Grazer Leseverständnistest – GraLeV

Lisa Paleczek, Susanne Seifert, Annelise Franz, Sylvia Riedl & David Wohlhart

2023

Illustrationen von Heike Skringer

Handanweisung/ Manual

verfasst von Susanne Seifert & Lisa Paleczek

September 2023

Inhaltsverzeichnis

Einleitung

Der Grazer Leseverständnistest (GraLeV) wurde als kostenfreier Leseverständnistest in den Projekten RegioDiff (https://regionen-kennenlernen.uni-graz.at/de/ueberblick/) und RegiNaDiff (https://regional-nachhaltig-differenziert.uni-graz.at/de/) entwickelt und erprobt. Er liegt in einer Digital-Version (Tabletversion als App) und einer Print-Version vor. Derzeit frei zugänglich und nutzbar ist jedoch nur die Print-Version, weshalb sich die im vorliegenden Manual gemachten Angaben auf diese Version beziehen (für einen Vergleich zwischen beiden Versionen, siehe Seifert & Paleczek, 2022).

Ziel der Entwicklung des vorliegenden Testverfahrens war es, für Lehrpersonen sowie für Forscher*innen ein Status-Diagnostik-Instrument bereit zu stellen, welches zeitökonomisch im Schulalltag eingesetzt werden kann und zuverlässig die Leseverständnis-Fähigkeiten auf Wort-, Satz- und Textebene aller Kinder einer Klasse misst.

Die Entwicklung dieses Tests hat L. Paleczek im Projekt RegioDiff verantwortet. An der Konstruktion dieses Tests waren nicht nur die Autorinnen dieses Manuals (S. Seifert & L. Paleczek) beteiligt, sondern auch A. Franz, S. Riedl und D. Wohlhart.

Den Prozess der Normierung hat S. Seifert verantwortet. Unser Dank gilt den Studierenden, die uns tatkräftig bei der Dateneingabe, Datenauswertung und bei dem Verfassen von Rückmeldungen an die Lehrpersonen unterstützt haben. Danken möchten wir aber auch allen teilnehmenden Kindern, die die Aufgaben gelöst und deren Eltern uns ihre Zustimmung gegeben haben. Unser Dank gilt aber vor allem auch den mit uns kooperierenden Schulen, in denen sowohl die Schulleitungen als auch die Lehrpersonen uns stets unterstützt und uns die Notwendigkeit eines solchen Verfahrens immer wieder vor Augen geführt haben. Erst durch sie wurden die Entwicklung dieses Verfahrens und die umfangreiche Normierungsstichprobe sowie viele weitere Vorzüge des GraLeV möglich.

1. Testkonzept

1.1 Theoretischer Hintergrund: Überprüfung des Leseverständnisses im Deutschen

Der Grazer Leseverständnistest (GraLeV) dient der differenzierten Erfassung des Leseverständnisses in der dritten und vierten Schulstufe. Perfetti, Landi und Oakhill (2005) definierten Leseverständnis als die Fähigkeit, die Bedeutung eines geschriebenen Wortes, Satzes oder Textes zu verstehen. Diese Prozesse sind hierarchisch aufgebaut, d. h. das Wortverständnis ist die Voraussetzung für das Satzverständnis, und das Satzverständnis eine Voraussetzung für das Textverständnis ist (z. B. Mullis & Martin, 2015; Richter & Christmann, 2009). Alle drei Ebenen (Wort-, Satz- und Textverständnis) basieren jedoch auf anderen Fertigkeiten oder Kenntnissen wie Dekodieren, Sprachprozessen, Wortschatz und Vorwissen (Kintsch, 1998; Richter et al., 2009; Stahl & Hiebert, 2005).

Für die meisten Kinder beginnt der systematische Erwerb von Lesefähigkeiten mit dem Schuleintritt. Zunächst werden grundlegende Lesefertigkeiten (einschließlich Dekodierung, Leseverständnis auf Wort- und Satzebene) erworben. Danach werden zunehmend komplexere Lese- und Verstehensprozesse trainiert (spätestens ab der 4. Schulstufe), während Leseflüssigkeit und -geschwindigkeit zunehmen (Klicpera et al., 2017). Insbesondere die Fähigkeit, Texte beim Lesen zu verstehen, ist entscheidend, um das Lernen aus Texten zu ermöglichen (Schnotz, 1994) – eine Fähigkeit, die von Schüler*innen am Ende der Grundschule erwartet wird. Beim umfassenden Lesen von Texten

ist es notwendig, satzübergreifende Bedeutungseinheiten zu interpretieren und lokale und globale Kohärenz herzustellen (Richter & Christmann, 2009).

Tests zur Überprüfung des Leseverständnisses können sich in Bezug auf ihren Zweck, ihre Zielgruppe und ihre Form voneinander unterscheiden (Afflerbach, 2016). In Bezug auf die Zielgruppe unterscheiden sich die Tests vor allem im Hinblick auf das Alter der Kinder. Für Grundschulkinder gibt es im deutschsprachigen Raum mehrere Testverfahren, die zur Überprüfung des Leseverständnisses eingesetzt werden können (für einen Überblick: Lenhard, 2013; Paleczek & Seifert, 2019). Dabei ist zu unterscheiden zwischen den umfangreicheren Verfahren, mit denen die verschiedenen Teilprozesse der Lesekompetenz getestet werden (z. B. ELFE II: Lenhard et al., 2020), und den kürzeren Verfahren, die einen groben Überblick über das Leseniveau der Klasse geben und Schüler*innen mit spezifischen Defiziten schnell identifizieren (z. B. SLS 2-9: Mayringer & Wimmer, 2014). Weitere Tests, die im Rahmen von CBM (Curriculum Based Measurement) eingesetzt werden, dienen der Überprüfung und Begleitung des Lernfortschritts (z. B. VSL: Walter, 2013).

Neben der Zielgruppe und dem Zweck lassen sich die Messinstrumente zur Erfassung der Lesefähigkeiten auch hinsichtlich der Messmethode unterscheiden. Diese messen entweder das Leseverständnis als Produkt des Leseprozesses (z. B. Beantwortung von Multiple-Choice-Fragen nach dem Lesen eines Textes) oder direkt während des Leseprozesses (z. B. Ausfüllen von Lücken in einem Text). Es besteht ein enger Zusammenhang zwischen der Messmethode und der Kompetenzmessung (Keenan, Betjemann & Olson, 2008). So hat sich beispielsweise gezeigt, dass die Merkmale des Tests und die in einem Leseverständnistest verwendeten Items einen Einfluss darauf haben, wie stark oder wenig andere Fähigkeiten (z. B. Dekodierfähigkeit, Leseflüssigkeit und sprachliche Fähigkeiten, Vorwissen) das Testergebnis beeinflussen (Brasher, 2017).

Auf Wort- und Satzebene erfassen Tests das Leseverständnis als Produkt des Leseprozesses. Die Entscheidung, welches Wort zu einem Bild oder Satz passt (z. B. in ELFE II auf Wort- bzw. Satzebene: Lenhard et al., 2020), oder die Entscheidung, ob ein Satz richtig oder falsch ist (z. B. in SLS 2-9: Mayringer et al., 2014), sind häufige Aufgabenformate in deutschen Lesetests.

Auf der Textebene werden in Tests oft Fragen zum gelesenen Text gestellt (z. B. bei der Überprüfung der Bildungsstandards in Österreich). Bei diesem Aufgabenformat sind die Antworten stark auf das Sprachverständnis und die sprachlichen Fähigkeiten bezogen (Keenan et al., 2008). Die Schwierigkeit der Aufgabe variiert je nach Fragetyp (z. B. direkte Informationsbeschaffung, einfache Verknüpfung von Informationen, Bildung von Schlussfolgerungen) sowie nach dem verfügbaren Antwortmodus (Ja-Nein-Fragen vs. offene Fragen) (Guthrie, Seifert, Burnham, & Caplan, 1974). Diese Methode wird auch in vielen standardisierten Verfahren zur Messung des Leseverständnisses eingesetzt, wobei der Fragetyp kontrolliert wird und der Antwortmodus meist in Form eines Multiple-Choice-Formats vorliegt (z.B. ELFE II: Lenhard et al., 2020).

Nur wenige Verfahren im deutschsprachigen Raum messen das Leseverstehen während der Leseaufgabe selbst (d.h. durch das Abfragen von Antworten während des Lesens). Das Maze-Verfahren ist eine solche Messmethode. Die Schüler*innen müssen eine bestimmte Textpassage innerhalb eines bestimmten Zeitlimits lesen. Normalerweise gibt es an der Stelle jedes siebten Wortes im Text eine Lücke und die Schüler*innen müssen das Zielwort identifizieren, indem sie aus mehreren Optionen (darunter ein Zielwort und 2 Ablenkern) wählen. Diese Methode zur Messung des Leseverständnisses ist in englischsprachigen Ländern weit verbreitet (für einen Überblick: Wayman, Wallace, Wiley, Tichà,

& Espin, 2007). Hierbei wird das Leseverständnis durch eine Reihe von Fähigkeiten beeinflusst, insbesondere durch solche, die mit dem Wortschatz zusammenhängen, oder mit der Leseflüssigkeit und dem Dekodieren (wenn ein Zeitlimit vorgegeben ist) (Brasher, 2017; Kendeou, Papadopoulos, & Spanoudis, 2012; Muijselaar, Kendeou, de Jong, & van den Broek, 2017; Spear-Swerling, 2004), wobei die Dekodierfähigkeiten besonders in den unteren Schulstufen wichtig sind (Garcia & Cain, 2014). Um eine bessere Differenzierung zwischen niedrigem und hohem Leseverständnis zu ermöglichen, variieren die Tests in Bezug auf Zeitlimit, Textlänge, Schwierigkeitsgrad des Zielworts und Verwendung von Distraktoren (Conoyer et al., 2017). Die schnelle und einfache Auswertung und Durchführung werden als spezifische Vorteile des Maze-Ansatzes immer wieder hervorgehoben (Brasher, 2017). Die Textverständnis-Fähigkeit Fähigkeit der Schüler*innen (u.a. das Ziehen von Schlussfolgerungen) ist allerdings mit dieser Methode nur eingeschränkt möglich (Muijselaar et al., 2017). Um das Textverständnis gut und umfassend zu erfassen, ist daher eine Kombination aus beiden Methoden (Fragen zum Text stellen und Maze-Verfahren) sinnvoll.

1.2 Testaufbau

Der GraLeV ist ein umfassender Leseverständnistest für die Schulstufen 3 und 4. Er prüft die drei verschiedenen Ebenen des Leseverständnisses (Wort-, Satz- und Textebene) anhand von vier verschiedenen Subtests (Wort, Satz, Text I und Text II). In jedem der Subtest werden zunächst Beispielaufgaben präsentiert, bevor die Schüler*innen aufgefordert werden, den eigentlichen Subtest zu beginnen und ihn innerhalb einer vorgegebenen Zeit zu bearbeiten.

Subtest Wort

Für den Subtest Wort gibt es 12 Itemsets, die jeweils aus drei Zielwörtern (Items) und drei Ablenkern bestehen (phonologisch-graphematisch mit den Zielwörtern verwandt). Den Kindern werden drei Bilder und sechs Wörter vorgelegt. In jedem Itemset müssen die Schüler*innen die drei Zielwörter den drei Bildern zuordnen, indem sie eine Linie vom Wort zum Bild ziehen (siehe Abb. 1). Für diesen Subtest gibt es ein Zeitlimit von drei Minuten.

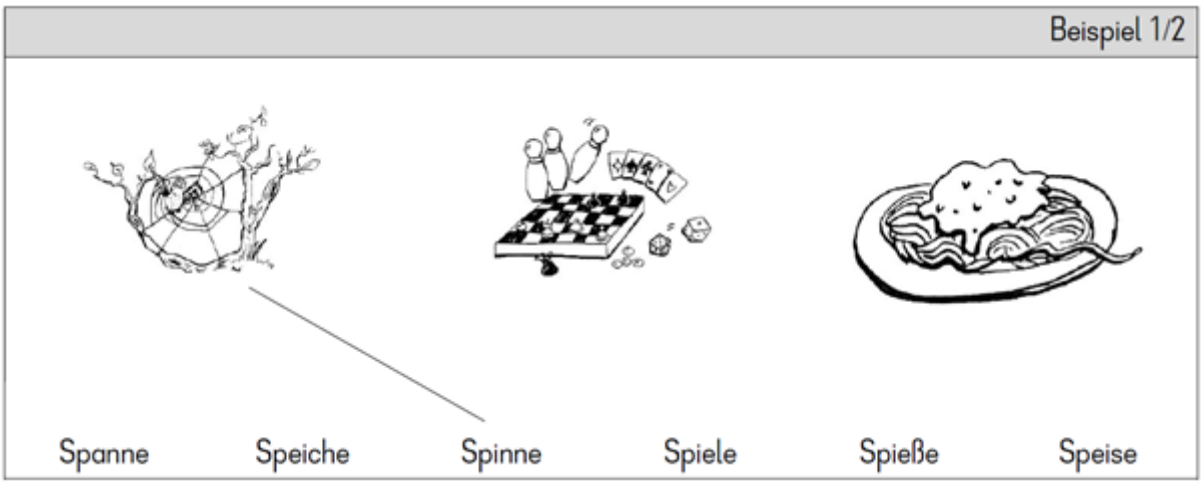

Abb. 1: Beispiel-Itemset des Subtests Wort im GraLeV.

Subtest Satz

Der Subtest Satz besteht aus 16 Items, die jeweils ein Situationsbild und vier Sätze enthalten. Ein Satz ist der Zielsatz, der zum Bild passt, und drei Sätze sind Distraktoren. Ein Distraktor zeigt syntaktische Nähe zum Zielsatz, der andere semantische Nähe und der letzte stellt eine Kombination beider dar. Nur der Zielsatz beschreibt genau die Situation auf dem Bild. Die Schüler*innen müssen den passenden Satz ankreuzen (siehe Abb. 2). Die Kinder haben drei Minuten Zeit, um so viele Aufgaben wie möglich zu lösen.

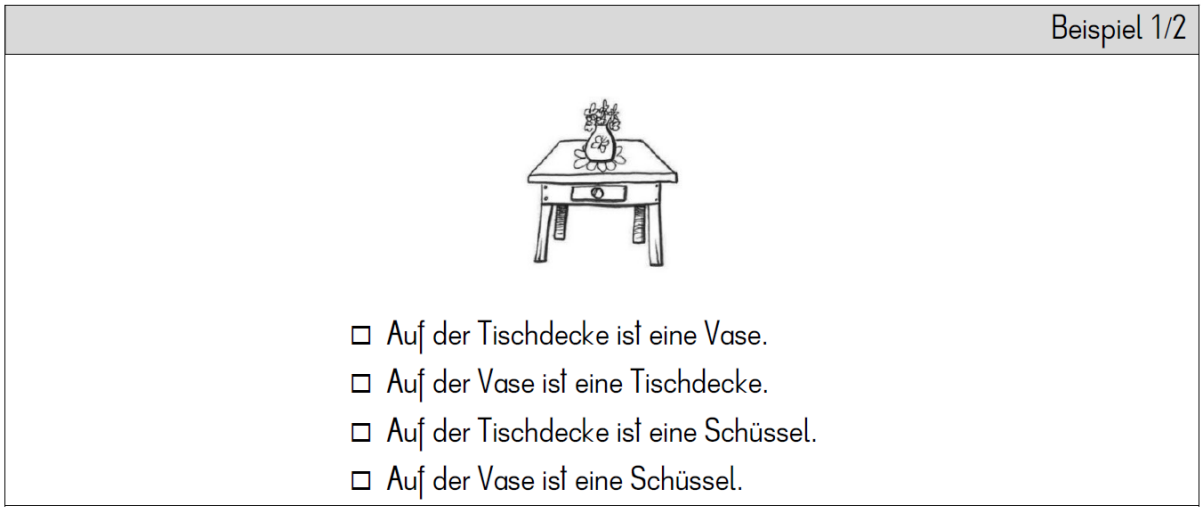

Abb. 2: Beispiel-Item des Subtests Satz im GraLeV.

Subtest Text I: Quatschgeschichten

Mit dem Subtest Text I wird geprüft, ob die Schüler*innen Informationen aus kurzen Texten entnehmen können (Beantwortung kurzer Fragen). Bei den Texten in diesem Subtest handelt es sich um acht Quatschgeschichten (über nichtexistierende Dinge, Tiere oder Handlungen). Nach jeder dieser Quatschgeschichten folgen zwei Fragen, von denen eine das direkte Abrufen von Informationen und eine weitere das Ziehen von Schlussfolgerungen erfordert. Die Schüler*innen kreuzen die richtige Antwort an (siehe Abb. 3) und haben für diesen Subtest drei Minuten Zeit.

