum-based measurement.* New York: Guilford Press. Retrieved from http://www.loc.gov/catdir/enhancements/fy0704/2006027886-b.html

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., … Valtin, R. (Hrsg.) (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster: Waxmann Verlag. https://content-select.com/de/portal/media/view/5aa259f2-7384-45b8-9192-27cab0dd2d03

Jungjohann, J., DeVries, J. M., Gebhardt, M. & Mühling, A. (2018). Levumi. A web-based curriculum-based measurement to monitor learning progress in inclusive classrooms. In K. Miesenberger & G. Kouroupetroglou (eds.), *Computers Helping People with Special Needs,* 369–378. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-94277-3_58

Jungjohann, J., DeVries, J. M., Mühling, A. & Gebhardt, M. (2018). Using theory-based test construction to develop a new curriculum-based measurement for sentence reading comprehension. *Frontiers in Education, (3)*1. https://doi.org/10.3389/feduc.2018.00115

Jungjohann, J., Diehl, K., Mühling, A. & Gebhardt, M. (2018). Graphen der Lernverlaufsdiagnostik interpretieren und anwenden – Leseförderung mit der Onlineverlaufsmessung Levumi. *Forschung Sprache, 6*(2), 84–91. https://doi.org/10.17877/DE290R-19806

Jungjohann, J. & Gebhardt, M. (2018). Lernverlaufsdiagnostik im inklusiven Anfangsunterricht Lesen – Verschränkung von Lernverlaufsdiagnostik, Förderplanung und Wochenplanarbeit. In F. Hellmich, G. Görel & M. F. Löper (Hrsg.), *Inklusive Schul- und Unterrichtsentwicklung. Vom Anspruch zur erfolgreichen Umsetzung,* 160–172. Stuttgart: Kohlhammer.

Jungjohann, J., Gegenfurtner, A. & Gebhardt, M. (2018). Systematisches Review von Lernverlaufsmessung im Bereich der frühen Leseflüssigkeit. *Empirische Sonderpädagogik, 10*(1), 100–118.

Jungjohann, J., Diehl, K. & Gebhardt, M. (2019). *SiL-Levumi – Tests der Leseflüssigkeit zur Lernverlaufsdiagnostik – „Silben lesen" der Onlineplattform www.levumi.de.* Trier: ZPID. https://doi.org/10.23668/psycharchives.2462

Jungjohann, J. & Gebhardt, M. (2019). *SinnL-Levumi – Tests zum sinnkonstruierenden Satzlesen als Lernverlaufsdiagnostik – „Sinnkonstruierendes Satzlesen" der Onlineplattform www.levumi.de.* Trier: ZPID. https://doi.org/10.23668/psycharchives.2463

Klauer, K.J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik, 3*(3), 207–224.

Klicpera, C., Schabmann, A., Gasteiger-Klicpera, B. & Schmidt, B. (2017). *Legasthenie – LRS. Modelle, Diagnose, Therapie und Förderung.* München: Ernst Reinhardt Verlag. http://www.utb-studi-e-book.de/9783838548166

Landerl, K. & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography. An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150–161. https://doi.org/10.1037/0022-0663.100.1.150

Lenhard, W. & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses,* 1–8. Göttingen: Hogrefe.

Lenhard, W., Lenhard, A. & Schneider, W. (2017). *ELFE II – ein Leseverständnistest für Erst- bis Siebtklässler. Version II.* Hogrefe Schultests. Göttingen: Hogrefe.

Mayer, A. (2016). *Lese-Rechtschreibstörungen (LRS).* München: Ernst Reinhardt Verlag.

Mühling, A., Jungjohann, J. & Gebhardt, M. (2019). Progress monitoring in primary education using Levumi: A case study. In H. Lane, S. Zvacek & J. Uhomoihi (eds.), *Proceedings of the 11th International Conference on Computer Supported Education,* 137–144. Setúbal: Science and Technology Publications.

Müller, B., Križan, A., Hecht, T., Richter, T. & Ennemoser, M. (2013). Leseflüssigkeit im Grundschulalter. Entwicklungsverlauf und Effekte systematischer Leseförderung. *Lernen und Lernstörungen, 2*(3), 131–146. https://doi.org/10.1024/2235-0977/a000039

Muthén, B.O. & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371–402. https://doi.org/10.1037/1082-989X.2.4.371

Nese, J.F.T., Biancarosa, G., Anderson, D., Lai, C.-F., Alonzo, J. & Tindal, G. (2012). Within-year oral reading fluency with CBM: a comparison of models. *Reading and Writing, 25*(4), 887–915. https://doi.org/10.1007/s11145-011-9304-0

Richter, T., Isberner, M.-B., Naumann, J. & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern. *Zeitschrift für Pädagogische Psychologie, 26*(4), 313–331. https://doi.org/10.1024/1010-0652/a000079

Rosebrock, C. & Nix, D. (2017). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung.* Baltmannsweiler: Schneider Verlag.

