Z Erziehungswiss (2022) 25:947–973
https://doi.org/10.1007/s11618-022-01068-1

ZfE

Mobile First? Ein Vergleich von Lese-/Rechtschreibtests in traditionellem Papier-und-Bleistift-Format versus App-Format

Josefine Rothe · Linda Visser · Ruth Görgen · Julia Kalmar · Gerd Schulte-Körne · Marcus Hasselhorn

Eingegangen: 10. März 2020 / Überarbeitet: 27. August 2021 / Angenommen: 7. Oktober 2021 / Online publiziert: 22. Februar 2022

Zusammenfassung Digitale Medien haben mittlerweile einen festen Platz im Alltag von Schülerinnen und Schülern. Sie dienen nicht nur der Vermittlung von Lehrinhalten, sondern werden zunehmend auch für das Erbringen von Leistungsnachweisen und zur Diagnostik eingesetzt. Für die Durchführung und Auswertung etablierter Testverfahren zur Erfassung schulischer Leistungen wäre es wünschenswert, wenn sie in einem digitalen Format vorliegen. Aber sind diese Testleistungen vergleichbar? Um dies zu prüfen, wurden verbreitete Papier- und Bleistift-Testverfahren zur Erfassung der Lesegenauigkeit (Verlaufsdiagnostik sinnerfassenden Lesens, VSL), Dekodiergeschwindigkeit (Würzburger Leise Leseprobe – Revision, WLLP-R) und Rechtschreibung (Weingartener Grundwortschatz Rechtschreib-Test, WRT 3+/4+) digitalisiert und in eine kindgerechte Rahmengeschichte eingebettet. Insgesamt führten 237 Kinder der dritten und vierten Klassenstufe aus Bayern und Hessen die

J. Rothe und L. Visser teilen sich die Ersautorinnenschaft.

Dr. Josefine Rothe · Dr. Ruth Görgen · Dr. Julia Kalmar · Prof. Dr. Gerd Schulte-Körne
Klinik und Poliklinik für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie, LMU Klinikum, Nußbaumstraße 5 a, 80336 München, Deutschland

Dr. Josefine Rothe
E-Mail: Josefine.Rothe@uniklinikum-dresden.de

Dr. Ruth Görgen
E-Mail: Ruth.Goergen@mercator.uni-koeln.de

Dr. Julia Kalmar
E-Mail: Julia.Kalmar@zpp.uni-hd.de

Prof. Dr. Gerd Schulte-Körne
E-Mail: Gerd.Schulte-Koerne@med.uni-muenchen.de

Dr. Josefine Rothe
Klinik und Poliklinik für Kinder- und Jugendpsychiatrie und -psychotherapie, Universitätsklinikum Carl Gustav Carus Dresden, Fetscherstraße 74, 01307 Dresden, Deutschland

Testverfahren zuerst in digitaler Version (via App am Tablet oder Smartphone) und danach als Papier-und-Bleistift-Version durch. Es zeigt sich ein hoher Zusammenhang zwischen den Testleistungen der beiden Testformate, der unterschiedliche Interpretationen zulässt. Chancen und Grenzen digitalisierter Leistungstests werden diskutiert.

Schlüsselwörter Digitalisierung · Lesen · Rechtschreibung · Lese- und/oder Rechtschreibstörung

Mobile first? A comparison of traditional paper-pencil based reading and spelling tests with a mobile app version

Abstract Digital media have become an integral part of everyday life of schoolchildren. They are not only used for transferring teaching content, but also in the context of school performance records. For the administration of well-established school performance tests and the calculation of test scores it would be desirable if these are available in a digital format. However, are test results based on paper-pencil-format comparable to those of a digital test? To answer this question, we digitalized existing German standardized tests for assessing reading comprehension (Diagnosis of progress in reading comprehension, VSL), decoding speed (Wuerzburger Silent Reading Test–Revised, WLLP-R), and writing (Weingarten spelling test for basic vocabulary, WRT 3+/4+) and embedded them in a child-friendly cover story. A total of 237 children from 3rd and 4th grade from the German federal states Bavaria and Hesse made these tests first in a digital (via an app for tablet or smartphone) and then in a paper-pencil version. The results show a strong relationship between the performances in the two test formats, which can be interpreted in different ways. Strengths and limitations of digitalized school performance tests are discussed.

Keywords Digitization · Reading · Spelling · Reading and/or spelling disorder

Dr. Linda Visser · Prof. Dr. Marcus Hasselhorn
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Straße 6, 60323 Frankfurt am Main, Deutschland
E-Mail: Hasselhorn@dipf.de

Center for Research on Individual Development and Adaptive Education of Children at Risk (IDeA), Rostocker Straße 6, 60323 Frankfurt am Main, Deutschland
E-Mail: Linda.Visser@dipf.de

Dr. Ruth Görgen
Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache, Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Deutschland

Dr. Julia Kalmar
Psychologisches Institut, Zentrum für Psychologische Psychotherapie, Universität Heidelberg, Bergheimer Str. 58a, 69115 Heidelberg, Deutschland

1 Einleitung

Kinder nutzen digitale Medien in zunehmend vielen Lebensbereichen. Dies dürfte zukünftig auch für den Erwerb und die Nutzung grundlegender schulischer Kompetenzen wie das Lesen, Schreiben und Rechnen gelten. Digitale Testverfahren zur Erfassung des Leistungsstands und des Lernverlaufs dieser Kompetenzen werden in der Schule voraussichtlich weiter an Bedeutung gewinnen. Dies entspricht nicht nur der aktuellen Lebenswelt der Kinder, sondern erleichtert auch die Leistungsbestimmung alleine durch die Möglichkeit automatisierter Auswertung. Die besondere Dringlichkeit der Verfügbarkeit digitaler Medien für schulische Bildungsangebote wurde jüngst durch das Aussetzen des Präsenzunterrichts zur Eindämmung der Corona-Pandemie deutlich (Eickelmann und Gerick 2020).

Mit Inkrafttreten des „Digitalpakt Schule“ am 17. Mai 2019 wurde mit einem flächendeckenden Ausbau der digitalen Infrastruktur an Schulen begonnen. Die technischen Voraussetzungen für die Verwendung digitaler Testverfahren wären damit in absehbarer Zeit gegeben. Neben der digitalisierten Lebenswelt der Kinder sprechen auch die ökonomischen Vorteile digitaler Testverfahren für deren Einsatz in der Schule. Alleine die zeitliche Entlastung bei der Auswertung schafft den Lehrkräften mehr Freiräume für die Begleitung individueller Lernverläufe. Im Kontext des aktuellen Lehrkräftemangels (Klemm und Zorn 2019) sowie der wachsenden Heterogenität innerhalb der Klassen ist dies von unschätzbarem Wert.