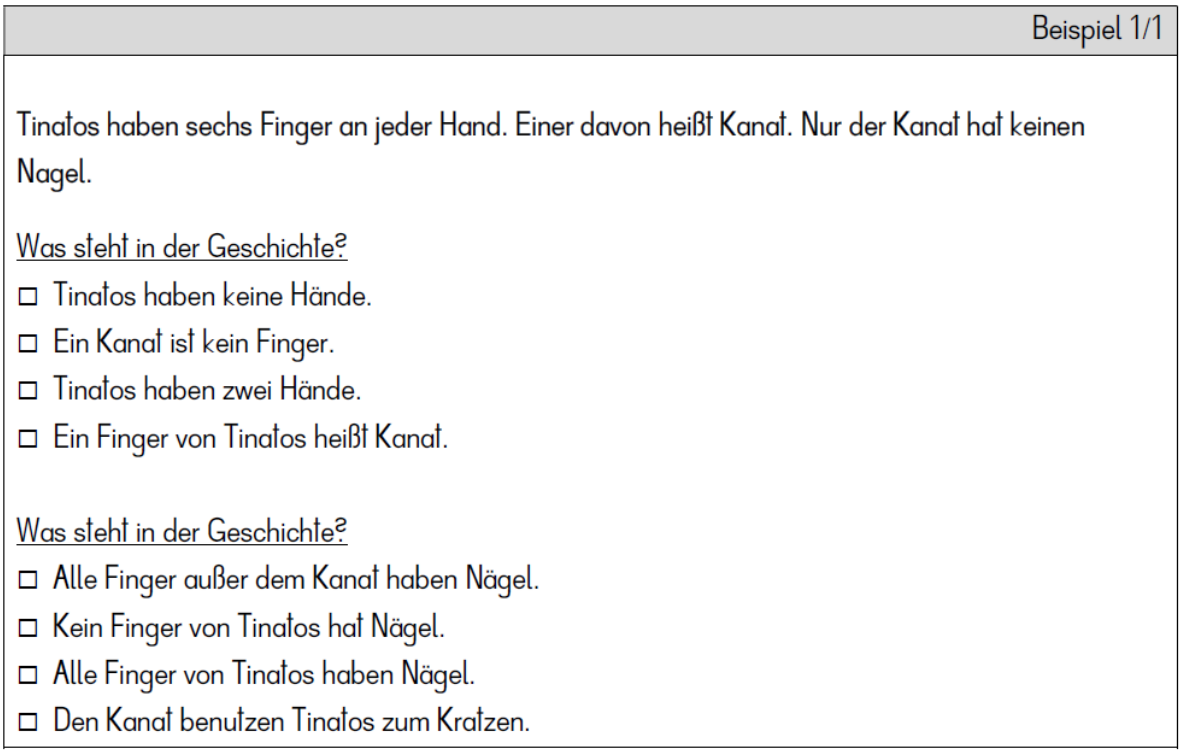

Abb. 3: Beispiel-Item des Subtests Text I im GraLeV.

Subtest Text II: Maze-Verfahren

Der Subtest Text II basiert auf dem Maze-Verfahren, bei dem während des Lesens eines Textes an der Stelle jeden siebten Wortes ein Wort aus drei Alternativen (Zielwort, sowie ein phonologischer/graphematischer und ein syntaktischer/morphologischer Distraktor) ausgewählt werden muss. Die drei Wörter werden in einer Klammer nebeneinander präsentiert und die Kinder müssen das am besten geeignete Wort einkreisen (siehe Abb. 4). Zwei Texte (je 100 Wörter) werden nacheinander dargeboten, wobei der zweite einen höheren Schwierigkeitsgrad hat als der erste. Die Kinder haben 100 Sekunden Zeit, um so viele Aufgaben zu lösen, wie sie können.

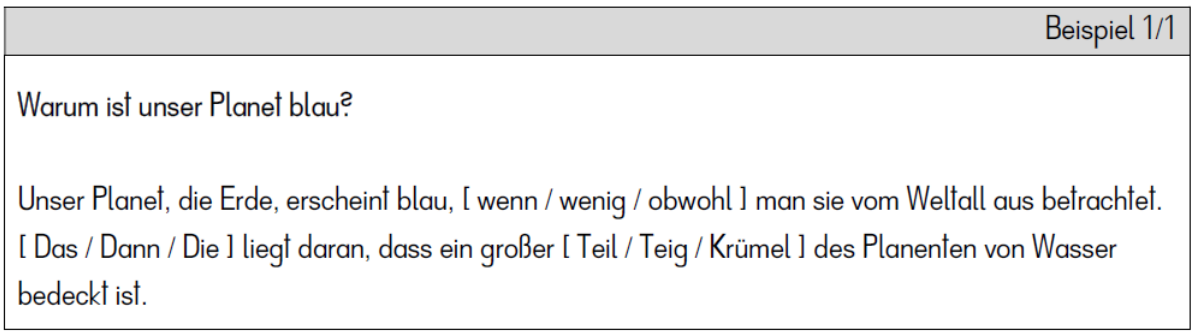

Abb. 4: Beispiel-Item des Subtests Text II im GraLeV.

1.3 Auswertungsmodus und Interpretation der Werte

1.3.1 Bestimmung der Rohwerte

Im ersten Auswertungsschritt wird je Subtest die **Anzahl richtig gelöster Items** innerhalb der vorgegebenen Zeit bestimmt. Bei diesen Werten handelt es sich um die **Rohwerte** der vier Subtests. Zu diesen gelangt man, indem man jeweils die Anzahl der Fehler und die Anzahl der Auslassungen von der Anzahl der insgesamt gelösten Items subtrahiert. Die Rohwerte werden also wie folgt berechnet:

Berechnung der Subtestrohwerte:

Gesamt Subtest Wort (maximal erreichbare Punktzahl: 36)

Anzahl richtiger Wörter = Wörter gesamt - Fehler - Auslassungen

*(dabei wichtig: jedes einzelne Item wird gezählt, **nicht** die Itemsets bestehend aus 3 Items)*

Gesamt Subtest Satz (maximal erreichbare Punktzahl: 16)

Anzahl richtiger Sätze = Sätze gesamt - Fehler - Auslassungen

Gesamt Subtest Text I (maximal erreichbare Punktzahl: 16)

Anzahl richtiger Antworten = Antworten gesamt - Fehler - Auslassungen

*(dabei wichtig: jede einzelne Antwort wird gezählt, **nicht** die Quatschgeschichten)*

Gesamt Subtest Text II (maximal erreichbare Punktzahl: 30)

Anzahl richtiger Klammern = Klammern gesamt - Fehler - Auslassungen

Die Ergebnisse werden auf dem Auswertungsbogen in die Spalte „Rohwert“ eingetragen.

1.3.2 Ablesen der Normwerte

Im nächsten Schritt werden für die vier Subtests Normwerte (Prozentränge und z-Werte) aus den Normtabellen A-1 bis A-4 abgelesen und in den dafür vorgesehenen Spalten im Auswertungsbogen eingetragen. Dabei handelt es sich um Normwerte, die hinsichtlich der Verteilung der Kinder in Bezug auf Geschlecht repräsentativ sind.

Diese Normwerte liegen für alle vier Subtests für vier verschiedene Zeitpunkte vor:

- Beginn 3. Schulstufe (erste 8 Wochen): Tabelle A-1

- Ende 3. Schulstufe (letzte 8 Wochen): Tabelle A-2
- Beginn 4. Schulstufe (erste 8 Wochen): Tabelle A-3
- Ende 4. Schulstufe (letzte 8 Wochen): Tabelle A-4

Bei einem Vergleich des Testergebnisses eines einzelnen Kindes mit den Normierungsdaten ist es wichtig, dass das Kind etwa zum selben Zeitpunkt im Schuljahr getestet wurde wie die Kinder der Normstichprobe. Wenn ein Kind nicht genau in der Zeitspanne, für die die Normen vorliegen, getestet wurde, sollte jene Normtabelle, die dem Testzeitpunkt am nächsten liegt, als Vergleichswert herangezogen werden.

Auch für Kinder, die eine Klasse wiederholen, wird empfohlen, die Normen jener Schulstufe, in der sie sich zum Zeitpunkt der Testung befinden, heranzuziehen, da Repetentinnen und Repetenten auch in der Testnormierung der jeweiligen Schulstufe zugeordnet wurden, in der sie sich zum Zeitpunkt der Testung befanden. Falls dies die Testleitung aus irgendeinem Grund nicht so handhaben möchte, könnten auch Vergleiche anhand des Alters, das in der Normtabelle für jeden Testzeitpunkt ersichtlich wird (Mittelwert und Standardabweichung), angestellt werden. Allerdings sollten jedoch bei Leistungen, die durch den Unterricht maßgeblich beeinflusst werden wie das Lesen, im Allgemeinen eher Schulstufennormen anstelle von Altersnormen zum Vergleich herangezogen werden (Marx, 2007).

1.3.3 Bestimmung des Gesamtwertes Leseverständnis

Um den Gesamtwert Leseverständnis zu erhalten, werden die z-Werte der vier Subtests addiert. Wie bereits erwähnt sind die z-Werte in den Tabellen ersichtlich, indem in der Zeile des Rohwerts des Kindes der dazugehörige z-Wert abgelesen wird. Im Ergebnis erhält man den so genannten kumulierten z-Wert, der in das entsprechende Eintragungsfeld (z_{kum}) auf dem Auswertungsbogen eingetragen wird.

Berechnung des Gesamtwertes

z_{kum} = z-Wert im Subtest Wort + z-Wert im Subtest Satz + z-Wert im Subtest Text I + z-Wert im Subtest Text II

Der Gesamtwert des Leseverständnisses wird somit nicht durch Addition der Rohwerte der Subtests gebildet, sondern ausnahmslos immer durch die Addition der z-Werte der vier Subtests. Dies ist dadurch begründet, dass die Subtests für Kinder unterschiedlich schwer sind und unterschiedlich viele Items zu lösen sind. Dadurch werden ein- und demselben Rohwert in den vier Subtests normalerweise unterschiedliche Prozentränge zugeordnet. Es würde daher zu schwer interpretierbaren Ergebnissen führen, wenn man die Rohwerte der Subtests einfach addiert. Dieses Problem wird durch Addition der z-Werte gelöst. Der kumulierte z-Wert kann nur gebildet werden, wenn die z-Werte aller Subtests vorliegen. Fehlt ein Ergebnis eines Subtests, kann kein Gesamtwert für das Leseverständnis des Kindes gebildet werden.

Mit Hilfe von Tabelle A-5 wird dem kumulierten z-Wert ein Prozentrang (PR) zugeordnet. Tabelle A-5 stellt Informationen zu den repräsentativen Normwerten in den beiden Schulstufen 3 und 4 bereit. Der PR bezieht sich auf das Leseverständnis des Kindes im Sinne einer Gesamtfähigkeit, die sich aus den Teilfähigkeiten auf Wort-, Satz- und Textebene zusammensetzt. Es fließen also Teilfähigkeiten auf allen drei Leseverständnisebenen in den Gesamtwert ein.

1.3.4 Interpretation der Ergebnisse

Mit Hilfe des GraLeV können Kinder mit unterdurchschnittlichen bzw. nicht altersgerechten Leseverständnisfähigkeiten relativ einfach identifiziert werden. Anhand der Anzahl der gelösten Items innerhalb einer vorgegebenen Zeit erkennt man, wie schnell ein Kind zu lesen vermag, aber auch, wie gut es das Gelesene versteht. Dabei können differenziert die drei Leseverständnisebenen betrachtet werden und Schwerpunkte für die Förderung in Therapie und Unterricht abgeleitet werden.

Einerseits kann ein niedriger Rohwert durch eine hohe Anzahl an Fehlern zustande kommen: einerseits fließen die nicht ausgebesserten Fehler direkt in die Subtraktion zur Erlangung des Rohwerts ein und andererseits machen sich die korrigierten Fehler durch einen Zeitverlust bemerkbar. Zudem kann eine schlechte Leistung im GraLeV auch durch ein sehr geringes Lesetempo zustande kommen.

Die Rohwerte in den vier Subtests sind aufgrund der Unterschiedlichkeit der Tests und des Lesematerials, das in der vorgegebenen Zeit bearbeitet werden kann, nicht vergleichbar. Durch die getrennte Normierung der Subtests wird dies berücksichtigt. Die relativen Leistungen in den vier Subtests (also die z-Werte bzw. Prozentränge) dienen dazu, die Fähigkeiten der vier Subtests besser miteinander vergleichen zu können. Aber auch die relativen Werte in den vier Subtests können bei einem Kind unterschiedlich sein. Wenn das Wortlesen besser ausgebildet ist als das Satz- und/oder Textlesen, deutet dies auf wenig gut ausgebildete höhere Leseverständnisfähigkeiten hin (vorausgesetzt die Testung lief in den Subtests vergleichbar ab).

Bedeutung der Prozentränge

Um einen einheitlichen Bezugsrahmen für die Interpretation zu schaffen, werden den Rohwerten im GraLeV Prozentränge (PR) zugeordnet. Die Interpretation der Ergebnisse sollte sich immer auf die Prozentränge stützen. Ein PR gibt an, wie viele Kinder aus der Vergleichsgruppe (das sind Kinder aus der Normstichprobe derselben Schulstufe zum selben Zeitpunkt der Testung: Schuljahresbeginn vs. Schuljahresende) *genauso viele oder weniger* Aufgaben korrekt bearbeitet haben.

Beispiel: Ein PR von 63 bedeutet, dass 63% der Kinder aus der herangezogenen Normstichprobe (vergleichbaren Alters) gleich gute oder schlechtere Leistungen zeigen, 37% zeigen hingegen eine bessere Leistung. Dementsprechend gilt: je höher der PR eines Kindes, desto besser waren seine Leistungen.

Ein PR von 50 weist auf eine durchschnittliche Leistung hin. Ein niedriger Wert ($PR \leq 15$) besagt hingegen, dass das Kind unterdurchschnittlich im (Sub-)Test abgeschnitten hat (deutlich weniger Items richtig bearbeitet hat im Vergleich zur Normstichprobe). Dieser Prozentrang entspricht etwa einer Standardabweichung unter dem Mittelwert. Ein solches Ergebnis deutet auf Defizite in den Leseverständnisfähigkeiten hin.

Vorsicht: Aus Prozentrangdifferenzen kann nicht auf Leistungsdifferenzen geschlossen werden (Bühner, 2011), da die Prozentrangskala Ordinalskalenniveau aufweist und keinen linearen Maßstab darstellt. Aufgrund der genannten Eigenschaften ist es auch mathematisch nicht zulässig, Prozentrangnormen zu mitteln. Lehrende sollen somit darauf verzichten, etwa den durchschnittlichen Prozentrang der Klasse im GraLeV zu ermitteln. Dies würde zu mathematisch unsinnigen Ergebnissen führen. Auch sollte darauf verzichtet werden, für ein und dasselbe Kind den PR der vier Subtests zu mitteln. Für die genannten Fälle sollten z-Werte verwendet werden.

Bedeutung der z-Werte

Für den GraLeV wurden für jeden Subtests neben PR auch z-Werte berechnet. Diese intervallskalierten Werte sind als Einheiten der Standardnormalverteilung zu interpretieren. Der Mittelwert der z-Skala liegt bei 0, die Standardabweichung bei 1. Abweichungen vom Mittelwert können anhand des Vorzeichens eindeutig einem überdurchschnittlichen oder unterdurchschnittlichen Ergebnis zugeordnet werden. Der Bereich von -1 bis +1 (also eine Standardabweichung um den Mittelwert) gilt im deutschen Sprachraum üblicherweise als Durchschnittsbereich. Niedrige Werte (z<-1) bzw. sehr niedrige Werte (z<-2) können als (weit) unterdurchschnittlich im Vergleich zur Normstichprobe interpretiert werden.

Anhand der z-Werte kann ein Vergleich zwischen den vier Subtests angestellt werden. Zudem ist es mathematisch auch zulässig, z-Werte zu addieren (wie bei der Berechnung des kumulierten z-Wertes für den Gesamtwert Leseverständnis) und Durchschnittswerte zu berechnen.