Scheerer-Neumann, G. (2015). *Lese-Rechtschreib-Schwäche und Legasthenie. Grundlagen, Diagnostik und Förderung.* Stuttgart: Kohlhammer.

Shin, J., Deno, S.L. & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*(3), 164–172. https://doi.org/10.1177/002246690003400305

Shinn, M.R. (2007). Identifying students at risk, monitoring performance and determining eligibility within response to intervention: Research on educational need and benefit from academic intervention. *School Psychology Review, 36*(4), 601–617.

Silberglitt, B. & Hintze, J.M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*(4), 304–325.

Silberglitt, B. & Hintze, J.M. (2007). How much growth can we expect? A conditional analysis of R-CB growth rates by level of performance. *Exceptional Children, 42*(1), 795–819. https://doi.org/10.1177/001440290707400104

Voß, S. & Blumenthal, Y. (2020). Assessing the word recognition skills of German elementary students in silent reading – Psychometric properties of an item pool to generate curriculum-based measurements. *Education Sciences, 10*(2), 35. https://doi.org/10.3390/educsci10020035

Voß, S. & Gebhardt, M. (2017). Schwerpunktthema: Verlaufsdiagnostik in der Schule: Editorial. *Empirische Sonderpädagogik, 9*(2), 95–97.

Walter, J. (2008). Adaptiver Unterricht erneut betrachtet: Über die Notwendigkeit systematischer formativer Evaluation von Lehr- und

Lernprozessen und die daraus resultierende Diagnostik und Neudefinition von Lernstörungen nach dem RTI-Paradigma. *Zeitschrift für Heilpädagogik, 59*(6), 202 – 215.

Walter, J. (2010a). *LDL – Lernfortschrittsdiagnostik Lesen. Ein curriculumbasiertes Verfahren.* Göttingen: Hogrefe.

Walter, J. (2010b). Lernfortschrittsdiagnostik am Beispiel der Lesekompetenz (LDL): Messtechnische Grundlagen sowie Befunde über zu erwartende Zuwachsraten während der Grundschule. *Heilpädagogische Forschung, 36*(4), 162 – 176.

Walter, J. (2011a). Die Entwicklung eines auch computerbasiert einsetzbaren Instruments zur formativen Messung der Lesekompetenz. *Heilpädagogische Forschung, 37*(3), 106 – 126.

Walter, J. (2011b). Die Messung der Entwicklung der Lesekompetenz im Dienste der systematischen formativen Evaluation von Lehr- und Lernprozessen. *Zeitschrift für Heilpädagogik, 62*(6), 204 – 217.

Walter, J. (2013). *VSL. Verlaufsdiagnostik sinnerfassendes Lesen.* Göttingen: Hogrefe.

Walter, J. (2014). Lernfortschrittsdiagnostik Lesen (LDL) und Verlaufsdiagnostik sinnerfassenden Lesens (VSL): Zwei Verfahren als Instrumente einer formativ orientierten Lesediagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik,* 166 – 201. Göttingen: Hogrefe.

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

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

Yeo, S., Fearrington, J. Y. & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes. *Assessment for Effective Intervention, 37*(3), 147 – 158. https://doi.org/10.1177/1534508411420129

Zurbriggen, C. (2016). *Schulklasseneffekte. Schülerinnen und Schüler zwischen komparativen und normativen Einflüssen.* Wiesbaden: Springer VS.

Anschrift der Autorin und der Autoren

Dr. Jana Jungjohann
Prof. Dr. Markus Gebhardt
Universität Regensburg
Lehrstuhl für Lernbehindertenpädagogik einschließlich inklusiver Pädagogik
Sedanstr. 1
D-93055 Regensburg
E-Mail: jana.jungjohann@paedagogik.uni-regensburg.de
markus.gebhardt@paedagogik.uni-regensburg.de

Dr. Michael Schurig
Technische Universität Dortmund
Entwicklung und Erforschung inklusiver Bildungsprozesse
Emil-Figge-Str. 50
D-44227 Dortmund
E-Mail: michael.schurig@tu-dortmund.de