Die Übertragung vorhandener analoger in digitale Testformate erfordert allerdings eine Untersuchung der sogenannten Testmoduseffekte. Von Testmoduseffekten spricht man, wenn es infolge eines Wechsels des Testformats zu Veränderungen psychometrischer Eigenschaften einer Messung kommt (vgl. Goldhammer et al. 2019). Auch wenn die Inhalte der beiden Testformate identisch sind, können Veränderungen aufgrund von Formaterfordernissen (z. B. wie die Reihenfolge oder Anordnung von Aufgaben und Items), die Handhabung (z. B. Texteingabe per Tastatur vs. handschriftlich oder Auswahl per drag and drop vs. Verbinden per Linie) oder die Erfahrungen im Umgang mit dem Testmedium sich unmittelbar auf die Testleistung auswirken.

Vorliegende Befunde zeigen, dass etwa das Eingabemedium (z. B. Tastatur, Maus, Touchpad, digitaler Stift; vgl. Cockburn et al. 2012; Connelly et al. 2007; Findlater et al. 2013; Gerth et al. 2016), die Anordnung der Testelemente (Shin et al. 2020; Steffener et al. 2020), sowie das Alter und die Erfahrung der Testperson im Umgang mit dem Eingabe- bzw. Testmedium die Bearbeitungsgeschwindigkeit und/oder das Leistungsergebnis beeinflussen können (vgl. Findlater et al. 2013; Schneider et al. 2008).

Die Vergleichbarkeit digital und analog erfasster Leseleistungen wurde in verschiedenen Studien untersucht (für Übersichten: Clinton 2019; Kong et al. 2018). Das *Lesetempo* scheint beim Lesen auf digitalen Medien im Vergleich zum Lesen auf Papier meist schneller zu sein (u. a. Elliott et al. 2019; Holzinger et al. 2011; Noyes und Garland 2003). Für das digital erfasste *Leseverständnis* werden hingegen überwiegend schlechtere Leistungen im Vergleich mit Papier und Bleistift erfassten Leseverständnis beschrieben (vgl. Clinton 2019; Delgado et al. 2018; Kong et al. 2018). Kong et al. (2018) berichten in ihrer Metaanalyse, dass die Differenz zwi-

schen digital und analog erfasstem Leseverständnis in den letzten Jahren immer geringer geworden ist. Dieser Effekt ist allerdings statistisch nicht abgesichert. In der Metaanalyse von Delgado et al. (2018) fällt der Effekt sogar umgekehrt aus. Mögliche Effekte der Unterschiede in der visuellen Darstellung (z. B. Bildschirmgröße) oder unterschiedlicher Eingabemedien (Tastatur, Maus, Touchscreen) wurden bisher in keiner Metaanalyse thematisiert.

Dass unterschiedliche Eingabemedien (Tastatur, Maus, Touchscreen) einen Einfluss auf die Antwortgeschwindigkeit haben, ist mittlerweile gut belegt. So konnte mehrfach nachgewiesen werden, dass das Auswählen von Objekten (z. B. Auswahl einer Antwortalternative) über ein Touchpad deutlich schneller erfolgt als mit einer Computermaus (u. a. Cockburn et al. 2012; Findlater et al. 2013; Schneider et al. 2008). Zumindest für *Speed*-Tests dürfte dies auch einen Effekt auf die Testleistung haben, was bei Testnormierungen zu berücksichtigen ist (Überprüfung der Parallelität bei unterschiedlichem Eingabemedium trotz gleichem Testmaterial). Findlater et al. (2013) haben auch gezeigt, dass ältere Personen (Altersspanne: 61–86 Jahre) generell langsamer im Umgang mit Computermaus und Touchpad sind als jüngere Personen (Altersspanne: 19–51 Jahre). Dieser Unterschied kann teilweise auf fehlende Routinen der älteren Personen im Umgang mit diesen Eingabemedien zurückgeführt werden und würde bei *Speed*-Tests zu einer systematischen Benachteiligung älterer Personen führen (fehlende Testfairness). Da jüngere Kinder in der Regel auch keine Routinen in der Bedienung einer Computermaus haben, könnte dieses Ergebnis möglicherweise auf jüngere Kinder übertragen werden. In der Studie von Findlater et al. (2013) wurde außerdem gezeigt, dass ältere Personen im Vergleich zu jüngeren von der Eingabe mit Touchpad mehr profitierten (35 % versus 16 % schnellere Eingabezeiten verglichen mit der Computermaus). Das Touchpad als Eingabemedium hat für *Speed*-Tests also im Vergleich mit der Computermaus den Vorteil, dass Routinen im Umgang mit dem Eingabemedium einen geringeren Einfluss auf die Testleistungen haben.

Untersuchungen zur Frage, ob die Verwendung einer Tastatur einen Einfluss auf die Anzahl oder Art der Rechtschreibfehler bei Kindern hat, sind uns nicht bekannt. Besonders Schreibanfänger profitieren jedoch von der handschriftlichen Bearbeitung von Aufgaben, indem Buchstaben und größere Schrifteinheiten selbst produziert werden (Berninger et al. 2015). Es lässt sich auch vermuten, dass die Eingabe mit einer Tastatur für Personen, die keine oder wenig Erfahrung in der Texteingabe per Tastatur haben, zusätzliche Fehlerquellen birgt. So ist für die Großschreibung beispielsweise die Bedienung der Umschalttaste notwendig, was eine zusätzliche Anforderung darstellt.

Neben den Eingabemedien müssen bei der Übertragung papierbasierter Tests in digitale Formate auch motivationale Aspekte berücksichtigt werden. So kann allein die Verwendung digitaler Geräte besonders bei Kindern motivational wirken (Irmert und Trutwin 2020). Werden die digitalen Testverfahren in eine spielerische Rahmengeschichte eingebettet und gar um ein Belohnungssystem erweitert, sind ebenfalls Auswirkungen auf die Testleistung zu erwarten. Positive Effekte wurden beispielsweise für exekutive Funktionen (z. B. Inhibition) bei Personen mit ADHS berichtet (Lauth-Lebens und Lauth 2020).

Effekte der Erfahrung im Umgang mit Computern auf die Leistungsdifferenz zwischen PP- und digitalen Testformaten konnte u. a. für Mathematiktests und die Textproduktion (Schreiben eines Aufsatzes) gezeigt werden (u. a. Bennett et al. 2008; Horkay et al. 2006). Schülerinnen und Schüler mit vergleichbaren Testleistungen in der papierbasierten Testung unterschieden sich in diesen Studien in ihren digitalen Testleistungen in Abhängigkeit zu ihrer Erfahrung im Umgang mit Computern. Auswertungen australischer Schulleistungsuntersuchungen (National Assessment Program in Literacy and Numeracy, NAPLAN) konnten außerdem Moduseffekte bei der Verwendung von Tablet und PC zeigen, die von der Vertrautheit mit dem Testgerät (sowohl bei Tablet als auch PC) abhängig sind (Davis et al. 2016), wobei der Moduseffekt für verschiedene Aufgabentypen (z. B. Antwortauswahl, Texteingabe, drag and drop) unterschiedlich stark ausfällt.