1.4 Auswertungshilfen

Für die Auswertung steht die Auswertungsschablone, ein Auswertungsbogen sowie Normtabellen (jeweils für die dritte und vierte Schulstufe und für Schuljahresbeginn und Schuljahresende) zur Verfügung.

1.5 Auswertungszeit

Die Auswertung benötigt ca. 15 Minuten pro Testheft. Eine Auswertungsschablone steht zur Verfügung.

1.6 Itembeispiele

Siehe Kapitel 1.2 Testaufbau

1.7 Items

Die Itemkennwerte (Itemschwierigkeit und Trennschärfe) finden sich nachfolgend in den Tabellen 1 bis 4 für die jeweiligen Subtests. Die Itemkennwerte wurden während der zweiten Pilotierung (Powertestung, siehe Kapitel 3.3) mit der Digitalversion1 ermittelt.

1 Aufgrund von Untersuchungen zum Vergleich zwischen Digital- und Printversion (siehe Seifert & Paleczek, 2022), gehen wir davon aus, dass die ermittelten Unterschiede zwischen den Versionen insbesondere auf den

Tabelle 1. Items des Subtests Wort.

Item-Set-Nr.	Wortart	Item-Nr.	Item	Frequenz[a]	Itemschwierigkeit	Trennschärfe
1	Nomen	1	Flasche	239	0,98	0,63
		2	Rutsche	37	0,98	0,42
		3	Tasche	422	0,97	0,56
2	Nomen Komposita	4	Hundefutter	13	0,97	0,75
		5	Hundeleine	9	0,94	0,51
		6	Hundeschnauze	2	0,96	0,58
3	Verben	7	fahren	553	0,99	0,60
		8	fallen	475	0,92	0,53
		9	falten	8	0,94	0,63
4	Verben	10	hämmern	5	0,95	0,59
		11	drehen	145	0,94	0,31
		12	hängen	234	0,95	0,53
5	Nomen	13	Spiegel	197	0,94	0,53
		14	Stiefel	51	0,97	0,45
		15	Ziege	80	0,93	0,39
6	Adjektive	16	steil	41	0,90	0,70
		17	stumpf	163	0,96	0,67
		18	still	650	0,95	0,38
7	Nomen	19	Abfall	21	0,92	0,67
		20	Abflug	5	0,95	0,41
		21	Abstand	17	0,90	0,54
8	Verben	22	regnen	54	0,98	0,27
		23	rodeln	5	0,88	0,59
		24	rasten	26	0,89	0,33
9	Verben	25	schnuppern	9	0,90	0,68
		26	schwimmen	120	0,94	0,25
		27	schnappen	68	0,91	0,54
10	Verben	28	wecken	58	0,93	0,35
		29	merken	167	0,85	0,28
		30	wedeln	7	0,87	0,64
11	Adjektive	31	müde	344	0,98	0,40
		32	mutig	95	0,83	0,43
		33	mächtig	80	0,83	0,53
12	Nomen Komposita	34	Bauchnabel	9	0,91	0,31
		35	Baumwipfel	21	0,86	0,57
		36	Baustelle	18	0,87	0,35

Faktor Zeit zurückzuführen sind. Die Itemkennwerte selbst sollten, wenn dann nur geringfügig beeinträchtigt sein. Dennoch ist eine Überprüfung der Kennwerte in der Printversion mit einer erneuten Powertestung angedacht.

Anmerkung: [a]ermittelt durch Childlex (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015), Angaben in pro Million Wörter

Tabelle 2. Items des Subtests Satz.

Item-Nr.	Ziel-Item	Grammatischer Fokus	Item-schwierigkeit	Trenn-schärfe
1	Die Frau öffnet den Gästen die Tür.	Satzstellung	0,95	0,41
2	Das Mädchen geht mit der Schildkröte, aber nicht mit dem Hund spazieren.	Negation	0,93	0,35
3	Der Schneemann wird vom Mädchen geschoben.	Satzstellung (bei Passivstrukturen)	0,92	0,40
4	Der Vogel sitzt auf einer Schaukel.	Plural	0,92	0,38
5	Das Mädchen zeigt dem Mann gerade etwas.	Satzstellung, Zeitformen	0,91	0,29
6	Der Pirat klatscht in die Hände, während die Prinzessin und der Clown tanzen.	Satzstellung bei Subordination	0,88	0,34
7	Mit der Weihnachtskugel schmückt der Junge den Weihnachtsbaum.	Satzstellung bei Topikalisierung	0,89	0,27
8	Der Junge hat einen Schneemann gebaut.	Zeitformen	0,86	0,46
9	Der Teppich ist zwischen dem Bett und dem Tisch.	Satzstellung	0,85	0,55
10	Der Spiegel ist hinuntergefallen.	Zeitformen	0,82	0,33
11	Eine Schranktür ist offen, deshalb sieht man die Kleidung.	Satzstellung bei Subordination (und Negation)	0,79	0,59
12	Die Kleiderbügel hängen an einer Kleiderstange.	Satzstellung	0,78	0,35
13	Das Kind badet in der Wanne, aber die Ente badet nicht mit.	Satzstellung bei Subordination (und Negation)	0,75	0,53
14	Die Blume ist schon umgeknickt.	Zeitformen	0,73	0,50
15	Während die Kinder mit der Achterbahn fahren, freuen sie sich.	Satzstellung bei Subordination	0,71	0,47
16	Das Wasser fließt in den Eimer.	Plural	0,53	0,39

Tabelle 3. Items des Subtests Text I.

Item-Set-Nr.	Nonsens-Wort/ Quatsch-Wort[a]	Item-Nr.	Item-Aussagen	Item-schwierigkeit	Trenn-schärfe
1	Relemis	1	Relemis haben schwarzes Fell.	0,84	0,64
		2	Relemis haben viel Fell.	0,76	0,46
2	branteln	3	Die Schule ist aus.	0,81	0,63
		4	Emre brantelt gern.	0,78	0,57

3	Rafiza	5	Nina isst gerne Rafiza.	0,88	0,60
		6	Nina freut sich, wenn es Rafiza zu essen gibt.	0,69	0,28
4	minnern	7	Wenn Adam minnert, ist das laut.	0,82	0,56
		8	Adam macht es Spaß zu minnern.	0,77	0,27
5	krolken	9	Vesta und Marin spielen heute drinnen.	0,83	0,39
		10	Bei Regen kann man nicht krolken.	0,63	0,59
6	Basati	11	Hudara rutscht aus.	0,73	0,57
		12	Nasses Basati ist rutschig.	0,76	0,66
7	Sinalas	13	Sinalas leben in Höhlen.	0,88	0,56
		14	Sinalas laufen weg, wenn man sich zu schnell bewegt.	0,60	0,42
8	Zünglis	15	Zünglis haben viele Feinde.	0,80	0,50
		16	Zünglis nehmen sich viel Zeit zum Fressen.	0,62	0,57

Anmerkung: [a] Die kurzen Texte, die aus 2 bis 4 Sätzen bestehen, wurden zu Nonsens-/ Quatschwörtern konstruiert. Für die vollständigen Texte, siehe Testheft.

Tabelle 4. Items des Subtests Text II.

Text	Überschrift	Item-Nr.	Item	Item-schwierigkeit	Trenn-schärfe
1	Was sind Nutztiere?	1	leben	0,92	0,33
		2	oder	0,83	0,28
		3	Wildtiere	0,81	0,41
		4	Diese	0,77	0,41
		5	heißen	0,71	0,45
		6	nützlich	0,80	0,38
		7	allem	0,65	0,48
		8	ist	0,89	0,49
		9	Wurst	0,85	0,47
		10	auch	0,83	0,49
		11	nicht	0,80	0,33
		12	geben	0,69	0,36
		13	jeder	0,76	0,41
		14	herstellen	0,84	0,51
		15	Nutztier	0,69	0,34
2	Wie entsteht Tomatensoße?	16	Dann	0,87	0,44
		17	Lastwagen	[a]	[a]
		18	transportiert	0,63	0,35
		19	werden	0,88	0,38
		20	zerkleinerten	0,69	0,31
		21	fertige	0,78	0,32
		22	Päckchen	0,88	0,40

		23	Kisten	0,87	0,32
		24	liefert	0,86	0,48
		25	dorthin	0,72	0,43
		26	warten	0,86	0,45
		27	die	0,87	0,36
		28	auspacken	0,74	0,43
		29	um	0,70	0,49
		30	zu	0,80	0,50

Anmerkung: Das Zielitem wurde nachträglich, nach der Erfassung der Kennwerte während der zweiten Pilotierung (siehe Kapitel 3.3) geändert, daher können für dieses Item keine Kennwerte angegeben werden.

2. Durchführung

2.1 Testformen

Hier beschrieben sind die Hinweise für die Printversion des GraLeV, der als Einzel- oder Gruppenverfahren eingesetzt werden kann. Zusätzlich existiert eine Digitalversion als App, die regelmäßige Wartung erfordert und daher nicht frei verfügbar ist (Anfragen bitte gern an susanne.seifert@uni-graz.at). 2

2.2 Altersbereiche

Der Test liegt normiert für den Schuljahresbeginn und das Schuljahresende der dritten und vierten Schulstufe vor. Die Altersangaben der Kinder der Normierungsstichprobe reichen von im Mittel 8,91 Jahren (SD = 0,48) zu Beginn der dritten Schulstufe bis durchschnittlich 10,55 Jahren (SD = 0,53) zum Ende der viertel Schulstufe. Weitere Angaben zur Normierungsstichprobe finden sich in Kapitel 4.4.

2.3 Durchführungszeit

Die Durchführungszeit beträgt im Gruppensetting mit der gesamten Klasse etwa 30 Minuten. Die reine Testzeit (ohne Instruktionen) beläuft sich auf weniger als 11 Minuten.

2.4 Material

- Testheft (19 Seiten pro Kind)
- 2 Bleistifte (oder andere Stifte, mit denen das Kind gut und leserlich schreiben kann)
- Stoppuhr (z. B. digitaler Timer am Smartphone)

2.5 Instruktion

Die Instruktion wird mündlich in folgender Form vorgegeben.

Nachdem die Testhefte ausgeteilt wurden, lassen sie die Kinder vorerst geschlossen vor sich liegen. Dann werden gemeinsam die Daten auf der Deckblatt-Seite ausgefüllt und auf die Stopp-Figur folgendermaßen hingewiesen:

„Seht ihr dieses Bild am Ende der Seite? Welches Schild hält denn dieses Mädchen in der Hand? Was könnte das bedeuten? (….) Genau: dort musst du stoppen. Du darfst nicht umblättern. Wir blättern dort dann erst gemeinsam um. Ich gebe euch dafür dann immer das Zeichen.“

Dann blättern alle gemeinsam auf die erste Seite um. Die Subtests werden immer einzeln erklärt und die Beispiele, für die unbegrenzte Zeit zur Verfügung steht, gemeinsam besprochen. Das erste Beispielitem wird jeweils gemeinsam gelöst. Es sollte sichergestellt werden, dass alle Kinder das

2 In der Publikation Seifert & Paleczek (2022) haben wir die beiden Versionen miteinander verglichen. Es zeigen sich Mittelwertsunterschiede, weshalb auch getrennte Normen für die Print- und Digitalversion vorliegen. Publiziert sind jedoch aufgrund der Unzulänglichkeit der Digitalversion nur die Normen der Printversion.

Prozedere des Markierens der Lösung (Wort: ziehen einer Verbindungslinie; Satz & Text I: ankreuzen; Text II: einkreisen) verstanden haben. Dann versuchen die Kinder, das zweite bzw. dritte Beispiel jeweils alleine zu lösen, um zu gewährleisten, dass die Kinder auch selbstständig zum Ausfüllen in der Lage sind. Es wird dann gemeinsam kontrolliert.

Durch Umhergehen im Klassenzimmer soll sichergestellt werden, dass die Kinder folgende Dinge berücksichtigen:

- Immer nur ein Kreuz (Subtest Satz, Text I) / Kreis (Subtest Text II) pro Frage setzen bzw. drei Verbindungsstriche pro Aufgabe (Subtest Wort) ziehen.
- Keine Items auslassen.
- Nicht vorausarbeiten.
- Nicht abschauen.
- Nicht herumblättern.

2.5.1 Subtest Wort

Die Instruktion zu den Beispielen des Subtests Wort, für die unbegrenzte Zeit zur Verfügung steht, sollte wie folgt lauten:

*„Jetzt machen wir ein Spiel mit Wörtern. Sieh dir dazu mal das erste Beispiel an. Du siehst immer drei Bilder und darunter stehen sechs Wörter. Zu jedem Bild suchst du das richtige Wort. Verbinde das Wort mit dem passenden Bild. Versuchen wir es mal gemeinsam. Das erste Bild ist schon mit dem passenden Wort verbunden. Wer liest mal bitte vor, welches Wort das ist? (…) Wer liest mir mal bitte die anderen Wörter vor beim ersten Beispiel? (…) Jetzt versucht jede*r für sich das richtige Wort zum zweiten Bild zu finden. Verbinde das Bild mit dem Wort. (…) Wer kann mir das passende Wort sagen? (…) Wenn du etwas falsch verbindest, dann radiere bitte nicht. Streiche einfach den Strich 2x durch.* (auf der Tafel zeigen) *Versuche jetzt mal das letzte Bild mit dem passenden Wort zu verbinden.“* (Die Kinder noch das letzte Bild im Beispiel 1 lösen lassen, gemeinsam korrigieren und sicherstellen, dass alle verstanden haben, was zu tun ist, Fehler klären).

„Jetzt versucht mal allein das zweite Beispiel zu lösen.“ (herumgehen und schauen, ob die Kinder wissen, was zu tun ist, Hilfestellungen sind hier noch erlaubt – wichtig: alle müssen verstehen, was sie zu tun haben)

Dann folgt der Übergang von den Beispielen zu den zu wertenden Aufgaben dieses Subtests. Für diese Aufgaben ist die Zeit begrenzt. Für alle Aufgaben haben die Kinder insgesamt 3 Minuten Zeit. Die Zeit muss manuell mitgestoppt werden:

„Super, jetzt haben alle verstanden, was zu tun ist. Auf den nächsten Seiten gibt es mehrere solche Aufgaben. Arbeite so schnell und genau, wie möglich. Fülle die Aufgaben von oben nach unten aus und beantworte sie allein. Verbinde so, wie du am ehesten denkst, dass es passt. Nicht vergessen: Bitte radiere nicht, sondern streich den Strich durch, wenn du deine Meinung änderst. Mach danach einfach einen neuen. Du darfst so lange arbeiten, bis ich STOPP sage. Wenn ich STOPP sage, dann legst du deinen Stift bitte hin und streckst deine Hände in die Höhe. Wenn du die STOPP-Figur siehst, bitte auch nicht

weiterblättern. Mach dir keine Sorgen, wenn du nicht fertig wirst. Fast kein Kind schafft es, fertig zu werden.

Sind alle bereit? Dann geht es jetzt los."

Nachdem die 3 Minuten vorbei sind:

„STOPP. Die Zeit ist jetzt vorbei. Lege bitte jetzt deinen Stift hin und strecke deine Hände in die Höhe. Wir gehen nun weiter zum nächsten Spiel. Bitte blättere dafür auf Seite 6." (kontrollieren, ob jedes Kind auf der Seite 6 ist)

2.5.2 Subtest Satz

Die Instruktion zu den Beispielen des Subtests Satz, für die unbegrenzte Zeit zur Verfügung steht, sollte wie folgt lauten:

„Jetzt machen wir ein Spiel mit Sätzen. Sieh dir dazu mal das erste Beispiel an. Du siehst immer ein Bild und darunter stehen vier Sätze. Aber nur ein Satz passt zum Bild. Kreuze den richtigen Satz zum Bild an. Versuchen wir es mal gemeinsam. Wer liest mir mal bitte die Sätze vor beim ersten Beispiel? (…) Versuche jetzt mal den richtigen Satz anzukreuzen. Wenn du etwas falsch angekreuzt hast, dann radiere bitte nicht. Male einfach das Kästchen schnell komplett aus." (auf der Tafel zeigen: nicht schön, sondern vor allem schnell muss es gehen, eher kritzeln – die Kinder müssen verstehen, dass es um Zeit geht) (Die Kinder Beispiel 1 lösen lassen, gemeinsam korrigieren und sicherstellen, dass alle verstanden haben, was zu tun ist, Fehler klären).