Während in den NEAP Studien (NEAP = National Assessment of Educational Progress; Bennett et al. 2008; Horkay et al. 2006) die Eingabegenauigkeit und Eingabegeschwindigkeit gemessen und zusätzlich die Computererfahrung mit einem Fragebogen erfasst wurde, verfügte die NAPLAN Studie (Davis et al. 2016) nur über die Information, welches Gerät im Unterricht zur Verfügung stand. In beiden Studien wurde damit nur ein Teilbereich digitaler Kompetenz erfasst, als deren Merkmale a) der Zugang zu digitalen Geräten, b) die Häufigkeit ihrer tatsächlichen Verwendung sowie c) entsprechendes Wissen, Fertigkeiten und Einstellungen gelten (Tumbas et al. 2019). Zusammenhänge der digitalen Kompetenz mit dem sozioökonomischem Status (Scherer und Siddiq 2019) und der sprachlichen Integration (Hatlevik et al. 2015) sind belegt und fallen durchschnittlich etwas höher bei Mädchen als bei Jungen aus (Siddiq und Scherer 2019). Die Befundlage spricht für die Notwendigkeit der Überprüfung von Moduseffekten.

Das digitale Angebot normierter Testverfahren zur Erfassung schulischer Leistungen ist bisher übersichtlich. Dies hat viele Gründe. So wurden etwa viele traditionelle Papier-Bleistift-Verfahren zur Leistungserfassung für eine Einzeltestsituation entwickelt (Lindberg et al. 2018), die im digitalen Format nicht einfach nachzubilden ist. Der Leseverständnistest für Erst- bis Siebtklässler – Version II (ELFE II, Lenhard et al. 2017) und die Verlaufsdiagnostik sinnerfassenden Lesens (VSL, Walter 2013) sind aktuell die einzigen uns bekannten deutschsprachigen Testverfahren zur Erfassung einer basalen schulischen Leistung (hier Leseverständnis), die sowohl im digitalen (hier PC mit Computermaus) als auch im analogen Format (Papier-und-Bleistift (PB)-Testung) vorliegen, und bei denen eine Äquivalenz der beiden Testformate nachgewiesen wurde sowie Gütekriterien und Normwerte für beide Testformate vorliegen.

Die Möglichkeit, einen Test sowohl im digitalen als auch PB-Format durchführen zu können, stellt für die diagnostische Praxis einen klaren Vorteil dar. Fachkräfte haben stets die Möglichkeit, zwischen den beiden Formaten zu wählen und flexibel auf den Bedarf der Kinder und die Situation zu reagieren und die in den unterschiedlichen Formaten erfassten Leistungen in Beziehung zueinander zu setzen. Für die VSL und für zwei Untertests des ELFE II konnte die Äquivalenz der beiden Testformate nachgewiesen werden, indem die erzielten Testleistungen in den beiden Testformaten verglichen und dabei keine bedeutsamen Unterschiede festgestellt wurden (Lenhard et al. 2017; Walter 2013). Für den Untertest „Wortverständnis“

des ELFE II zeigten sich Unterschiede zwischen den beiden Testformaten, sodass für diesen Untertest getrennte Normen für die zwei Testformate vorliegen. Lenhard et al. (2017) vermuten, dass Anforderungsunterschiede der beiden Testformate beim Untertest „Wortverständnis“ zu den Unterschieden geführt haben. So berichten sie, dass in der digitalen Testversion die Möglichkeit fehlt, schon bearbeitete Items nochmals zu bearbeiten. Weitere Unterschiede bestehen darin, dass das Umblättern entfällt oder die Markierung der richtigen Lösung eine andere Zeitdauer in Anspruch nimmt. Dies zeigt, dass trotz der Bemühungen, die digitale Testversion in ihrer Handhabung und Darstellung möglichst nah an der analogen Version umzusetzen, Unterschiede in der Darstellung und Bedienung unumgänglich sind, was sich auf die Testergebnisse auswirken kann. So ergibt sich die Frage, ob es für die Entwicklung digitaler Testformate ausreichend ist, vorhandene Tests in ein digitales Format zu überführen oder ob diese speziell für die digitale Verwendung konstruiert werden müssen (Geiger und Wilhelm 2018; Kuhn und Schwenk 2018; Schulte-Körne et al. 2018). Insbesondere stellt sich diese Frage für die bisher kaum untersuchte Übertragbarkeit papierbasierter Tests auf mobile digitale Geräte (Tablet und Smartphone), welche durch die variierende und teils geringe Bildschirmgröße, sowie die Eingabe per Touchscreen besondere Herausforderungen birgt.

Das Hauptziel der vorliegenden Studie war es, vorhandene papierbasierte Leistungstests zur Dekodiergeschwindigkeit, zum Leseverständnis und zur Rechtschreibung mit einem digitalen Format (App) für Tablets und Smartphones zu vergleichen. Im Einzelnen wurden folgende Fragestellungen und Hypothesen untersucht:

1. Werden im analogen und digitalen Format (papierbasiert und App) von Testverfahren zur Erfassung schriftsprachlicher Kompetenzen (Leseverständnis, Dekodiergeschwindigkeit, Rechtschreibung) im Grundschulalter vergleichbare Testleistungen erzielt? Für das Leseverständnis und die Dekodiergeschwindigkeit liegen bisher keine eindeutigen Evidenzen für Testmoduseffekte als Folge der digitalen Umsetzung vor. Bei der Übertragung der papierbasierten Tests in die mobile App wurde in der vorliegenden Studie darauf geachtet, so wenig Veränderungen wie möglich an der Darstellung der Aufgaben vorzunehmen. Die Auswahl von Items (Leseverständnis und Dekodiergeschwindigkeit) erfolgte über ein Touchpad. Für die Rechtschreibleistung konnten wir keine Untersuchungen zu der Frage finden, ob die Verwendung einer Tastatur einen Einfluss auf die Anzahl der Rechtschreibfehler bei Kindern hat. Im Einzelnen bestanden unsere Erwartungen darin, dass
   a. der Zusammenhang der Testleistungen in den beiden Testformaten hoch ist;
   b. es zu Unterschieden der Testleistungen durch Lerneffekte der Messwiederholung oder Entwicklungsfortschritte der Kinder kommt;
   c. mit beiden Testformaten die gleichen Kinder als solche mit unterdurchschnittlichen Lese-/Rechtschreibleistungen identifiziert werden; und
   d. der statistische Zusammenhang zwischen Testleistung und Schulnoten bei beiden Testformaten vergleichbar und moderat ist ($r > 0{,}30$).
2. Besteht Testfairness gegenüber der Vertrautheit mit digitalen Endgeräten? Bisherige Studien berichten übereinstimmend, dass Erfahrungen mit dem jeweiligen digitalen Testmedium einen positiven Effekt auf die digitalen Testleistungen haben.

Daher vermuten wir, dass Kinder mit einem eigenen Tablet und/oder Smartphone geringere Leistungsunterschiede zwischen den beiden Testformaten zeigen.

3. Zeigen sich in den unterschiedlichen Testformaten verschiedene Fehlerschwerpunkte? Ob es Fehlerschwerpunkte in der Rechtschreibleistung von Kindern als Folge des Verwendens einer Tastatur gibt, wurde bisher nicht überprüft. Allerdings vermuten wir, dass im digitalen Format des Rechtschreibtests Großschreibungen seltener richtig ausgeführt werden, da dies die zusätzliche Anforderung des Bedienens der Umschalttaste erfordert und dies insbesondere bei Kindern mit geringer Computererfahrung ungewohnt ist.

2 Methode

2.1 Datenerhebung

Die Datenerhebung erfolgte in zwei Schritten. Im ersten Schritt wurden 52.734 Familien in Hessen und Bayern per Brief eingeladen, an einer digitalen Testung per App teilzunehmen. Die Auswahl der Familien erfolgte über Einwohnermeldeämter (Bayern) bzw. das Kultusministerium (Hessen) mit Kontrolle des Alters und des Geschlechts (Bayern, Altersrange 8;8–10;8, Geschlecht zu gleichen Teilen verteilt) bzw. der Klassenstufe (Hessen, Dritt- und Viertklässler zu gleichen Teilen). Die Bearbeitung der digitalen Leistungstests erfolgte selbstständig von zu Hause am Tablet oder am Smartphone im Mai und Juni 2017. Dabei wurden die Kinder von der Kunstfigur „Meister Cody“ in einer kindgerechten Rahmengeschichte durch die App geleitet. Neben den Leistungstests wurden mit Fragebögen die Einschätzung der psychischen Belastungen der Kinder durch die Eltern als auch durch die Kinder selbst erfasst. Die Eltern füllten außerdem einen allgemeinen Anamnesefragebogen aus. Die digitale Testbearbeitung durch die Kinder wurde über vier Tage verteilt, mit einer Bearbeitungszeit von 30–45 min pro Tag.

Im zweiten Schritt wurde ein Teil der Familien zur Bearbeitung der Tests im analogen Format eingeladen. Somit haben alle teilnehmenden Kinder zuerst die Tests im digitalen Format durchgeführt und danach im analogen Format. Ein „counterbalanced Design“ war aus organisatorischen Gründen nicht realisierbar. In Hessen hatten alle teilnehmenden Familien die Möglichkeit, sich dazu eigeninitiativ nach Abschluss der digitalen Testungen anzumelden. In Bayern wurde eine selegierte Teilstichprobe eingeladen. Die analogen Testungen fanden zwischen Juni 2017 und Anfang Januar 2018 statt. Die Anzahl der Tage zwischen der digitalen und analogen Testung betrug durchschnittlich 101 ($SD = 62$; Bereich = 19–237 Tage). Die Studie wurde von den Ethikkommissionen der durchführenden Institutionen genehmigt. Alle Eltern und Kinder gaben eine schriftliche Einwilligung nach Aufklärung gemäß der Deklaration von Helsinki.

2.2 Stichprobe

Insgesamt haben 4542 Familien mit der Bearbeitung der App begonnen. Dies entspricht einer Rücklaufquote von 8,6 % und deckt sich mit den Rücklaufquoten vor-

Tab. 1 Verteilung von Bundesland, Klassenstufe, Geschlecht, und Alter in den Stichproben der digitalen und analogen Testungen

		Digitale Testung ($N = 3911$)	Analoge Testung ($N = 237$)
Bundesland	Bayern	3072 (78,5 %)	79 (33,3 %)
	Hessen	839 (21,5 %)	158 (66,7 %)
Klassenstufe	3. (4.*)	1812 (46,3 %)	69 (29,1 %)
	4. (5.*)	2099 (53,7 %)	168 (70,9 %)
Geschlecht	Mädchen	1868 (47,8 %)	100 (42,2 %)
	Jungen	2043 (52,2 %)	137 (57,8 %)
Alter (Jahre;Monate)	*M* (*SD*)	9;9 (0;7)	10;0 (0;7)
	Range	8;1–11;10	8;8–11;10
Tage zwischen den Testungen	*M* (*SD*)	–	101 (62)
	Range	–	19–237

hergehender Datenerhebungen der Institutionen (bei Einladung per Brief). 3911 Kinder bearbeiteten mindestens eines der relevanten digitalen Testverfahren. In der Stichprobe der digitalen Testung sind Geschlecht und Klassenstufe (3. vs. 4.) annähernd gleich verteilt (vgl. Tab. 1).

Für die Testungen im analogen Format wurden alle 177 Familien aus dem Stadtgebiet München, deren teilnehmendes Kind in mindestens einem der drei Leistungsbereiche (Lesen, Rechtschreiben, Mathematik) im digitalen Format unterdurchschnittliche Leistungen aufwiesen ($T \leq 40$), angeschrieben. Außerdem wurden 112 Kinder mit unauffälligen Leistungen aus dem Stadtgebiet München eingeladen (gewichtet nach Klassenstufe und Geschlecht), was einer Quote von 20 % der leistungsunauffälligen Kinder aus dem Stadtgebiet München entspricht. Weiterhin konnten alle Familien aus Hessen teilnehmen, die sich eigeninitiativ zur analogen Testung angemeldet hatten.

Insgesamt nahmen an der analogen Testung 237 Kinder teil. In dieser Teilstichprobe gibt es eine Überrepräsentation von Jungen, Kindern der vierten Klassenstufe und Kindern aus Hessen (siehe Tab. 1). Die Eltern von 115 der 237 Kinder (48,5 %) gaben an, dass ihr Kind im Besitz eines eigenen Handys oder Tablets ist. Dies deckt sich in etwa mit den Angaben der Kindheit, Internet, Medien-Studie (KIM-Studie; Medienpädagogischer Forschungsverbund Südwest 2018), wonach in Deutschland etwas über 50 % der 8–11-Jährigen über ein eigenes Smartphone verfügen.