„Jetzt versucht mal allein das zweite Beispiel zu lösen." (herumgehen und schauen, ob die Kinder wissen, was zu tun ist, Hilfestellungen sind hier noch erlaubt – wichtig: alle müssen verstehen, was sie zu tun haben)

Dann folgt der Übergang von den Beispielen zu den zu wertenden Aufgaben dieses Subtests. Für diese Aufgaben ist die Zeit begrenzt. Für alle Aufgaben haben die Kinder insgesamt 3 Minuten Zeit. Die Zeit muss manuell mitgestoppt werden:

„Super, jetzt haben alle verstanden, was zu tun ist. Auf den nächsten Seiten gibt es mehrere solche Aufgaben. Arbeite so schnell und genau, wie möglich. Fülle die Aufgaben von oben nach unten aus und beantworte sie allein. Kreuze das an, was für dich am besten passt. Nicht vergessen: Bitte radiere nicht, sondern male das Kästchen ganz schnell an, wenn du deine Meinung änderst. Mach danach einfach ein neues Kreuz. Du darfst so lange arbeiten, bis ich STOPP sage. Wenn ich STOPP sage, dann legst du deinen Stift bitte hin und streckst deine Hände in die Höhe. Wenn du die STOPP-Figur siehst, bitte auch nicht weiterblättern. Mach dir keine Sorgen, wenn du nicht fertig wirst. Fast kein Kind schafft es, fertig zu werden.

Sind alle bereit? Dann geht es jetzt los."

Nachdem die 3 Minuten vorbei sind:

„STOPP. Die Zeit ist jetzt vorbei. Lege bitte jetzt deinen Stift hin und strecke deine Hände in die Höhe. Wir gehen nun weiter zum nächsten Spiel. Bitte blättere dafür auf Seite 13." (kontrollieren, ob jedes Kind auf der Seite 13 ist)

2.5.3 Subtest Text I

Die Instruktion zu den Beispielen des Subtests Text I, für die unbegrenzte Zeit zur Verfügung steht, sollte wie folgt lauten:

„Jetzt machen wir ein Spiel mit Quatschgeschichten. Du siehst immer eine kurze Geschichte zu irgendetwas, was es eigentlich nicht gibt – z.B. eine Tätigkeit, ein Tier oder ähnliches. Und darunter stehen zwei Fragen mit je vier Antwortmöglichkeiten. Aber zu jeder Frage passt immer nur eine Antwort. Kreuze die richtige Antwort zur Frage an. Versuchen wir es mal gemeinsam. Wer liest mir bitte die erste Quatschgeschichte und die erste Frage vor beim Beispiel? (…) Wer liest mir mal bitte die vier Antworten vor? (…) Versucht jetzt mal bei der ersten Frage den richtigen Satz anzukreuzen. Wenn du etwas falsch angekreuzt hast, dann radiere bitte nicht. Male einfach das Kästchen komplett und schnell aus." (auf der Tafel zeigen: nicht schön, sondern vor allem schnell muss es gehen, eher kritzeln – die Kinder müssen verstehen, dass es um Zeit geht) (Die Kinder Frage 1 lösen lassen, gemeinsam korrigieren und sicherstellen, dass alle verstanden haben, was zu tun ist, Fehler klären).

„Jetzt versucht mal allein die zweite Frage zu der Geschichte zu lösen." (herumgehen und schauen, ob die Kinder wissen, was zu tun ist, Hilfestellungen sind hier noch erlaubt – wichtig: alle müssen verstehen, was sie zu tun haben)

Dann folgt der Übergang von den Beispielen zu den zu wertenden Aufgaben dieses Subtests. Für diese Aufgaben ist die Zeit begrenzt. Für alle Aufgaben haben die Kinder insgesamt 3 Minuten Zeit. Die Zeit muss manuell mitgestoppt werden:

„Super, jetzt haben alle verstanden, was zu tun ist. Auf den nächsten Seiten gibt es mehrere solche Geschichten und immer zwei Fragen dazu. Arbeite so schnell und genau, wie möglich. Fülle die Aufgaben von oben nach unten aus und beantworte sie allein. Kreuze das an, was für dich am besten passt. Nicht vergessen: Bitte radiere nicht, sondern male das Kästchen an, wenn du deine Meinung änderst. Mach danach einfach ein neues Kreuz. Du darfst so lange arbeiten, bis ich STOPP sage. Wenn ich STOPP sage, dann legst du deinen Stift bitte hin und streckst deine Hände in die Höhe. Wenn du die STOPP-Figur siehst, bitte auch nicht weiterblättern. Mach dir keine Sorgen, wenn du nicht fertig wirst. Fast kein Kind schafft es, fertig zu werden.

Sind alle bereit? Dann geht es jetzt los."

Nachdem die 3 Minuten vorbei sind:

„STOPP. Die Zeit ist jetzt vorbei. Lege bitte jetzt deinen Stift hin und strecke deine Hände in die Höhe. Wir gehen nun weiter zum nächsten und letzten Spiel. Bitte blättere dafür auf Seite 18." (kontrollieren, ob jedes Kind auf der Seite 18 ist)

2.5.4 Subtest Text II

Die Instruktion zu den Beispielen des Subtests Text II, für die unbegrenzte Zeit zur Verfügung steht, sollte wie folgt lauten:

„Jetzt machen wir ein Spiel mit Lückentexten. Du siehst einen kurzen Text und darin sind einige Wörter ersetzt durch eine Klammer. In dieser Klammer stehen immer drei Wörter. Aber es passt immer nur ein Wort aus der Klammer in den Text. Kreise das richtige Wort ein. Versuchen wir es mal gemeinsam. Wer liest mir bitte die Überschrift und den ersten Satz vor beim Beispiel, mit allen Wörtern in der Klammer? (…) Versucht jetzt mal beim ersten Satz das richtige Wort einzukreisen. Wenn du etwas falsch eingekreist hast, dann radiere bitte nicht. Streiche einfach den Kreis 2x durch." (auf der Tafel zeigen) (Die Kinder den ersten Satz lösen lassen, gemeinsam korrigieren und sicherstellen, dass alle verstanden haben, was zu tun ist, Fehler klären).

„Jetzt versucht mal allein den zweiten Satz zu lösen." (herumgehen und schauen, ob die Kinder wissen, was zu tun ist, Hilfestellungen sind hier noch erlaubt – wichtig: alle müssen verstehen, was sie zu tun haben)

Dann folgt der Übergang von den Beispielen zu den zu wertenden Aufgaben dieses Subtests. Für diese Aufgaben ist die Zeit begrenzt. Für alle Aufgaben haben die Kinder insgesamt 100 Sekunden Zeit. Die Zeit muss manuell mitgestoppt werden:

„Super, jetzt haben alle verstanden, was zu tun ist. Auf der nächsten Seite gibt es zwei solche Lückentexte mit vielen solchen Klammern. Arbeite so schnell und genau, wie möglich. Fülle die Aufgaben von oben nach unten aus und beantworte sie allein. Kreise in den Klammern immer das Wort ein, das für dich am besten in den Satz passt. Nicht vergessen: Bitte radiere nicht, sondern streiche den Kreis 2x durch, wenn du deine Meinung änderst. Mach danach einfach einen neuen Kreis. Du darfst so lange arbeiten, bis ich STOPP sage. Wenn ich STOPP sage, dann legst du deinen Stift bitte hin und streckst deine Hände in die Höhe. Wenn du die STOPP-Figur siehst, bitte auch nicht weiterblättern. Mach dir keine Sorgen, wenn du nicht fertig wirst. Fast kein Kind schafft es, fertig zu werden.

Sind alle bereit? Dann geht es jetzt los."

Nachdem die 100 Sekunden vorbei sind:

„STOPP. Die Zeit ist jetzt vorbei. Lege bitte jetzt deinen Stift hin und strecke deine Hände in die Höhe. Wir sind jetzt fertig mit unserem Spiel und ich sammle die Lesehefte wieder ein."

2.6 Durchführungsvoraussetzungen

In der Gestaltung des Testsettings sollte der Testleiter bzw. die Testleiterin (also u.U. der Lehrer bzw. die Lehrerin) bemüht sein, den Lärmpegel in der Klasse möglichst gering zu halten. Um Ablenkungen zu vermeiden, ist es ratsam, dass die testende Person die Kinder auffordert, lediglich zwei Stifte bereit zu halten und alles andere vom Tisch zu entfernen. Es kann hinzugefügt werden, dass der zweite Stift im Bedarfsfall eingesetzt werden kann, wenn der andere Stift nicht mehr funktioniert. Korrekturen können durch Durchstreichen vorgenommen werden (so dass eine Benutzung des Radiergummis oder

Tintenkillers unnötig wird). Es empfiehlt sich zudem der Einsatz von aufstellbaren Sichtschutzwänden oder Schultaschen zwischen den Kindern, um ein Abschauen zu vermeiden.

3. Testkonstruktion

Im Sommer 2020 wurden die Subtests konstruiert, so dass entsprechende Testvorformen entstanden, die anschließend in der Pilotierung überprüft wurden.

3.1 Testvorformen

3.1.1 Subtest Wort

Der Subtest Wort des GraLeV erfasst das Leseverständnis auf Wortebene. Beim Itemformat handelte es sich um Zuordnungsaufgaben, die in Form von Verbindungslinien gelöst werden. Es wurden Itemsets konstruiert, in denen drei Items bearbeitet wurden. Die Itemsets bestanden aus sechs Wörtern und drei Bildern. Zu jedem Bild gab es ein passendes Wort, das mit dem jeweiligen Bild verbunden werden sollte. Bei den restlichen drei Wörtern handelte es sich um Ablenker. Diese wurden so gewählt, dass sie dieselbe Silbenanzahl wie die gesuchten Wörter hatten und sich durch semantische (z. B. Hundefutter/Hundehütte) oder phonologische Ähnlichkeit mit dem Zielwort (z. B. Knopf/Kopf) auszeichneten. Die Wörter waren ein- bis viersilbig, wobei die sechs Wörter eines Itemsets stets dieselbe Silbenanzahl aufwiesen. Die verwendeten Wörter sollten mittel- bis hochfrequent in Kinderliteratur vertreten sein (Verwendung des childlex-Korpus unter http://alpha.dlexdb.de/query/childlex/childlex2/typ/filter/, Altersgruppe 9-10, Typefrequenz absolut bei mindestens 1 pro Mio. Wörter; Schroeder et al., 2015).

Für die Erstellung des Bildmaterials wurden vorrangig Bilder verwendet, die bereits im Projekt DiLu (Differenzierter Leseunterricht, https://differenzierter-leseunterricht.uni-graz.at/de/) generiert wurden oder sie wurden von der Zeichnerin Heike Skringer zu diesem Zwecke angefertigt. Die Bilder sind prototypische Schwarz-Weiß-Zeichnungen, was sowohl die Motivation als auch Konzentration der Kinder steigern soll. Die Vorteile von schwarz-weiß-Zeichnung bei Verwendung in diagnostischen Verfahren liegen in ihrer Klarheit und der Vermeidung einer Mischung der grafischen Elemente Kontur, Textur und Farbe (Glück, 2007). Auf Kontextinformationen, die mitunter vom Zielitem ablenken können, konnte nicht gänzlich verzichtet werden, da die Abbildbarkeit einiger Items (z.B. für das Wort Abflug) einen Kontext erfordert. Bei einigen Bildern wurde ein Pfeil hinzugefügt, um die Aufmerksamkeit der Kinder auf den relevanten Bildteil zu lenken.

Insgesamt gab es für den Subtest Wort in der Testvorform zwei Beispielaufgaben und 38 Itemsets. Letztere konnten aufgeteilt werden in 16 Itemsets mit Nomen (Tab. 5), 14 mit Verben (Tab. 6) und 8 mit Adjektiven (Tab. 7). Bei 3 der 16 Nomen-Itemsets handelte es sich bei den präsentierten Wörtern um Komposita.

Tabelle 5. Zielitems und Ablenker der Itemsets Nomen in der Testvorform des Subtest Wort, Angaben in Klammern bezeichnen die Frequenz (ermittelt mit childlex-Korpus, Schroeder et al., 2015) pro Million Wörter.

Silben	Zielitem 1	Ablenker 1	Zielitem 2	Ablenker 2	Zielitem 3	Ablenker 3
1	Knopf (132)	Kopf (4001)	Topf (126)	Top (16)	Zopf (10)	Zoff (3)
2	Handtuch (58)	Handbuch (3)	Handschuh (43)	Handstand (12)	Handschrift (17)	Handgriff (2)
2	Abflug (5)	Abfluss (1)	Abfall (21)	Abfahrt (17)	Abstand (17)	Abspann (2)
1	Schloss (242)	Schoß (138)	Floß (13)	Fluss (176)	Fuß (429)	Schluss (269)
2	Ziege (80)	Ziegel (4)	Spiegel (197)	Siegel (8)	Stiefel (51)	Stiege (9)
1	Ohr (424)	Rohr (29)	Chor (78)	Chlor (3)	Ort (294)	Tor (404)
2	Lücke (24)	Mücke (17)	Brücke (129)	Krücke (1)	Stücke (113)	Spucke (39)
2	Flagge (32)	Flamme (46)	Falle (128)	Falte (19)	Farbe (188)	Faden (52)
1	See (338)	Schnee (354)	Reh (15)	Zeh (34)	Tee (179)	Fee (219)
3	Sonnen-schirm (10)	Sonnen-creme (3)	Sonnen-schein (63)	Sonnen-schutz (1)	Sonnen-hut (2)	Sonnen-uhr (10)
1	Hahn (59)	Hand (2352)	Hund (1037)	Mund (1028)	Mond (317)	Mohn (3)
4	Hunde-futter (13)	Hunde-hütte (6)	Hunde-leine (9)	Hunde-schlitten (2)	Hunde-schnauze (2)	Hunde-schule (2)
4	Bade-anzug (10)	Bade-hose (14)	Bade-wanne (40)	Bade-wasser (8)	Bade-mantel (14)	Bade-meister (4)
2	Rutsche (37)	Kutsche (79)	Flasche (239)	Lasche (6)	Tasche (422)	Masche (2)
1	Blitz (119)	Spitz (2)	Kitz (1)	Witz (104)	Sitz (82)	Schlitz (17)
3	Baumwipfel (21)	Baumwolle (3)	Baustelle (18)	Baumeister (2)	Bauchnabel (9)	Bauchredner (42)

Tabelle 6. Zielitems und Ablenker der Itemsets Verben in der Testvorform des Subtests Wort, Angaben in Klammern bezeichnen die Frequenz (erfasst mit childlex-Korpus, Schroeder et al., 2015) pro Million Wörter.

Silben	Zielitem 1	Ablenker 1	Zielitem 2	Ablenker 2	Zielitem 3	Ablenker 3
2	fahren (553)	fahnden (1)	fallen (475)	fällen (14)	falten (8)	fangen (197)
2	spiegeln (12)	spielen (583)	springen (130)	singen (183)	spritzen (6)	sprießen (6)
2	rudern (22)	rubbeln (2)	ruhen (14)	rühren (50)	rutschen (22)	rupfen (9)
2	schreiben (362)	schreien (113)	schneiden (31)	schreiten (3)	schniefen (2)	schleifen (2)
2	ausschalten (5)	aushalten (49)	aufspannen (31)	ausspannen (3)	aufbauen (17)	aufsaugen (2)
2	hämmern (5)	dämmern (8)	drehen (145)	dehnen (3)	drängen (16)	hängen (234)
2	gucken (171)	ducken (8)	pflücken (23)	drücken (74)	drucken (10)	spucken (26)
2	schneien (19)	scheinen (61)	schießen (79)	schielen (5)	schieben (58)	scheiden (14)
2	platzen (40)	platschen (3)	quatschen (10)	quetschen (5)	blasen (11)	basteln (33)
2	fliegen (313)	fliehen (58)	lieben (65)	liegen (313)	flitzen (15)	flicken (6)
2	schnuppern (9)	schnattern (9)	schnappen (68)	schwappen (5)	schwimmen (120)	schimmern (13)
2	wecken (58)	wechseln (56)	merken (167)	meckern (18)	melken (8)	welken (1)
2	regnen (54)	regeln (13)	rodeln (5)	radeln (7)	rasten (26)	rosten (1)
2	sägen (9)	sagen (1949)	segeln (19)	sengen (2)	sehen (2656)	sehnen (5)

Tabelle 7. Zielitems und Ablenker der Itemsets Adjektive in der Testvorform des Subtest Wort, Angaben in Klammern bezeichnen die Frequenz (erfasst mit childlex-Korpus, Schroeder et al., 2015) pro Million Wörter.