Von den teilnehmenden Geschwisterpaaren, die im Rahmen der Zufallsziehung zur Studienteilnahme eingeladen wurden, ist jeweils ein Geschwisterkind randomisiert ausgeschlossen worden ($N = 47$ für die digitale Testung; $N = 4$ für die analoge Testung).

2.3 Material

Zur Erfassung der Testleistungen im digitalen Format wurde eine App entwickelt. Vor der Datenerhebung wurde eine erste Version der App in Bayern und Hessen mit

25 Kindern sowie 7 Eltern pilotiert. Aufgrund der Beobachtungen und Rückmeldungen wurde die App für die eigentliche Datenerhebung optimiert.

2.3.1 Leseverständnis

Zur Erfassung des Leseverständnisses wurde die „Verlaufsdiagnostik sinnerfassenden Lesens“ (VSL; Walter 2013; Paralleltestreliabilität r = 0,78–0,80 für die 3. und 4. Klassenstufe) verwendet. Die Kinder lesen bei diesem Test einen Text, bei dem sich an der Stelle jedes siebten Wortes eine Klammer mit drei Auswahlwörtern (zwei Distraktoren und das Zielwort) befindet. Das Zielwort ist einzukreisen. Zur Abbildung des Lernverlaufes über einen längeren Zeitraum stehen 20 parallele Lesehefte mit jeweils 2 Geschichten zur Verfügung. Für die aktuelle Studie wurden zwei der Texte (*Die Elefanten* und *Die Feldhasen*) verwendet.

Digitale Testversion Bei der visuellen Darstellung der Aufgabe musste die Textmenge an die Bildschirmgrößen von Tablets und Smartphones angepasst werden. Dadurch war es nicht möglich, den gesamten Text auf dem Bildschirm darzustellen. Um sicherzustellen, dass Textzusammenhänge dennoch uneingeschränkt erfasst und selbstständig überprüft und korrigiert werden konnten, gab es die Möglichkeit, mit Hilfe von Pfeiltasten zwischen den Seiten zu wechseln (Abb. 1). Die Antwortauswahl erfolgte durch Antippen des entsprechenden Wortes. Das ausgewählte Wort wurde dann farbig unterlegt. Die Korrektur eines Items erfolgte durch erneutes Antippen eines Wortes. Diese Funktionen wurden vor Beginn der Aufgabe erklärt und von den Kindern getestet.

Durch einen technischen Fehler wurden im digitalen Testformat lediglich die Antworten von 46 der 50 Items der analogen PB-Version abgespeichert. Durchschnittlich bearbeiteten die Kinder innerhalb der verfügbaren vier Minuten 27 Items (Standardabweichung = 8; Spanne = 4–46).

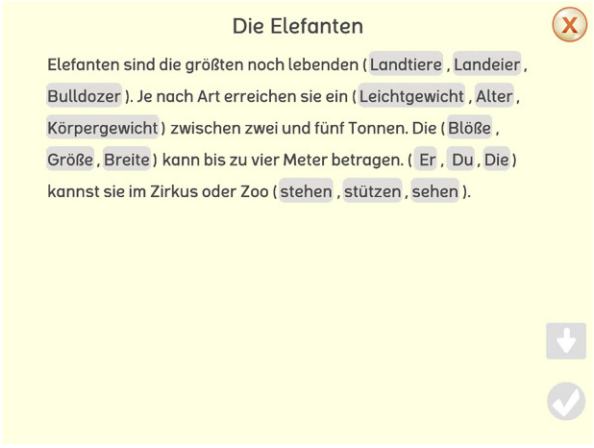

Abb. 1 Digitale Adaptation der VSL

Abb. 2 Digitale Adaptation der WLLP-R

2.3.2 Dekodiergeschwindigkeit

Die Dekodiergeschwindigkeit wurde mit der „Würzburger Leise Leseprobe – Revision“ erfasst (WLLP-R; Schneider et al. 2011; Paralleltestreliabilität r = 0,93 für die 3. Klassenstufe und r = 0,82 für die 4. Klassenstufe). Den Kindern werden in diesem Test eine Reihe von geschriebenen Wörtern präsentiert. Jedem Wort werden jeweils vier Bilder gegenübergestellt. Die Kinder sollen zu jedem Wort das passende Bild aus den vier Optionen anstreichen. Eine Testzeit von fünf Minuten ist vorgegeben.

Digitale Testversion Für die digitale Darstellung musste die Anzahl der präsentierten Items an die Bildschirmgrößen von Tablets und Smartphones angepasst werden. Dadurch konnte stets nur eine Aufgabe auf dem Bildschirm angezeigt werden (Abb. 2). Die Antwortauswahl erfolgte durch Antippen des entsprechenden Bildes. Nach der Auswahl eines Items wurde direkt die nächste Aufgabe angezeigt. Aufgrund der Vielzahl an Items (180) wurde auf die Möglichkeit eines Wechsels zu den vorherigen Items, wie in der VSL, verzichtet. Folglich konnten einmal ausgewählte Items nicht korrigiert werden. Durchschnittlich bearbeiteten die Kinder innerhalb der verfügbaren fünf Minuten 105 Items (Standardabweichung = 18; Spanne = 9–178).

2.3.3 Rechtschreibleistung

Die Rechtschreibleistung wurde mit den Langfassungen des „Weingartener Grundwortschatz Rechtschreib-Test“ erfasst (WRT 3+ für Drittklässler; Birkel 2007a; Paralleltestreliabilität r > 0,91 und WRT 4+ für Viertklässler; Birkel 2007b; Paralleltestreliabilität r > 0,90). Der Test besteht aus 8 (WRT 3+) bzw. 7 (WRT 4+) Texten mit insgesamt 55 (WRT 3+) bzw. 60 (WRT 4+) Wortlücken. Den Kindern wird von der Testleitung zuerst der jeweilige gesamte Text, danach jeder einzelne Satz sowie jedes fehlende Wort einzeln vorgesprochen. Die Kinder schreiben das jeweils fehlende Wort in die entsprechende Lücke. Eine zeitliche Begrenzung ist für diesen Test nicht vorgesehen; die Bearbeitungszeit beträgt in der Regel maximal 45 min.

Abb. 3 Digitale Adaptation des WRT

Digitale Testversion Die Anzahl der dargestellten Items musste an die Bildschirmgrößen von Tablets und Smartphones angepasst werden. Hierbei wurde jeweils nur ein Item präsentiert. Die Verschriftlichung der Zielitems erfolgte mit einer Tastatur. Bei der Symbolauswahl für die Tasten „löschen", „bestätigen" und „Groß-/Kleinschreibung" wurde darauf geachtet, Symbole zu verwenden die auch für Kinder ohne Erfahrungen im Schreiben mit einer Tastatur leicht verständlich sind (Abb. 3). War für die Tastatur die Kleinschreibung aktiviert, wurden die Buchstaben der Tastatur als Kleinbuchstaben dargestellt. Gleiches gilt für die Großschreibung. Alle Funktionstasten der Tastatur wurden vor Beginn der Aufgabe erklärt und von den Kindern getestet. Der Wechsel zu vorherigen Items war nicht möglich. Die Bearbeitungszeit in der App war durchschnittlich 33 min für den WRT 3+ und 42 min für den WRT 4+.

2.3.4 Eigenes Tablet oder Smartphone (ToS)

Die Vertrautheit der Kinder mit digitalen Endgeräten wurde über die Antwort der Eltern auf die Fragen „Besitzt ihr Kind ein eigenes Smartphone?" und „Besitzt ihr Kind ein eigenes Tablet?" operationalisiert.

2.4 Statistische Verfahren

Da die Kinder während der digitalen Testung nicht beobachtet werden konnten, wurden Plausibilitätsprüfungen durchgeführt. Daten wurden als „unplausibel" gewertet, wenn 1.) die tatsächliche Testzeit von der im Manual vorgegebenen Testzeit abwich (d. h. wenn die Testzeit beispielsweise aufgrund eines Bugs zu kurz/lang war, z. B. WLLP-R: –1 Sek./+4 Sek. der im Manual vorgegebenen Testzeit); 2.) die gleiche Antwort zu oft in Folge gegeben wurde (gilt nur für die WLLP-R, wenn mehr als 10 Mal in Folge die gleiche Antwortalternative gewählt wurde); 3.) die Reaktionszeit unrealistisch schnell war (>3 × mittlere absolute Abweichung vom Median), und 4.) im Lückendiktat willkürliche Zeichenfolgen getippt wurden (z. B. skjdcn). Dazu

wurde für jede gegebene Antwort die Ähnlichkeit zum Zielitem durch Berechnung der Jaro-Winkler-Distanz bestimmt, bei der die Ähnlichkeit zweier Zeichenketten über die Anzahl und Reihenfolge gemeinsamer Zeichen bestimmt wird (Jaro 1989; Winkler 1990). Es können Werte von 0 bis 1 erzielt werden, wobei der Wert 0 für eine genaue Übereinstimmung der Zeichenketten steht und der Wert 1 für eine maximale Distanz. Antworten mit einem Jaro-Winkler-Abstand von mehr als 0,41 (Ausnahme für das Wort ‚Axt' = 0,49) wurden als zufällige Eingabe betrachtet. Testergebnisse von Kindern, die bei mehr als 15 % der Items willkürlich tippten, wurden als unplausibel gekennzeichnet und ausgeschlossen. Außerdem wurden für die digitalen Versionen der Testverfahren neue Normen entwickelt (jeweils separat für die 3. und 4. Klassenstufe). Der Normierungszeitraum entspricht dabei dem Testzeitraum, Anfang Mai bis Ende Juni. Weitere Informationen zur Plausibilitätsprüfung sowie zu den Normen der digitalen Version können Visser et al. (2020) entnommen werden.

In den digitalen und analogen Testungen wurden jeweils identische Testformen verwendet (keine Parallel-Testformen). Die T-Werte der analog erfassten Leistungen wurden mit Hilfe der Manual-Normen ermittelt. Für den VSL wurden die Normtabellen für die Mitte der 3. bzw. 4. Klassenstufe verwendet; für den WRT und WLLP-R die Normtabellen für das Ende der 3. bzw. 4. Klassenstufe. Im Sinne der Vergleichbarkeit der T-Werte innerhalb der Stichprobe wurden für die analogen Testungen die gleichen Normierungszeiträume gewählt. Dadurch wurden für einige Kinder, die sich schon am Anfang der nächsten Klassenstufe befanden (Anfang der 4. bzw. 5. Klasse), nicht die individuell passenden Normen verwendet, was Betrachtungen auf individueller Ebene erschwert. Für die Prüfung der formulierten Hypothesen ist diese Vorgehensweise jedoch geeignet. Für Rückmeldungen der Testergebnisse an die Eltern wurden die individuell passenden Normen verwendet.

Da das Zeitintervall zwischen der digitalen und analogen Testung sehr variierte, wurde zunächst regressionsanalytisch geprüft, inwiefern durch den Rohwert der digitalen Testung, das Alter, das Zeitintervall und die Auffälligkeit der digitalen Testleistung (UV) der Rohwert der Leistung bei der analogen Testung (AV) vorhergesagt wird. Dies erlaubte eine Prüfung der plausiblen Annahme, dass das Zeitintervall zwischen den Testungen einen Einfluss auf den Rohwert bei der analogen Testung hat. Außerdem wurde in dieser Regression geprüft, ob die Unterteilung in mindestens durchschnittliche ($T > 40$) und unterdurchschnittliche ($T \leq 40$) Testleistung im digitalen Format zu Unterschieden im Rohwert bei der analogen Testung geführt hat, unter Kontrolle des digitalen Rohwerts (als UV). Sollte sich nämlich hier kein Unterschied zwischen den Kindern mit auffälligem und unauffälligem digitalen Testergebnis zeigen, kann bei den anschließenden Hypothesenprüfungen auf eine Unterscheidung zwischen diesen Gruppen verzichtet werden. Bei den Regressionsanalysen wurde über die *false discovery rate procedure* (FDR) nach Benjamini und Yekutieli (2001) für mehrfaches Testen korrigiert.

Für jedes Testverfahren wurden die deskriptiven Statistiken der Rohwerte (und für den WRT 3+/4+ die Itemschwierigkeiten) für die digitale und die analoge Testung ermittelt. Für die Substichprobe der Kinder, die auch an der analogen Testung teilnahmen, wurden die Leistungen bei digitaler und analoger Testung (Rohwerte) mit Hilfe einer Varianzanalyse mit Messwiederholung auf Unterschiedlichkeit geprüft (Hypothese 1b) sowie deren statistischer Zusammenhang berechnet (Hypothese 1a).

Anhand der Rohwertdifferenzen zwischen den Klassennormen der Testmanuale können grobe Richtwerte abgeleitet werden, welche Rohwertdifferenzen aufgrund von Entwicklungseffekten zu erwarten sind. Dies wurde für alle Tests (VSL, WLLP-R, WRT 3+ und WRT 4+) durchgeführt und kann im Onlinematerial 1 eingesehen werden.