Silben	Zielitem 1	Ablenker 1	Zielitem 2	Ablenker 2	Zielitem 3	Ablenker 3
1	rund (162)	bunt (55)	rot (379)	roh (8)	wund (4)	wüst (7)
2	borstig (1)	buschig (1)	duftend (3)	dösend (1)	düster (87)	durstig (12)
2	eisig (15)	eilig (203)	edel (6)	eben (1020)	eckig (4)	eifrig (118)
1	steil (41)	steif (57)	stumm (163)	stramm (11)	stumpf (163)	stur (29)
1	groß (619)	grob (39)	grün (29)	kühn (6)	krumm (27)	kross (2)
2	müde (344)	mürbe (2)	mutig (95)	muffig (12)	mächtig (80)	mäßig (2)
3	ungesund (15)	umgehend (9)	ungestüm (6)	unbequem (10)	unglücklich (66)	ungültig (1)
2	witzig (61)	winzig (72)	windig (8)	würzig (2)	wichtig (238)	willig (4)

3.1.2 Subtest Satz

Der Subtest Satz erfasst das Leseverständnis auf Satzebene in Form von Mehrfach-Wahlaufgaben mit jeweils einer richtigen Antwortmöglichkeit (Single-Choice-Aufgabe). Den Schüler*innen wurden pro Item ein Bild und vier Sätze präsentiert, wovon einer das Bild passend beschreibt und die korrekte Lösung darstellt. Die drei übrigen Sätze sind Ablenker, wobei einer davon ein lexikalischer, einer ein grammatischer und einer ein Ablenker ist, der eine Kombination aus den beiden anderen Ablenkern darstellt. Die einzelnen Ablenker-Kategorien können der Tabelle 8 entnommen werden.

Tabelle 8. Zielsätze und Ablenker in der Testvorform des Subtest Satz.

Zielsatz	Grammatischer Ablenker (Angabe des grammatischen Fokus)	Semantischer Ablenker	Kombination beider Ablenker
Auf der Tischdecke ist eine Vase.	**Auf der Vase** ist eine Tischdecke. (Subjekt-Objekt-Vertauschung)	Auf der Tischdecke ist eine **Schüssel**.	Auf der Vase ist eine Schüssel.

Der Wanderer marschiert mit seinem Stock in den Wald.	Der Wanderer marschiert mit **seinen Stöcken** in den Wald. (statt Singular Plural)	Der Wanderer **kriecht** mit seinem Stock in den Wald.	Der Wanderer kriecht mit seinen Stöcken in den Wald.
Der Spiegel ist hinuntergefallen.	Der Spiegel **wird hinunterfallen**. (statt Präsens Futur)	Der Spiegel ist **hinuntergesprungen**.	Der Spiegel wird hinunterspringen.
Während die Kinder mit der Achterbahn fahren, freuen sie sich.	**Nachdem** die Kinder mit der Achterbahn gefahren sind, freuen sie sich. (statt Gleichzeitigkeit Nachzeitigkeit)	Während die Kinder mit dem **Lift** fahren, freuen sie sich.	Nachdem die Kinder mit dem Lift gefahren sind, freuen sie sich.
Der Kleiderbügel hängt an einer Kleiderstange.	**Die Kleiderstange** hängt an einem Kleiderbügel. (Subjekt-Objekt-Vertauschung)	Die **Schuhe** hängen an einer Kleiderstange.	Die Schuhe hängen an einem Kleiderbügel.
Das Mädchen baut eine Sandburg, während der Junge mit dem Bagger spielt.	Das Mädchen baut eine Sandburg, **nachdem** der Junge mit dem Bagger gespielt hat. (statt Gleichzeitigkeit Nachzeitigkeit)	Das Mädchen baut eine Sandburg, während der **Hund** mit dem Bagger spielt.	Das Mädchen baut eine Sandburg, nachdem der Hund mit dem Bagger gespielt hat.
Das Mädchen flüstert dem Jungen etwas ins Ohr.	Der Junge flüstert dem Mädchen etwas ins Ohr. (Subjekt-Objekt-Vertauschung)	Das Mädchen **legt** dem Jungen etwas ins Ohr.	Der Junge legt dem Mädchen etwas ins Ohr.
Der Junge hat einen Schneemann gebaut.	Der Junge **baut gerade** einen Schneemann. (statt Perfekt Präsens)	Der Junge hat eine **Sandburg** gebaut.	Der Junge baut gerade eine Sandburg.
Der Junge trägt Kopfhörer, um den Lärm nicht zu hören.	Der Junge trägt Kopfhörer, um den Lärm **zu hören**. (Negation)	Der Junge trägt ein **Kopftuch**, um den Lärm nicht zu hören.	Der Junge trägt ein Kopftuch, um den Lärm zu hören.
Die Blume ist schon umgeknickt.	Die Blume **wird erst** umknicken. (statt Präsens Futur)	Die **Vase** ist schon umgeknickt.	Die Vase wird erst umknicken.
Das Kind badet in der Wanne, aber die Ente badet nicht mit.	Die Ente badet in der Wanne, aber das Kind badet **nicht** mit. (Subjekte in Teilsätzen vertauscht, Negation)	Das Kind **duscht** in der Wanne, aber die Ente duscht nicht mit.	Die Ente duscht in der Wanne, aber das Kind duscht nicht mit.
Das Mädchen zeigt dem Mann gerade etwas.	**Der Mann** zeigt dem Mädchen gerade etwas.	Das Mädchen **kauft** dem Mann gerade etwas.	Der Mann kauft dem Mädchen gerade etwas.

	(Subjekt-Objekt-Vertauschung)		
Der Schneemann wird vom Mädchen geschoben.	**Der Schneemann** schiebt das Mädchen. (Subjekt-Objekt-Vertauschung, Aktiv statt Passiv)	Der **Schneelöwe** wird vom Mädchen geschoben.	Der Schneelöwe schiebt das Mädchen.
Mit der Weihnachtskugel schmückt der Junge den Weihnachtsbaum.	**Mit dem Weihnachtsbaum** schmückt der Junge die Weihnachtskugel. (Objektvertauschung im Zusammenhang mit Topikalisierung)	Mit der Weihnachtskugel schmückt der Junge den **Weihnachtsmann**.	Mit dem Weihnachtsbaum schmückt der Junge den Weihnachtsmann.
Eine Schranktür ist offen, deshalb sieht man die Kleidung.	Eine Schranktür ist offen, deshalb sieht man die Kleidung **nicht**. (Negation)	Eine **Lade** ist offen, deshalb sieht man die Kleidung.	Eine Lade ist offen, deshalb sieht man die Kleidung nicht.
Der Teppich ist zwischen dem Bett und dem Tisch.	**Der Tisch** ist zwischen dem Bett und dem Teppich. (Subjekt-Objekt-Vertauschung)	Der Teppich ist **auf** dem Bett und dem Tisch.	Der Tisch ist auf dem Bett und dem Teppich.
Das Wasser fließt in den Eimer.	Das Wasser fließt in **die Eimer**. (statt Singular Plural)	Das Wasser **plumpst** in den Eimer.	Das Wasser plumpst in die Eimer.
Der Mann trifft die Krone mit dem Schneeball.	Der Mann trifft **den Schneeball** mit der Krone. (Objektvertauschung)	Der Mann trifft die Krone mit dem **Fußball**.	Der Mann trifft den Fußball mit der Krone.
Der Vogel sitzt auf einer Schaukel.	**Die Vögel** sitzen auf einer Schaukel. (statt Singular Plural)	Der Vogel sitzt auf einer **Treppe**.	Die Vögel sitzen auf einer Treppe.
Die Kuh und die Ziege schauen den Bauern an.	Die Kuh und die Ziege **schaut** der Bauer an. (Subjekt-Objekt-Vertauschung bei Topikalisierung)	Die Kuh und die Ziege schauen den **Esel** an.	Die Kuh und die Ziege schaut der Esel an.
Der Pirat klatscht in die Hände, während die Prinzessin und der Clown tanzen.	Der Pirat klatscht in die Hände, **nachdem** die Prinzessin und der Clown fertig getanzt haben. (statt Gleichzeitigkeit Nachzeitigkeit)	Der Pirat **klatscht sich auf die Knie**, während die Prinzessin und der Clown tanzen.	Der Pirat klatscht sich auf die Knie, nachdem die Prinzessin und der Clown fertig getanzt haben.
Das Mädchen geht mit der Schildkröte, aber nicht mit dem Hund spazieren.	Das Mädchen geht **nicht** mit der Schildkröte, aber mit dem Hund spazieren.	Das Mädchen geht mit der **Katze**, aber nicht mit dem Hund spazieren.	Das Mädchen geht nicht mit der Katze, aber mit dem Hund spazieren.

	(Negation in komplexer Satzkonstruktion)		
Die Katze läuft unter dem Tisch.	Die Katze läuft unter **den** Tisch. (statt Dativ Akkusativ)	Die Katze läuft unter dem **Bett**.	Die Katze läuft unter das Bett.
Die Frau öffnet den Gästen die Tür.	**Die Gäste** öffnen der Frau die Tür. (Subjekt-Objekt-Vertauschung)	Die Frau öffnet den Gästen das **Fenster**.	Die Gäste öffnen der Frau das Fenster.

3.1.3 Subtest Text I

Der Subtest Text I dient dazu, herauszufinden, ob Schüler*innen in der Lage sind, Informationen aus kurzen Texten zu entnehmen. Der Test besteht aus acht kurzen Geschichten, zu denen jeweils zwei Multiple-Choice-Fragen beantwortet werden müssen. Die Texte enthalten Informationen zu Dingen, Lebewesen oder Handlungen, die es nicht gibt, weshalb diese Texte auch als „Quatschgeschichten" bezeichnet werden. Wir haben Nonsens-Inhalte verwendet, um zu verhindern, dass die Schüler*innen bei der Beantwortung der Fragen auf Hintergrundwissen zurückgreifen können.

Für die Testkonstruktion wurden zunächst 16 Quatschgeschichten geschrieben. Die Texte waren entweder als leicht lesbare Texte mit zwei bis drei Sätzen (n = 9) oder als anspruchsvollere Texte (z. B. mit mehr Nebensätzen, längeren Sätzen) mit vier bis fünf Sätzen (n = 7) aufgebaut.

Für die Analyse der Lesbarkeit der Texte für Schüler*innen der Schulstufen 3 und 4 wurde die Lesbarkeitsformel des Regensburger Index (RIX: Wild & Pissarek, 2019) verwendet. Diese berücksichtigt charakteristische Werte für die Lesbarkeit (z. B. mehrsilbige Wörter, Anzahl der Sätze) und bezieht Schwierigkeitsparameter (z. B. Passivformen, Satzkomplexität) mit ein. Diese Lesbarkeitsformel wurde für deutsche Texte getestet und gibt Auskunft über die Eignung für bestimmte Schulstufen. Zur Bestimmung von Textmerkmalen (siehe Tabelle 9) haben wir das Regensburger Analysetool für Texte (Ratte: Wild & Pissarek, o.J.) verwendet.

Tabelle 9. Merkmale und Lesbarkeitsindex (RIX) der Texte in der Testvorform des Subtests Text I.

Text Typ	Nonsens-Wort/ Quatsch-Wort	Word-Typ	Anzahl Wörter	Anzahl Sätze	RIX
kurz	Tinatos Kanat	Lebewesen Ding	17	3	2,3
	krolken	Handlung	20	2	2,93
	Stasmir	Ding	20	2	2,94

	minnern	Handlung	16	3	2,09
	Delliwam	Ding	26	3	2,74
	Relemis	Lebewesen	14	2	2,44
	Rafiza	Ding	12	2	2,22
	Basati	Ding	18	3	2,22
	branteln	Handlung	13	3	1,86
M (SD)			**17,33 (4,33)**	**2,56 (0,53)**	**2,40 (0,39)**
lang	Sinalas	Lebewesen	41	4	3,98
	Fenati	Ding	41	3	4,47
	Zünglis	Lebewesen	44	4	4,1
	Makentas	Ding	38	5	3,54
	Wanila	Lebewesen	42	4	4,02
	Tentaris	Lebewesen	52	5	4,41
	frijaben Frijabis	Handlung Ding	51	4	4,34
M (SD)			**44,14 (5,34)**	**4,14 (0,69)**	**4,12 (0,32)**

Die Multiple-Choice-Fragen zu den Texten deckten die beiden wichtigen Verstehensprozesse ab, die in internationalen Large-Scale-Studien (z. B. Ende der 4. Schulstufe in PIRLS: Widauer & Wallner-Paschon, 2017) und in etablierten Lesetests (z. B. ELFE II: Lenhard et al., 2020) überprüft werden. Der erste Prozess erfordert das Abrufen von explizit angegebenen Informationen. Der zweite Prozess erfordert das Ziehen einfacher Schlüsse. So ist es möglich, mit nur zwei Fragen zu jedem Text das Grundniveau des Textverständnisses eines Kindes zu erfassen.

Für die zwei angebotenen Fragen werden neben der richtigen Antwort (Ziel) drei Ablenker geboten. Die Ablenker wurden so konstruiert, dass sie zumindest theoretisch möglich waren und die Schüler*innen dazu anregten, sich auf das Lesen und Verstehen des Textes zu konzentrieren. Um die Schüler*innen mit der Textsorte (Quatschgeschichten) und mit dem Aufgabenformat (zwei Fragen zu jedem Text) vertraut zu machen, wurde ein Beispieltext mit den zwei entsprechenden Fragen konstruiert. Die Lehrkräfte lösen die Beispielaufgaben gemeinsam mit ihren Schüler*innen, um zu überprüfen, ob diese die Anweisungen verstanden haben.

3.1.4 Subtest Text II

Für den Subtest Text II wurde das Maze-Verfahren verwendet, also eine Aufgabe, bei der Schüler*innen beim Lesen eines Textes an der Stelle jedes siebten Wortes eine Auswahl aus drei Wörtern treffen müssen, so dass der Text Sinn ergibt. Der Test besteht dabei aus zwei Texten (einem etwas leichteren zu Beginn und einem etwas schwierigeren an zweiter Stelle) mit je ca. 100 Wörtern, die nacheinander präsentiert werden.

Als Basis dienten vier verschiedene Sachtexte aus differenzierten Materialien zur Förderung der Lesekompetenz von Schüler*innen der Schulstufe 3 (entwickelt vom Arbeitsbereich Inklusive Bildung und Heilpädagogische Psychologie der Universität Graz von 2012 bis 2014, kostenfrei downloadbar

unter https://differenzierter-leseunterricht.uni-graz.at/). Die Themen wurden dabei bewusst so gewählt, dass sie altersspezifische Interessen abdecken und geschlechtsneutral sind. Von den vier zur Verfügung stehenden Niveaustufen wurde jenes für durchschnittliche Leser*innen gewählt.

Nach der Auswahl der vier Sachtexte wurden die Textlänge und das voraussichtliche Zeitlimit für den Test festgelegt. Da der GraLeV auch als digitale Version entwickelt wurde, war die Vorgabe, dass ein gesamter Text auf einem Tabletbildschirm in einer angemessenen Schriftgröße ohne Scrollen lesbar sein soll, wesentlich. Dadurch war die Textlänge auf etwa 100 Wörter festgelegt. Da in der finalen Testversion ein leichter und ein schwieriger Text miteinander kombiniert werden sollten, wurden die vier ausgewählten Texte leicht umgeschrieben und gekürzt, so dass zwei relativ leichte Texte und zwei anspruchsvollere Texte zur Verfügung standen. Zur Bestimmung des Lesbarkeitsniveaus wurden zwei verschiedene Formeln verwendet. Eine davon war der gSmog (Simple measure of Gobbledygook - Deutsch; Bamberger & Vanecek, 1984), der die Anzahl der mehrsilbigen Wörter (mehr als drei Silben) im Verhältnis zur Anzahl der Sätze misst. Die andere Formel war der RIX (Wild & Pissarek, 2019). Beide Lesbarkeitsformeln wurden für die deutsche Sprache getestet und ermöglichten die Auswahl eines geeigneten Textes für die betreffende Schulstufe. Zur Ermittlung der Texteigenschaften wurde das Tool Ratte (Wild & Pissarek, o.J.) verwendet. Auf der Grundlage dieser Indizes wurden dann zwei leichtere (Texte 1 und 2) und zwei anspruchsvollere Texte (Texte 3 und 4) entworfen (siehe Tabelle 10). Schließlich fügten wir Fragen hinzu, die als Überschriften das Thema des Textes angeben. Die Verwendung solcher Überschriften sollte das Interesse der Schüler*innen wecken und sie zum Lesen des jeweiligen Textes motivieren.

Tabelle 10. Informationen über die vier Texte der Testvorform des Subtests Text II (inklusive Textmerkmale und Lesbarkeitsindices gSmog und RIX).

	Titel	Anzahl Wörter (ohne Ablenker)	Anzahl Sätze	gSmog	RIX
1	Was können wir in der Natur entdecken?	104	12	3,48	4,72
2	Was sind Nutztiere?	107	15	3,83	3,87
3	Wie entsteht Tomatensoße?	106	13	5,29	5,29
4	Wo bekommen wir unser Essen her?	104	15	6,37	4,45
M (SD)		**105,25 (1,5)**	**13,75 (1,5)**	**4,74 (1,34)**	**4,82 (0,43)**

Wie in Maze-Verfahren üblich (Fuchs & Fuchs, 1992) wurde jedes siebte Wort (ohne die Überschrift) durch eine Auswahl-Option ersetzt, die das Ziel-Item (passend zum Text) und zwei Ablenker enthielt. Jeder Text enthielt 15 Auswahl-Optionen. Wie für dieses Aufgabenformat typisch (siehe Walter, 2013), verwendeten wir einen Ablenker, der dem Zielwort graphemisch-phonemisch ähnelte, und einen Ablenker, der dem Zielwort semantisch-syntaktisch ähnelte. Wir stellten sicher, dass die Ablenker (zumindest bei den Wortarten Substantiv, Verb und Adjektiv) (a) grammatikalisch korrekt, (b) syntaktisch möglich und (c) kontextbezogen waren (Ketterlin-Geller et al., 2006). Dies diente dazu, den Schwierigkeitsgrad der Ablenker zu erhöhen. Es hat sich gezeigt, dass eine höhere Schwierigkeit der

Ablenker mit einer höheren Konstruktvalidität bei der Beurteilung des Leseverständnisses einhergeht (Conoyer et al., 2017).

3.2 Erste Pilotierung der Subtests

Im Oktober 2019 wurden alle Items der Vorformen der vier Subtests sechs Kindern einzeln angeboten (Schulstufe 3: n = 3, 2 Jungen, 1 durchschnittlicher und 1 schwacher Leser; 1 Mädchen starke Leserin; Schulstufe 4: n = 1 Mädchen starke Leserin, Schulstufe 5: n = 2 Jungen, sehr schwache Leser). Die Items wurden in der digitalen Version präsentiert3. Wir beobachteten die Kinder bei der Durchführung. Sie sollten laut mitdenken (think-aloud-Methode) sowie nach Abschluss des Tests ein paar Fragen beantworten (leitfadengestütztes Interview). Dadurch wollten wir Informationen über die Lösbarkeit der Aufgaben und die Benutzer*innenfreundlichkeit des Instruments erhalten. Anschließend wurden Texte und Items leicht verändert, um unnötige Mehrdeutigkeiten aus den Zielitems zu entfernen und die Items eindeutiger zu machen.

Nach diesen ersten Anpassungen wurde ein Pilotversuch mit einer ganzen Klasse der Schulstufe 3 durchgeführt (November 2019). Eine Beobachtungsphase und eine Diskussion in der Klasse nach dem Test ergaben, dass kein weiterer Bedarf an Item-Anpassungen bestand.

3.3 Zweite Pilotierung der Subtests: Ermittlung des finalen Item-Sets, der internen Konsistenz und Festlegung des Zeitlimits

Der GraLeV ist in seiner endgültigen Fassung als Geschwindigkeitstest konzipiert. Um Informationen über die Itemcharakteristika und die interne Testkonsistenz zu erhalten, ist es jedoch notwendig, in der Pilotierung die Tests als Powertests, bei denen (ohne Zeitbegrenzung) jedes Item von den meisten Kindern gelöst wird, durchzuführen. Um Frustration bei sehr schwachen Leser*innen zu vermeiden, wurde der Test abgebrochen, sobald 80 % der Schüler*innen in der Klasse den betreffenden Untertest abgeschlossen hatten. Das Powertest-Verfahren lieferte genügend Informationen, um die Schwierigkeit der Items und die Trennschärfe zu analysieren. Die Schwierigkeit der Items wurde in Excel als korrigierte Itemschwierigkeit unter Berücksichtigung der Anzahl der Distraktoren berechnet (Eid & Schmidt, 2014). Die Schwierigkeit der Items sollte vergleichsweise niedrig sein (mit hohen Koeffizienten von mindestens 0,6), da die endgültigen Tests als Geschwindigkeitstests durchgeführt werden sollten. Die Trennschärfe der Items wurde mit Hilfe der Reliabilitätsanalyse in SPSS ermittelt, die mit Cronbachs Alpha ein Maß für die interne Testkonsistenz liefert. Um die Zuverlässigkeit des Tests zu gewährleisten, muss die Trennschärfe eines Items so hoch wie möglich sein und darf niemals unter 0,3 liegen. Außerdem wurde eine Distraktorenanalyse durchgeführt, um problematische Distraktoren zu identifizieren. Alle Items, die die Gütekriterien nicht erfüllten, wurden aus dem Item-Set der Vorversionen ausgeschlossen und die interne Konsistenz des resultierenden Item-Sets (als Maß für die Reliabilität) wurde anschließend berechnet. Nach Bühner (2011) werden Reliabilitätswerte von 0,7 als akzeptabel und Werte über 0,8 als gut angesehen.

3 Aufgrund von Untersuchungen zum Vergleich zwischen Digital- und Printversion (siehe Seifert & Paleczek, 2022), gehen wir davon aus, dass die ermittelten Unterschiede zwischen den Versionen insbesondere auf den Faktor Zeit zurückzuführen sind. Die Itemkennwerte selbst sollten, wenn dann nur geringfügig beeinträchtigt sein. Dennoch ist eine Überprüfung der Kennwerte in der Printversion mit einer erneuten Powertestung angedacht.

Die Zeit, die die Schüler*innen für die Bearbeitung der einzelnen Aufgaben benötigten, wurde ebenfalls erfasst. Diese Informationen halfen bei der Festlegung von Zeitlimits für die Geschwindigkeitstests.

Die Datenerhebung fand von Oktober bis Dezember 2019 mit 273 Schüler*innen (117 Schulstufe 3; 156 Schulstufe 4; 49,1% Mädchen; 21,2% Kinder mit anderer Erstsprache als Deutsch) statt. Für jede Klasse gab es zwei verschiedene Testtage (an jedem Tag wurde etwa eine Unterrichtsstunde benötigt), da die vorläufige Powertest-Version des GraLeV aus einer relativ großen Anzahl von Items bestand. An Tag 1 bearbeiteten die Schüler*innen den Subtest Wort und den Subtest Text II. Am zweiten Tag bearbeiteten die Schüler*innen den Subtest Satz und den Subtest Text I. Projektmitarbeiter*innen führten den Test durch. Sie wurden geschult, um die Schüler*innen durch das Testverfahren zu führen. Für die Pilotierung wurde die Digitalversion verwendet.

Die Ergebnisse der Powertestung werden nachfolgend für jeden Subtest einzeln dargestellt, wobei für die Itemkennwerte des finalen Item-Sets auf das Kapitel 1.7 verwiesen wird.

3.3.1 Subtest Wort

Im Subtest Wort bearbeiteten die Schüler*innen insgesamt 38 Item-Sets (mit je 3 Zielitems). Nach der Analyse wurde ein Item-Set entfernt aufgrund einer zu niedrigen Itemschwierigkeit eines Items von 1, weitere fünf Sets aufgrund einer zu niedrigen Trennschärfe einzelner Items unter 0,2 und bei zwei weiteren Sets wurden häufig bestimmte Distraktoren gewählt, so dass anzunehmen ist, dass die Bilder mit einem Distraktor assoziiert werden konnten. Von den verbliebenen 30 Item-Sets wurden Items für geplante zwei Parallelversionen entsprechend einer ausgewogenen Verteilung (u.a. hinsichtlich Wortarten, Silbenzahlen, Frequenzen, Itemschwierigkeiten, Trennschärfen, orthographische Besonderheiten) festgelegt. Aufgrund des Fehlens einer ausreichend großen Anzahl von Items für zwei Parallelversionen in den anderen Subtests wurde jedoch letztlich nur eine Version realisiert. Eine spätere Veröffentlichung dieser Items in einer zweiten Version ist nachträglich noch möglich.

Die interne Konsistenz (Cronbachs Alpha) der finalen Testversion mit zwölf Item-Sets (mit je drei Zielitems) liegt bei $\alpha = 0{,}91$ und kann damit als gut bezeichnet werden. Die Items wurden dann nach ihrer korrigierten Itemschwierigkeit geordnet, beginnend mit dem leichtesten Item.

Die durchschnittliche Bearbeitungszeit der Schüler*innen pro Itemset (à drei Items) konnte anhand der aufgezeichneten Bearbeitungszeit pro Item berechnet werden. Diese lag bei den gewählten zwölf Item-Sets dieser Version bei 15,31 Sekunden ($SD = 2{,}54$). Die minimale Zeit, die für die Bearbeitung dieser zwölf Item-Sets aufgezeichnet wurde (bei Kindern, die alle dieser Item-Sets bearbeitet hatten, $N = 158$), lag bei 363 Sekunden, die Maximalzeit bei 1268 Sekunden. Die Zeit, die die 25% schnellsten Leser*innen benötigten, lag bei 557 Sekunden. Wir wollten, dass leseschwache Schüler*innen ausreichend viele Items bearbeiten können, um innerhalb dieser Gruppe differenzierte Aussagen zu den Lesefähigkeiten machen zu können. Auf der Grundlage dieser Ergebnisse und dem Bedürfnis, insgesamt eine ökonomische Testdurchführung in kurzer Zeit zu ermöglichen, beschlossen wir, das Zeitlimit auf drei Minuten festzulegen. Dieses Limit würde die schnellsten Leser*innen davon abhalten, alle Aufgaben zu lösen, jedoch schwächeren Leser*innen die Möglichkeit geben, zumindest einige Aufgaben zu lösen.

3.3.2 Subtest Satz

Im Subtest Satz bearbeiteten die Schüler*innen insgesamt 22 Items. Nach der Analyse wurden zwei Items aufgrund unzureichender Trennschärfen und vier weitere aufgrund eines zu plausiblen Distraktors, der in über 20% der Fälle gewählt wurde, ausgeschlossen. Die interne Konsistenz (Cronbachs Alpha) der Testversion mit den übrigen 16 Items liegt bei $\alpha = 0{,}81$ und kann damit als gut bezeichnet werden. Die Items wurden dann nach ihrer korrigierten Itemschwierigkeit geordnet, beginnend mit dem leichtesten Item. Jedoch wurde darauf geachtet, dass nicht zweimal hintereinander der gleiche grammatische Fokus verwendet wurde.

Die durchschnittliche Bearbeitungszeit der Schüler*innen pro Item konnte anhand der aufgezeichneten Bearbeitungszeit pro Item berechnet werden. Diese lag bei den gewählten 16 Items dieser Version bei 18,74 Sekunden (*SD* = 3,01). Die minimale Zeit, die für die Bearbeitung dieser 16 Items aufgezeichnet wurde (bei Kindern, die alle diese Items bearbeitet haben, *N* = 204), lag bei 278,34 Sekunden (Minimum: 114 Sekunden, Maximum: 601 Sekunden). Die Zeit, die die 25% schnellsten Leser*innen benötigten, lag bei 228,25 Sekunden. Auf der Grundlage dieser Ergebnisse und der Ökonomie in der Durchführung (kein Umdenken bei den verschiedenen Subtests notwendig) beschlossen wir, das Zeitlimit ebenfalls, wie bei Subtest Wort, auf 3 Minuten festzulegen. Dieses Limit würde die schnellsten Leser*innen davon abhalten, alle Aufgaben zu lösen, jedoch schwächeren Leser*innen die Möglichkeit geben, zumindest ein paar Aufgaben zu lösen.

3.3.3 Subtest Text I

Im Subtest Text I bearbeiteten die Schüler*innen 30 Aufgaben (15 Geschichten mit je zwei Fragen). Nach der Analyse der Aufgaben wurden acht Aufgaben mit einem Schwierigkeitsgrad unter 0,6 ermittelt. Außerdem wurden die Distraktoren von drei dieser Aufgaben von mehr als 20% der Schüler*innen gewählt. Die Trennschärfe eines dieser Items lag zudem unter 0,2. Diese Items wurden nicht in das finale Item-Set aufgenommen. Ein Text wurde nur dann in das finale Item-Set aufgenommen, wenn beide zugehörigen Fragen die Qualitätskriterien erfüllten. So wurden sieben Texte ausgeschlossen, was zu einem finalen Item-Set von acht Texten führte (sechs kurze Texte, zwei lange Texte mit jeweils zwei Fragen: 16 Items). In Bezug auf Cronbachs Alpha ($\alpha = 0{,}87$) war die interne Konsistenz dieses endgültigen Item-Sets gut. Die Items wurden dann nach ihrer korrigierten Itemschwierigkeit geordnet, beginnend mit dem leichtesten Item.

Die Zeit, die die Schüler*innen zur Beantwortung beider Fragen für jeden Text benötigten, wurde aufgezeichnet. Daraus errechneten wir die durchschnittliche Zeit, die für die Bearbeitung der 16 Items des finalen Item-Sets benötigt wurde: 333 Sekunden (*SD* = 118). Die minimale Zeit, die für die Bearbeitung dieses Item-Sets aufgezeichnet wurde (72 Sekunden), wurde als nicht zuverlässig erachtet. Wir gingen davon aus, dass der*die Schüler*in sich nur durch die Antworten geklickt hatte, ohne den Text gelesen zu haben (z. B. brauchte er*sie nur 2 Sekunden für die Lösung der beiden Fragen zu Text 12). Wir zogen also die Zeit heran, die die 25% schnellsten Leser*innen benötigten. Diese brauchten 258 Sekunden. Die Maximalzeiten wurden auch berücksichtigt, um zu bestimmen, wie lange langsame Leser*innen für einen Text mit zwei Fragen brauchten. Wir wollten erreichen, dass leseschwache Schüler*innen zumindest die Zeit hatten eine oder zwei Fragen bearbeiten zu können. Dies ermöglicht es uns auch, innerhalb der Gruppe der schwachen Leser*innen zu differenzieren. So wurde die maximal benötigte Zeit für den einfachsten Text im letzten Item-Set (Kurztext 6) analysiert. Der*die langsamste Leser*in der Stichprobe löste dieses Item in 116 Sekunden. Auf der Grundlage dieser Ergebnisse

beschlossen wir, das Zeitlimit auf 3 Minuten festzulegen. Dieses Limit würde die schnellsten Leser*innen davon abhalten, alle Aufgaben zu lösen, und schwächeren Leser*innen zusätzlich die Möglichkeit geben, zumindest eine Aufgabe zu bearbeiten.

3.3.4 Subtest Text II

Im Subtest Text II bearbeiteten 246 bis 251 Schüler*innen die vier Texte und damit die 60 Items der Vorversion. Eine Analyse der Items ergab, dass Text 1 und Text 4 Items mit einer korrigierten Itemschwierigkeit unter 0,6 (zwei Items) bzw. einer Trennschärfe unter 0,2 (vier Items) enthielten. Text 2 (leicht) und Text 3 (anspruchsvoller) wiesen hingegen gute Itemkennwerte auf und wurden in die endgültige Testversion aufgenommen. Die internen Konsistenzen von Text 2 (Cronbachs $\alpha = 0{,}80$) und 3 (Cronbachs $\alpha = 0{,}78$) waren zufriedenstellend. Bis auf ein Item in Text 3 wiesen alle Items akzeptable Werte auf. Eine getrennte Betrachtung der Itemcharakteristika in den Schulstufen 3 und 4 ergab jedoch, dass in der Stichprobe der Schüler*innen der Schulstufe 3 drei weitere Items in Text 2 eine Trennschärfe unter 0,2 aufwiesen. Zur Verbesserung der Itemcharakteristika in Text 2 wurde daher ein Satz leicht verändert und drei Distraktoren ausgetauscht. Bei einem Item in Text 3 wurde ein Distraktor geändert.