Auch wenn sich die nicht transformierten Rohwerte am besten für einen direkten Vergleich zwischen den Testergebnissen eignen, kann mit Hilfe der Normwerte ein Bezug zu einer Grundgesamtheit hergestellt werden und somit der individuelle Leistungsstand innerhalb einer Bezugsgruppe interpretiert werden. Aus diesem Grund wurden auch für die Normwerte deskriptive Statistiken erstellt und geprüft, ob unter beiden Testformaten dieselben Kinder (d. h. Kinder mit einer Minderleistung in dem jeweiligen Bereich, die einem T-Wert ≤40 entspricht), identifiziert werden. Basierend auf den Leistungen bei der analogen Testung wurden auch die Trefferquote und der Ratz-Index (Relativer Anstieg der Trefferquote gegenüber der Zufallstrefferquote) berechnet (Marx und Lenhard 2010). Dies erlaubt eine Abschätzung, wie gut die Vorhersage der Minderleistungen im analogen Testformat durch die digitalen Testergebnisse gelingt. Ab einem RATZ-Index von 66 % wird eine sehr gute Vorhersage attestiert (Jansen et al. 1999). Für die Leistungen bei digitaler Testung wurden schließlich die Korrelationen mit den berichteten Schulnoten im Fach Deutsch berechnet. Zusätzlich wurden die Ergebnisse von Kindern mit und ohne eigenem ToS mit Hilfe einer Varianzanalyse mit Messwiederholung verglichen.

3 Ergebnisse

Für alle berichteten Testverfahren kann von einer Normalverteilung der Daten ausgegangen werden (Abb. E1–E12 im Onlinematerial 2). Die deskriptiven Statistiken für die Testverfahren sind den Tab. 2 (auf Rohwertbasis) und Tab. 3 (auf T-Wertbasis) zu entnehmen.

Tab. 2 Stichprobengrößen (N), Mittelwerte (M) und Standardabweichungen (SD) sowie Wertebereiche (Range) der Testleistungen für beide Testformate auf Basis der Rohwerte

	Digitale Testung			Digitale Testergebnisse der Substichprobe der analogen Testung			Analoge Testung[a]		
	N	*M* (*SD*)	Range	*N*	*M* (*SD*)	Range	*N*	*M* (*SD*)	Range
VSL	3814	24,6 (8,1)	2–44	236	24,2 (8,4)	7–44	236	25,7 (8,3)	4–46
WLLP-R	3787	98,3 (17,0)	7–166	233	96,9 (18,6)	50–139	233	104,7 (24,3)	50–177
WRT 3+	1782	36,5 (11,3)	1–55	68	31,5 (12,2)	9–53	68	36,0 (10,9)	6–53
WRT 4+	2077	40,7 (13,1)	2–60	165	37,7 (13,9)	2–60	165	40,5 (12,7)	2–60

[a]Auf Basis der gleichen 46 Items der VSL wie in der digitalen Version

Tab. 3 Deskriptive Statistik der Testleistungen für die beiden Testformate auf Basis der T-Normwerte

	Digitale Testung			Digitale Testergebnisse der Substichprobe der analogen Testung			Analoge Testung		
	N	*M* (*SD*)	Range	*N*	*M* (*SD*)	Range	*N*	*M* (*SD*)	Range
VSL	3814	50,0 (9,6)	29–71	236	48,2 (9,8)	29–71	236	61,5 (9,8)	30–75
WLLP-R	3787	50,1 (9,7)	29–71	233	48,3 (10,4)	29–71	233	51,7 (11,2)	28–81
WRT 3+	1782	50,0 (9,6)	29–71	68	46,0 (10,3)	29–69	68	52,6 (9,6)	29–73
WRT 4+	2077	50,1 (9,7)	29–71	165	47,9 (9,8)	29–71	165	52,1 (10,0)	22–78

3.1 VSL

Von den 3911 Kindern der Stichprobe haben 97 Kinder die digitale Testung der VSL nicht oder nicht vollständig absolviert. Zusätzlich wurden für 4 Kinder die Daten als unplausibel kategorisiert, womit die Analysestichprobe 3814 Kinder umfasste. Von diesen Kindern zeigten 3153 (82,7 %) mindestens durchschnittliche (T-Wert >40) und 661 (17,3 %) unterdurchschnittliche Leistungen in der VSL.

Die Ergebnisse der Regressionsanalysen bestätigen für die VSL, dass das Zeitintervall zwischen den Testungen vernachlässigt werden kann sowie die Zuordnung „mindestens durchschnittliche“ und „unterdurchschnittliche“ Testleistungen über beide Testformate vergleichbar ausfällt. In Tab. 4 sind die Ergebnisse der Regressionsanalyse dargestellt. Von den Kindern, die beide Testformate bearbeitet ha-

Tab. 4 Ergebnisse der Regressionsanalysen zur Vorhersage des Rohwerts der analogen Testung

Test	Effekt	B	SE	*p*
VSL	Intercept	10,562	6,227	0,091
	Rohwert der digitalen Testung	0,860	0,053	0,000*
	Alter	–0,053	0,051	0,297
	Auffälligkeit der digitalen Testleistung	0,541	0,978	0,581
	Zeitintervall	0,006	0,006	0,310
WLLP-R	Intercept	–15,422	19,412	0,428
	Rohwert der digitalen Testung	1,001	0,085	0,000*
	Alter	0,216	0,161	0,181
	Auffälligkeit der digitalen Testleistung	0,737	3,652	0,840
	Zeitintervall	–0,028	0,017	0,103
WRT 3+	Intercept	50,106	22,736	0,031
	Rohwert der digitalen Testung	0,488	0,115	0,000*
	Alter	–0,267	0,196	0,178
	Auffälligkeit der digitalen Testleistung	–7,205	3,013	0,020
	Zeitintervall	0,019	0,014	0,160
WRT 4+	Intercept	7,859	14,562	0,590
	Rohwert der digitalen Testung	0,779	0,064	0,000*
	Alter	0,014	0,113	0,902
	Auffälligkeit der digitalen Testleistung	–0,986	2,047	0,631
	Zeitintervall	0,021	0,010	0,035

*Signifikant nach FDR-Korrektur

ben, wurden die Ergebnisse der digitalen Testung von einem Kind als unplausibel kategorisiert und von der Auswertung der Ergebnisse bei der analogen Testung ausgeschlossen. Die Substichprobe besteht somit aus 236 Kindern, wovon 177 (75,0 %) eine mindestens durchschnittliche und 59 (25,0 %) eine unterdurchschnittliche Leistung in der digitalen Testung gezeigt haben.