Betrachtet man die Lösungszeiten der 172 Schüler*innen, die Text 2 und 3 vollständig bearbeitet haben, so zeigt sich, dass für Text 2 etwa 167 Sekunden ($SD = 63{,}59$) und für Text 3 174 Sekunden ($SD = 62{,}22$) benötigt wurden. Die schnellsten Leser*innen brauchten 48 Sekunden für Text 2 und 66 Sekunden für Text 3. Auch hier haben wir uns an der Zeit orientiert, die die 25% schnellsten Leser*innen benötigten, d. h. etwa 122 Sekunden für Text 2 und 130 Sekunden für Text 3. Um sicherzustellen, dass auch langsame Leser*innen zumindest einige Items bearbeiten können, wurde die für die Texte aufgezeichnete Maximalzeit durch die 15 Items pro Text geteilt. Bei Text 2 benötigte der*die langsamste Leser*in demnach etwa 30 Sekunden für die Lösung eines Items. Für Text 3 benötigte der*die langsamste Leser*in 29 Sekunden. Auf der Grundlage dieser Ergebnisse beschlossen wir, das Zeitlimit für das Lesen beider Texte, die nacheinander präsentiert werden, auf insgesamt 100 Sekunden festzulegen. Dies ermöglicht es sehr langsamen Leser*innen, zumindest einige Aufgaben zu lösen, und verhindert, dass schnelle Leser*innen die Items beider Texte vor Ablauf der Zeit bearbeiten.

4. Gütekriterien

In der klassischen Testtheorie werden neben den drei Hauptgütekriterien Objektivität, Reliabilität und Validität eine Reihe an Nebengütekriterien angeführt. Zu diesen zählen die Ökonomie, die Zumutbarkeit, die Nützlichkeit und das Vorhandensein von Normen (für einen detaillierten Überblick siehe Bühner, 2011; Lienert & Raatz, 1998). Für den vorliegenden GraLeV kann festgehalten werden, dass dieser als sehr ökonomisch zu beurteilen ist, da die Bearbeitungszeit für einen komplexen Blick auf die Leseverständnisfähigkeiten nur knapp über 10 Minuten (ohne Instruktion) beträgt. Dadurch ist der Test sowohl den Testleiter*innen als auch den Schüler*innen zumutbar. Zudem liegt eine umfassende Normierung anhand einer repräsentativen Normstichprobe vor. Die entsprechenden Normtabellen finden sich im Anhang des Manuals.

Im Folgenden werden die Hauptgütekriterien Objektivität, Reliabilität und Validität für den GraLeV dargestellt. Dabei ist zu beachten, dass die einzelnen Gütekriterien anhand unterschiedlicher Stichproben bestimmt wurden.

Zusammenfassend zeigen die Ergebnisse der testtheoretischen Analyse, dass der GraLeV sowohl im Hinblick auf Nebengütekriterien (Ökonomie, Nutzen, etc.) als auch bezüglich seiner Hauptgütekriterien (Objektivität, Reliabilität und Validität) ein hochwertiges Verfahren zur Erhebung der Leseverständnisfähigkeiten bei Kindern der dritten bis vierten Schulstufe darstellt.

4.1 Objektivität

Im Sinne der klassischen Testtheorie ist Objektivität durch drei Aspekte bestimmt:

- Durchführungsobjektivität
- Auswertungsobjektivität
- Interpretationsobjektivität

Ein Test ist objektiv, wenn seine Ergebnisse sowohl unabhängig von der durchführenden als auch der auswertenden Person sind.

Die *Durchführungsobjektivität* ist durch eine standardisierte Durchführungsanleitung (siehe Kapitel 2.5) gewährleistet, die eine wörtliche Instruktion enthält, mit Hilfe derer die Lehrpersonen bzw. Testleiter*innen den GraLeV durchführen sollen. Trotzdem besteht bei Gruppentestverfahren die Möglichkeit, dass Einschränkungen der Durchführungsobjektivität, beispielsweise aufgrund der Vorbereitung auf das Testverfahren (z. B. Motivationsaufbau) oder wegen des Geräuschpegels während der Testung entstehen können. Um dies zu vermeiden, sollte der GraLeV in einer möglichst ruhigen Umgebung durchgeführt und Fragen sollten mit den Schüler*innen bereits während der Instruktionsphase, in der die Beispielitems besprochen werden, geklärt werden.

Die *Auswertungsobjektivität* des vorliegenden Tests ist gegeben, da anhand einer Auswertungsschablone bestimmt werden kann, welche Antworten als korrekt bzw. falsch zu werten sind. Ebenfalls dargestellt ist, wie der Gesamtwert pro Subtest berechnet wird. Dadurch ist die eindeutige Bestimmung des Gesamtwertes gesichert und es gibt keinen Interpretationsspielraum für die auswertende Person (Hinweise zur Auswertung, siehe Kapitel 1.3)

Auch die *Interpretationsobjektivität* des GraLeV kann als gesichert gelten, da die individuellen Testwerte einer Person mit der jeweiligen Normstichprobe verglichen werden und dieser Vergleich korrekte Interpretationen möglich macht (Hinweise zur Interpretation, siehe Kapitel 1.3).

Insgesamt kann der GraLeV bei genauer Durchführung, Auswertung und Interpretation gemäß Testmanual als objektiv eingeschätzt werden.

4.2 Reliabilität

Reliabilitätskennwerte liefern Hinweise über die Messgenauigkeit eines Tests. Zur Erfassung der Reliabilität eines Testverfahrens stehen verschiedene Möglichkeiten zur Verfügung.

Die Ermittlung der internen Konsistenz (siehe Kapitel 3.3) wurde im Zuge der zweiten Pilotierung bei Darbietung als Powertest (ohne Zeitlimit) erfasst, um zu prüfen, inwiefern die Items jedes Subtests miteinander zusammenhängen.

Für die Erfassung der Reliabilität des Verfahrens unter Angabe des Zeitlimits wurden einerseits die Retest-Reliabilität, andererseits die Split-Half-Reliabilität berechnet.

4.2.1 Retest-Reliabilität

Die Retestung zur Erhebung der Reliabilität der print-Version fand im Zusammenhang mit der Normierung des GraLeV im Herbst 2021 statt. Alle Lehrpersonen, die an der Normierung teilgenommen haben, wurden gefragt, ob sie uns helfen könnten und zusätzlich eine wiederholte Testung im Abstand von 2 Wochen durchführen würden. Zugesagt haben die Lehrpersonen von insgesamt acht Klassen. Dadurch konnten die Daten von 112 Kindern (72 Schulstufe 3, 40 Schulstufe 4) für die Analyse der Retest-Reliabilität verwendet werden.

Reliabilitätskoeffizienten über 0,7 werden als akzeptabel angesehen (Bühner, 2011). Wie aus Tabelle 11 hervorgeht, wurden für die Gesamtstichprobe für alle Subtests akzeptable Reliabilitäts-Koeffizienten erzielt.

Table 11. Retest-Reliabilität.

Wort			Satz			Text I			Text II		
t1 *M* *(SD)*	t2 *M* *(SD)*	*r*	t1 *M* *(SD)*	t2 *M* *(SD)*	*r*	t1 *M* *(SD)*	t2 *M* *(SD)*	*r*	t1 *M* *(SD)*	t2 *M* *(SD)*	*r*
26,33 (7,54)	30,35 (6,43)	.81**	8,51 (3,34)	10,35 (3,42)	.80**	7.31 (3,25)	9.06 (3,45)	.84**	8,86 (5,69)	9,75 (4,82)	.79**

Anmerkung: ** p<.01

4.2.2 Split-Half-Reliabilität

Anhand der Daten der Normstichprobe zu Beginn des Schuljahres 21/22 wurde die Odd-Even-Split-Half-Reliabilität ermittelt. Die Testitems wurden dabei immer abwechselnd den beiden Testhälften zugeordnet (Testhälfte 1: Item 1, 3, 5, 7, usw.; Testhälfte 2: Item 2, 4, 6, 8, usw.). Bei dem Subtest Text I werden dabei beide zu einem Text gehörenden Fragen in der gleichen Testhälfte belassen. Jeweils die gültigen Fälle sind in Tabelle 12 aufgelistet. Die Spearman-Brown-Koeffizienten (korrigierte Reliabilitätsschätzung, r_{korr}) zeigen jeweils sehr hohe Split-Half-Reliabilitätswerte an (siehe Tabelle 12).

Tabelle 12. Split-Half-Reliabilität.

Wort				Satz				Text I				Text II			
N = 300[a]				*N* = 221[a]				*N* = 186[a]				*N* = 140[a]			
Teil 1	Teil 2	*r*	r_{korr}	Teil 1	Teil 2	*r*	r_{korr}	Teil 1	Teil 2	*r*	r_{korr}	Teil 1	Teil 2	*r*	r_{korr}
,87	,85	,90	,95	,81	,83	,92	,96	,83	,86	,90	,95	,93	,93	,97	,99

Anmerkung: r_{korr} = korrigierte Reabilitätsschätzung, Spearman-Brown-Koeffizient; [a] = Anzahl der gültigen Fälle

4.3 Validität

Die Validität eines Tests ist nach Lienert and Raatz (1998) als das Ausmaß definiert, in dem ein Test das Merkmal, das er vorgibt zu messen (in diesem Fall die Leseverständnisfähigkeit), auch wirklich misst. Für den GraLeV wird im Folgenden die inhaltliche Validität kurz diskutiert und die Konstruktvalidität anhand von Korrelationen mit Tests und Einschätzungen durch Lehrer*innen, die entweder dasselbe Konstrukt oder ein anderes Konstrukt messen bzw. einschätzen, analysiert.

4.3.1 Inhaltliche Validität

In Bezug auf die inhaltliche Validität kann angemerkt werden, dass der GraLeV Aufgaben beinhaltet, die die Leseverständnisfähigkeit des Kindes auf Wort-, Satz- und Textebene messen. Dies geschieht durch die vier Subtests, die als unmittelbare Indikatoren für die entsprechenden Fähigkeiten angesehen werden können. Augenscheinlich ist somit eine hohe inhaltliche Validität gegeben.

4.3.2 Konstruktvalidität

Die Konstruktvalidität beschreibt, ob ein Test das erfasst, was er zu messen beansprucht. Hierzu wird überprüft, ob die Korrelationen zu ähnlichen Tests, die dasselbe oder ein ähnliches Konstrukt messen (konvergente Validität), höher ausfallen als Korrelationen mit Tests anderer Gültigkeitsbereiche (diskriminante Validität). Im Folgenden werden die Analysen des Zusammenhangs zwischen dem GraLeV und dem Leseverständnistest für Erst- bis Sechstklässler 2 (ELFE II: Lenhard, Lenhard, & Schneider, 2020) und den Einschätzungen durch Lehrpersonen dargestellt.

Die Datenerhebung fand in dritten und vierten Schulstufen von September bis November 2020 statt. Projektmitglieder führten die Tests an zwei aufeinanderfolgenden Tagen durch, um die Testzeit nicht zu lang zu halten und damit Überforderung zu vermeiden (Tag 1: GraLeV und ELFE II auf Textebene; Tag 2: GraLeV und ELFE II Wort- und Satzebene). Da nicht alle Kinder an beiden Tagen anwesend waren, kommt es zu unterschiedlichen Stichprobenzahlen für die einzelnen Subtests.

Insgesamt nahmen 534 Schüler*innen an der Erhebung für die Konstruktvalidität teil, 333 davon waren Schüler*innen der dritten Schulstufe. Wenn die Lehrpersonen angaben, dass der*die Schüler*in zu Hause mindestens eine andere Sprache als Deutsch spricht, wurde der*die Schüler*in als Kind mit einer anderen Erstsprache als Deutsch (L2) definiert.

Die Einschätzung der Lehrpersonen zu den Lese-, Mathematik- und sozial-emotionalen Fähigkeiten lag für 458 Schüler*innen vor. Die Tabelle 13 enthält weitere Einzelheiten.

Tabelle 13. Angaben zur Validierungsstichprobe.

Schulstufe	N	Alter *M (SD)*	% weiblich	% L2	% SPF
3	333	8,78 (0,47)	43,0	33,5	0,7
4	201	9,82 (0,46)	54,8	43,4	1,3
Gesamt	**534**	**9,14 (0,68)**	**47,1**	**37,0**	**1,0**

Anmerkung: Die Angaben zu Alter, Geschlecht und der zu Hause gesprochenen Sprache wurden nur für 337, 352 bzw. 356 Schüler*innen gemacht. Die prozentualen Berechnungen beruhen ausschließlich auf den erhobenen Daten. Die fehlenden Schüler*innendaten werden ignoriert.

Konvergente Validitätsmaße umfassten die Einschätzung der Lehrpersonen (LE) der Lesekompetenz der Schüler*innen und die Leistung der Schüler*innen im Leseverständnistest ELFE II (Lenhard et al. 2020). Nach Bühner (2011) muss der Korrelationskoeffizient der konvergenten Validität über 0,5 liegen, um darauf schließen zu können, dass die Tests dieselbe Fähigkeit messen.

Die divergente Validität wurde auf der Grundlage der LE der mathematischen und sozial-emotionalen Fähigkeiten der Schüler*innen berechnet.

Tabelle 14 zeigt die Korrelationen der vier GraLeV-Subtests mit dem ELFE-II-Leseverständnistest und der LE zum Leseverständnis (konvergente Validität) sowie die Korrelationen mit der LE zu den mathematischen und sozial-emotionalen Kompetenzen (divergente Validität). Erwartungsgemäß korrelierten die GraLeV-Subtests am stärksten mit den ELFE-II-Subtests (.60 bis .76). Ebenso korrelierten die GraLeV-Werte stark mit der LE des Leseverständnisses (.33 bis .46). Zwar gibt es signifikante Korrelationen der GraLeV-Subtests mit der LE der mathematischen Fähigkeiten (.26 bis .33), doch sind diese meist geringer als die für die konvergente Validität gefundenen Korrelationen der jeweiligen Subtests (z. B. korreliert Subtest Text I signifikant höher mit der LE des Textverständnisses als mit der LE der räumlich-visuellen Fähigkeiten: $z = 1.74$, $p < .05$). Im Allgemeinen korrelieren die Werte der Untertests am wenigsten mit der LE der sozial-emotionalen Fähigkeiten (.14 bis .21, siehe Tabelle 14).

Tabelle 14. Validität.

	ELFE II			LE Leseverständnis			LE mathematische Fähigkeiten		LE sozial-emotionale Fähigkeiten
	Wort	Satz	Text	Wort	Satz	Text	Numerisches Verständnis	Räumlich-visuelle Fähigkeiten	
Wort	,62**	,64**	,64**	,38**	,40**	,42**	,26**	,27**	,18**
Satz	,60**	,69**	,66**	,43**	,43**	,46**	,28**	,33**	,14**
Text I	,63**	,70**	,69**	,33**	,35**	,40**	,31**	,33**	,16**
Text II	,69**	,76**	,75**	,39**	,41**	,46**	,29**	,32**	,21**

Anmerkung: LE = Lehrer*inneneinschätzung; ** $p<.01$

Die LE waren hoch miteinander korreliert. Erwartungsgemäß korrelierten die LE der verschiedenen Ebenen des Leseverständnisses hoch miteinander (.87 bis .96), ebenso wie die LE der beiden mathematischen Fähigkeiten (.77). Interessanterweise gab es auch hohe Interkorrelationen zwischen der LE der Lesekompetenz und der LE der mathematischen Fähigkeiten (.51 bis .61) sowie zwischen den LE der Lesekompetenz und der sozial-emotionalen Fähigkeiten (.41 bis .43). Dieses Ergebnis stimmt mit Studien überein, die gezeigt haben, dass Lehrpersonen dazu neigen, ein eher ganzheitliches Bild von ihren Schüler*innen zu haben und nicht zwischen verschiedenen Teilfähigkeiten zu unterscheiden (u.a. Paleczek, Seifert, & Gasteiger-Klicpera, 2017).

4.4 Normierung

Der GraLeV wurde zu Beginn des Schuljahres 2021/22 (Oktober-November 2021), sowie zum Ende des Schuljahres 2022/23 (Mai-Juni 2023) normiert. Die Daten wurden nach intensiver Online-Einschulung von Lehrpersonen selbstständig erhoben, die Auswertung übernahmen überwiegend Projektmitarbeiter*innen der Universität Graz. In einigen Fällen wurde auch die Auswertung (nach Einschulung) von Lehrpersonen selbstständig gemacht und Eintragungen in eine leere Excel-Datenmaske vorgenommen. Im gesamten Normierungsprozess stand eine Autorin des GraLeV fortwährend für Fragen zur Verfügung.

Die Daten für die Normierung wurden größtenteils in Österreich (Region: Steiermark, Kärnten), aber auch teilweise in Deutschland (Region: Nordrhein-Westfalen, Niedersachsen, Brandenburg) erhoben. Innerhalb beider Regionen wurde sowohl mit Schulen aus städtischen Bezirken als auch mit Schulen aus ländlichen Gegenden kooperiert. An den Testungen nahmen immer die gesamten Schulklassen teil, sofern für die Kinder das Einverständnis der Eltern vorlag. Schüler*innen mit einem sonderpädagogischen Förderbedarf (SPF) wurden in die Normierung einbezogen, um auch die Heterogenität, die in Klassen herrscht, abzubilden und den GraLeV als Instrument, das in allen Klassen mit allen Kindern einsetzbar ist, zu normieren. Eine Übersicht über die Stichprobengrößen und Verteilung von Merkmalen (Erhebung in Österreich/Deutschland, Geschlecht, Deutsch als Erstsprache (L1)/ andere Erstsprache als Deutsch (L2), SPF, Alter) befindet sich in Tabelle 15.

Die Stichprobengrößen für die dritte Schulstufe zu beiden Messzeitpunkten, sowie die der vierten Schulstufe zu Schuljahresbeginn können als repräsentativ angesehen werden. Diejenige der vierten Schulstufe zum Schuljahresende ist vergleichsweise klein. Dies ist vermutlich der Tatsache geschuldet, dass die Schüler*innen die Grundschule in den teilnehmenden Bundesländern nach der vierten Schulstufe verlassen und somit die Ergebnisse für die Lehrpersonen weniger Relevanz hatten als zu Schuljahresbeginn in dieser Schulstufe. Dadurch ist bei der Einschätzung eines zu testenden Kindes in diesem Erhebungszeitraum keine gesicherte Aussage über die tatsächliche Leistung möglich.

Tabelle 15. Übersicht über die Normierungs-Stichprobe Schuljahresbeginn und -ende.

Schulstufe	Anzahl der Kinder	Anteil österreichische Kinder[a]	Anteil weiblich	Anteil L2 Deutsch[b]	Anteil Kinder mit SPF[c]	Alter *MW (SD)*	Alter Range
Schuljahresbeginn							
3.	723	80,1%	50,6%	19,1%	1,4%	8,91 (0,48)	8,08 – 11,00
4.	513	96,9%	50,0%	10,5%	1,2%	9,86 (0,50)	8,25 – 12,08
Gesamt	**1236**	**87,1%**	**50,4%**	**15,6%**	**1,4%**	**9,30 (0,68)**	**8,08 – 12,08**
Schuljahresende							
3.	241	91,3%	49,0%	24,5%	1,2%	9,47 (0,54)	8,75 – 11,41
4.	90	100%	47,8%	41,1%	1,1%	10,55 (0,53)	9,83-12,33
Gesamt	**331**	**93,7%**	**48,6%**	**29,0%**	**1,2%**	**9,78 (0,73)**	**8,75 – 12,33**

Anmerkung: [a] vorrangig Bundesländer Steiermark und Kärnten, die Daten der übrigen Kinder wurden in einzelnen Schulen in Deutschland (Bundesländer Niedersachsen, Nordrhein-Westfalen und Brandenburg) erhoben; [b] L2 Deutsch = vereinfacht für Kinder mit einer anderen Erstsprache als Deutsch; [c] Sonderpädagogischer Förderbedarf (SPF) festgestellt oder beantragt

5. Anwendungsmöglichkeiten

Der GraLeV wurde als Gruppentestverfahren entwickelt und erfasst die Leistung im Leseverständnis bei Grundschulkindern der dritten und vierten Schulstufe (eine Ausweitung auf Mitte der zweiten Schulstufe, sowie auf die fünfte Schulstufe ist geplant, Untersuchungen dazu laufen bereits). Der Test kann jeweils zu Beginn und am Ende des dritten bzw. vierten Schuljahres eingesetzt werden. Er ermöglicht eine separate Betrachtung der Leistungen im Wort-, Satz- und Textverständnis. Das Verfahren beschränkt sich somit auf einen Teilaspekt der Lesefähigkeiten, der nach dem Erwerb grundlegender Lesefertigkeiten zunehmend benötigt und vorausgesetzt wird, um selbstständig Texte zu lesen und sich Wissen erschließen zu können. Mit dem GraLeV können Auffälligkeiten einzelner Kinder ökonomisch erkannt werden, um entsprechende Fördermaßnahmen, orientiert an den Bedürfnissen der jeweiligen Kinder, einleiten und den Lernfortschritt dokumentieren zu können.

Der GraLeV kann sowohl von Regelschul- als auch Sonder- bzw. Förderschullehrpersonal, Inklusionspädagog*innen, Lese- bzw. Sprachtherapeut*innen gleichermaßen wie auch Forscher*innen sowohl im Klassenverband (Gruppentestung) als auch in einem Setting der Einzeltestung angewendet werden. Der Test kann zum Schuljahresbeginn und -ende der dritten und vierten Schulstufe (jeweils erste und letzte acht Wochen) eingesetzt werden.

Das Verfahren bietet folgende Vorteile:

1 Aufgrund seines geringen Durchführungs- und Zeitaufwandes kann der GraLeV auch als Bestandteil einer Testbatterie eingesetzt werden (beispielsweise in Verbindung mit einem Dekodiertest, Wortschatztest, Grammatiktest, etc.).

2 Der Test wurde an einer großen und repräsentativen Stichprobe (N_{gesamt} = 1567) normiert, die sowohl Kinder mit sonderpädagogischem Förderbedarf (SPF), als auch Kinder mit einer anderen Erstsprache als Deutsch einschloss. Für letztere wird vereinfachend der Begriff Deutsch als Zweitsprache verwendet (L2 Deutsch).

3 Bei der Normierung wurde der GraLeV von allen beteiligten Lehrpersonen nach Einschulung selbstständig durchgeführt. Die Normen sind damit unter realen Testbedingungen entstanden.

4 Das Testverfahren wird Open Access angeboten und steht damit allen Anwender*innen kostenfrei zur Verfügung.

6. Kurzfassung

Diagnostische Zielsetzung: Der GraLeV ist ein Verfahren zur Messung des Leseverständnisses (auf Wort-, Satz- und Textebene) von Kindern der dritten und vierten Schulstufe.

Aufbau: Der Test besteht aus vier Subtests. Der *Subtest Wort* (Zeitlimit 3 Minuten, 36 Items in 12 Itemsets, 2 Beispiel-Itemsets) ist eine Wort-Bild-Zuordnungsaufgabe, bei der die Kinder in jedem Itemset drei Bilder mit den passenden Wörtern, die sie lesen, verbinden (sechs Antwortalternativen pro Itemset: drei Zielitems, drei Distraktoren). Der *Subtest Satz* (Zeitlimit 3 Minuten, 16 Items, 2 Beispiel-Items) ist eine Satz-Bild-Zuordnungsaufgabe, bei der die Kinder zu einem Bild aus vier Sätzen jenen wählen und ankreuzen, der am besten zum Bild passt. Der *Subtest Text I* (Zeitlimit 3 Minuten, 16 Items in 8 Itemsets, 1 Beispiel-Itemset) besteht aus kurzen Quatschgeschichten (Texte über Nonsens-Wörter) mit jeweils 2 Fragen, die immer gleich formuliert sind („Was steht in der Geschichte?“). Eine der Fragen prüft explizite Informationsentnahme, die zweite Frage prüft das Ziehen einfacher Schlussfolgerungen aus dem Text. Die Kinder kreuzen zu jeder Frage den korrekten Satz bei einer Auswahl aus vier Sätzen an. Der *Subtest Text II* (Zeitlimit 100 Sekunden, 30 Items in 2 Texten, Beispieltext mit drei Items) besteht aus zwei Lückentexten, in welchen an der Stelle jedes siebten Wortes eine Auswahl aus drei Wörtern vorzunehmen ist. Die Kinder kreisen dasjenige Wort, welches in den Text passt, ein.

Grundlagen und Konstruktion: Leseverständnis wird als vielschichtiger, komplexer und hierarchischer Prozess angesehen. Bei der Konstruktion des GraLeV wurde dies durch die Überprüfung des Leseverständnisses auf Wort-, Satz- und Textebene berücksichtigt. Bei der Konstruktion wurden gängige Methoden der Leseverständnismessung berücksichtigt.

Empirische Prüfung und Gütekriterien:

Reliabilität: Der GraLeV weist sehr zufriedenstellende Reliabilitätswerte auf (Split-Half-Reliabilität: .95-.99; Retest-Reliabilität: .79 - .84).

Validität: Neben der augenscheinlichen inhaltlichen Validität weist der GraLeV auch zufriedenstellende Werte in der konvergenten und der diskriminanten Validität auf. Es bestehen erwartungskonforme Zusammenhänge zwischen GraLeV und ELFE II (.60 bis .76) sowie mit Einschätzung des Leseverständnisses durch Lehrpersonen (.33 bis .46) und niedrigere Korrelationen mit der Lehrpersoneneinschätzung der mathematischen Fähigkeiten (.26 bis .33) und der sozial-emotionalen Fähigkeiten (.14 bis .21).

Normen: Es liegen Prozentrangnormen und z-Werte für den GraLeV aus einer repräsentativen Stichprobe (N = 1567) vor, die hauptsächlich in Österreich (Schuljahresbeginn 87,1%, Schuljahresende 93,7%), jedoch auch Deutschland erhoben wurden. Die Normen werden getrennt für die dritte und vierte Schulstufe und jeweils für den Schuljahresbeginn und das Schuljahresende berichtet.

7. Bewertung

Der GraLeV ist ein ökonomisches Verfahren zur Überprüfung des Leseverständnisses im Klassenverband in der dritten und vierten Schulstufe. Er liegt derzeit frei verfügbar in der Printversion vor. Angedacht ist, auch die Digitalversion langfristig frei nutzbar zu machen. Eine Ausweitung für die zweite Schulstufe sowie fünfte Schulstufe ist, ggf. mit Adaptierung der Zeit, in Vorbereitung.

8. Literatur

Publikationen über das Verfahren GraLeV:

Paleczek, L., & Seifert, S. (2021). Diagnostik als Basis für Förderung: Der Grazer Leseverständnistest. *Sprachtherapie aktuell: Forschung - Wissen – Transfer, 8*(2), e2021-36.

Seifert, S., & Paleczek, L. (2020). Development of a German Digital Assessment of Reading Comprehension in Grades 3-4. Proceedings of the 19th European Conference on e-Learning. Reading: Academic Conferences International Limited. DOI: 10.34190/EEL.20.014.

Seifert, S., & Paleczek, L. (2021). Digitally Assessing Text Comprehension in Grades 3-4: Test Development and Validation. *Electronic Journal of E-Learning*, *19*(5), 336-348. DOI: 10.34190/ejel.19.5.2467.

Seifert, S., & Paleczek, L. (2022). Comparing tablet and print mode of a German reading comprehension test in grade 3: Influence of test order, gender and language. *International Journal of Educational Research, 113*. DOI: 10.1016/j.ijer.2022.101948.

Verwendete Literatur:

Afflerbach, P. (2016). Reading Assessment. Looking Ahead. *The Reading Teacher, 69*(4), 413-419. doi:10.1002/trtr.1430

Bamberger, R. & Vanecek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache*. Wien: Jugend und Volk.

Brasher, C.F. (2017). *Beyond Screening and progress monitoring: An examination of the reliability and concurrent validity of maze comprehension assessments for fourth-grade students*. Dissertation, Middle Tennessee State University.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.). *Psychologie*. München, Boston [u.a.]: Pearson Studium.

Conoyer, S.J., Lembke, E.S., Hosp, J.L., Espin, C.A., Hosp, M.K., & Poch, A.L. (2017). Getting More From Your Maze: Examining Differences in Distractors. *Reading & Writing Quarterly, 33*(2), 141-154. doi:10.1080/10573569.2016.1142913

Fuchs, L.S. & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*(1), 45-58.

Garcia, J.R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84*, 74-111. doi:10.3102/003465313499616

Glück, C. W. (2007). *WWT 6-10 Wortschatz- und Wortfindungstest für 6- bis 10-Jährige*. München: Urban & Fischer.

Guthrie, J.T., Seifert, M., Burnham, N.A., & Caplan, R.I. (1974). The Maze Technique to Assess, Monitor Reading Comprehension. *The Reading Teacher, 28*(2), 161-168.

Keenan, J.M., Betjemann, R.S., & Olson, R.K. (2008). Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension. *Scientific Studies of Reading, 12*(3), 281–300. doi:10.1080/10888430802132279

Kendeou, P., Papadopoulos, T.C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*, 354-367. doi:10.1016/j.learninstruc.2012.02.001

Ketterlin-Geller, L.R., McCoy, J.D., Twyman, T. & Tindal, G. (2006). Using a Concept maze to Assess Student Understanding of Secondary-Level Content. Assessment for Effective Intervention, 31(2), 39-50. 10.1177/073724770603100204.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Klicpera, C., Schabmann, A., Gasteiger-Klicpera, B. & Schmidt, B. (2017). *Legasthenie – LRS. Modelle, Diagnose, Therapie und Förderung.* 5th edn. Stuttgart: UTB.

Lenhard, W. (2013). *Leseverständnis und Lesekompetenz: Grundlagen – Diagnostik – Förderung*. Stuttgart: Kohlhammer.

Lenhard, A., Lenhard, W., & Schneider, W. (2020). ELFE II. *Ein Leseverständnistest für Erst- bis Siebtklässler*: Version II (4. unveränderte Auflage). Göttingen: Hogrefe.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl., Studienausg). Weinheim: Beltz, Psychologie Verl.-Union.

Marx, P. (2007). *Lese- und Rechtschreiberwerb*. *StandardWissen Lehramt: Vol. 2946*. Paderborn, München, Wien, Zürich: Schöningh.

Mayringer, H., & Wimmer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9*: SLS 2-9. Göttingen: Hogrefe.

Muijselaar, M.M.L., Kendeou, P., de Jong, P.F., & van den Broek, P.W. (2017). What Does the CBM-Maze Test Measure? *Scientific Studies of Reading, 21*(2), 120-132. doi:10.1080/10888438.2016.1263994

Mullis, I.V.S., & Martin, M.O. (Eds.). (2015). *PIRLS 2016 Assessment Framework*. 2nd ed. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Paleczek, L., & Seifert, S. (2019). Pädagogische Diagnostik und deren Bedeutung für inklusiven Leseunterricht. In L. Paleczek, & S. Seifert (eds.), Inklusive(r) Leseunterricht: Leseentwicklung, Diagnostik und Konzepte (pp. 125-147). Wiesbaden: Springer VS.

Paleczek, L., Seifert, S. & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: a multilevel analysis. *Psychology in the Schools*, 54(3), 228-245. https://doi.org/10.1002/pits.21993

Perfetti, C.A., Landi, N., & Oakhill, J. (2005). The Acquisition of Reading Comprehension Skill. In M.J. Snowling, & C. Hulme (eds.), *The Science of Reading: A Handbook* (pp. 227-247). Oxford: Blackwell.

Richter, T., & Christmann, U. (2009). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben, & B. Hurrelmann (eds.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (pp. 25-85). Weinheim: Beltz.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen: Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten.* Weinheim: Psychologie Verlags Union.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015a). childLex (German Children's Book Corpus). Retrieved from https://www.mpib-berlin.mpg.de/de/forschung/max-planck-forschungsgruppen/mpfg-read/projekte/childlex

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology, 25*(2), 121-148. doi:10.1080/02702710490435727

Stahl, S.A., & Hiebert, E.H. (2005). The "word factors": A problem for reading comprehension assessments. In S.G. Paris, & S.A. Stahl (eds.), *Children's reading comprehension and assessment* (pp. 161–186). Mahwah, NJ: Erlbaum.

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

Wayman, M.M., Wallace, T., Wiley, H.Y., Tichá, R., & Espin, C.A. (2007). Literature Synthesis on Curriculum-Based Measurement in Reading. *Journal of Special Education, 41*(2), 85-120. doi:10.1177/00224669070410020401

Widauer, K. & Wallner-Paschon, C. (2017). Entwicklung und Aufbau der Testinstrumente und Kontextfragebögen. In Wallner-Paschon, C. & Itzlinger-Bruneforth, U. (eds.), *PIRLS 2016. Technischer Bericht* (pp.9-22). Salzburg.

Wild, J. & Pissarek, M. (2019). *Regensburger Analysetool für Texte. Dokumentation*, available at:<https://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/downloads/ratte/index.html> [Accessed 25 July 2019].

Wild, J. & Pissarek, M. (o.J.). *Ratte. Regensburger Analysetool für Texte*, available at https://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/ratte/index.html [Accessed 28 August 2019].