Für die beiden Testversionen der VSL ergibt sich ein hoher Zusammenhang der Testrohwerte ($r = 0{,}82$, $p < 0{,}01$) und T-Werte ($r = 0{,}76$, $p < 0{,}01$). Die Rohwerte im analogen Testformat fallen jedoch signifikant höher als im digitalen Testformat aus ($F(1{,}235) = 21{,}9$, $p < 0{,}01$, partielles $\eta^2 = 0{,}09$). Die Rohwertdifferenz (RD) der beiden Testformate fällt im Vergleich zu der aufgrund von Entwicklungseffekten zu erwartenden Rohwertdifferenz sowohl für die 3. Klasse als auch die 4. Klasse etwas niedriger aus (beobachtet: 3. Klasse RD = 2,6, 4. Klasse RD = 1,1; erwartet: 3. Klasse RD = 4,8, 4. Klasse RD = 3,2; Details siehe Onlinematerial 1). Ein signifikanter Unterschied, mit höheren Werten in der analogen Testung, findet sich auch für die T-Werte ($F(1{,}235) = 8409{,}1$, $p < 0{,}01$, partielles $\eta^2 = 0{,}79$).

Von den fünf Kindern, die im analogen Testformat T-Werte ≤40 erzielten, zeigten drei auch in der digitalen Testversion T-Werte ≤40 (siehe Tab. 4). In der analogen Testung wurden 231 Kinder mit T-Werten >40 identifiziert, von denen 175 Kinder auch in der digitalen Testversion T-Werte >40 erzielten. Dies entspricht, gemessen an den Testleistungen der analogen Testung, einer Trefferquote von 75,4 %. Der RATZ-Index beträgt 46,7 %.

Für die digitale Version der VSL zeigt sich ein mittlerer Zusammenhang zur Deutschnote (3. Klasse: $N = 1769$, $r = -0{,}389$, $p < 0{,}01$; 4. Klasse: $N = 2045$, $r = -0{,}421$, $p < 0{,}01$). Ein ebenfalls mittlerer Zusammenhang besteht zwischen den Testleistungen der analogen Version und den Deutschnoten (3. Klasse: $N = 67$, $r = -0{,}57$, $p < 0{,}01$; 4. Klasse: $N = 165$, $r = -0{,}51$, $p < 0{,}01$).

Kinder mit und ohne eigenes ToS unterscheiden sich bedeutsam in ihrer Rohwertdifferenz zwischen den beiden Testversionen mit einer höheren Differenz für Kinder ohne eigenes ToS (vgl. Tab. 5). Dabei zeigen Kinder ohne eigenes ToS in der analogen Testung deutlich höhere Testwerte als in der digitalen Testung. Bei Kindern mit eigenem ToS sind die Rohwerte der beiden Versionen eher vergleichbar. Die Varianzanalyse mit Messwiederholung erbrachte einen statistisch bedeutsamen Interaktionseffekt zwischen Testzeitpunkt und ToS ($F(1{,}234) = 10{,}0$, $p < 0{,}01$, partielles $\eta^2 = 0{,}04$) (siehe auch Abb. E13).

Tab. 5 Unterschiede in den Rohwerten zwischen Kindern mit eigenem Tablet/Smartphone und ohne eigenes Tablet/Smartphone

	Eigenes Tablet oder Smartphone				Kein eigenes Tablet oder Smartphone			
	N	Digital *M* (*SD*)	Analog *M* (*SD*)	Differenz *M* (*SD*)	*N*	Digital *M* (*SD*)	Analog *M* (*SD*)	Differenz *M* (*SD*)
VSL	115	24,5 (8,4)	24,9 (8,6)	0,5 (5,5)	121	23,9 (8,3)	26,5 (7,9)	2,5 (4,4)
WLLP-R	114	97,3 (18,7)	105,2 (25,3)	7,8 (16,0)	119	96,6 (18,6)	104,3 (23,5)	7,8 (13,6)
WRT 3+	26	30,3 (13,4)	33,0 (12,1)	2,8 (9,8)	42	32,3 (11,5)	37,8 (9,8)	5,5 (7,3)
WRT 4+	87	37,3 (14,5)	39,8 (12,6)	2,5 (7,9)	78	38,1 (13,2)	41,2 (12,9)	3,1 (6,0)

3.2 WLLP-R

Von den 3911 Kindern der Stichprobe haben 124 Kinder die digitale Testung der WLLP-R nicht oder nicht vollständig absolviert (eigeninitiierter oder technischer Abbruch der App vor oder während der Testung). Zusätzlich wurden für 43 Kinder die Daten als unplausibel kategorisiert. Die Datenanalyse der digitalen WLLP-R basiert damit auf den Testergebnissen von 3787 Kindern. Davon zeigten 3146 Kinder (83,1 %) Leistungen im mindestens durchschnittlichen Bereich (T-Wert >40), während 641 Kinder (16,9 %) Leistungen im unterdurchschnittlichen Bereich zeigten.

Das Zeitintervall zwischen den Testungen erwies sich in der Regressionsanalyse bei der WLLP-R als vernachlässigbar. Die Ergebnisse der Regressionsanalyse (siehe Tab. 4) bestätigen außerdem, dass die Zuordnung „mindestens durchschnittliche" und „unterdurchschnittliche" Testleistungen über beide Testformate der WLLP-R vergleichbar ausfällt.

In der analogen Testung der WLLP-R waren die Daten von 2 Kindern unplausibel (beurteilt auf Basis von Beobachtungen der Testleitung). Von zwei weiteren Kindern fehlten die Daten. Somit besteht die Analysestichprobe aus 233 Kindern. Davon haben in der App 179 Kinder (76,8 %) Leistungen im mindestens durchschnittlichen Bereich gezeigt und 54 Kinder (23,2 %) Leistungen im unterdurchschnittlichen Bereich.

Die Testergebnisse der WLLP-R zeigen einen hohen statistischen Zusammenhang der Rohwerte ($r = 0{,}80$, $p < 0{,}01$) und T-Werte ($r = 0{,}76$, $p < 0{,}01$) der beiden Testformate und sind vergleichbar mit den im Manual berichteten Retest-Reliabilitäten (3. Klasse: r = 0,82; 4. Klasse: r = 0,80). Die Mittelwerte und Standardabweichungen für die Rohwerte und T-Werte der beiden Testversionen sind den Tab. 2 und 3 zu entnehmen. Die Rohwerte bei analoger Testung übertreffen die der digitalen Version ($F(1{,}232) = 64{,}8$, $p < 0{,}01$, partielles $\eta^2 = 0{,}22$). Die Rohwertdifferenz der beiden Testformate entspricht für die 3. Klasse dem, was aufgrund von Entwicklungseffekten zu erwarten ist. Für die 4. Klasse liegt die Rohwertdifferenz deutlich über dem, was auf Basis der Normwerte des Testmanuals zu erwarten wäre, wobei hier methodische Einschränkungen zu beachten sind (beobachtet: 3. Klasse RD = 5,1, 4. Klasse RD = 8,9; erwartet: 3. Klasse RD = 4,9, 4. Klasse RD = 2,4; Details siehe Onlinematerial 1).

Die T-Werte der beiden Testversionen unterscheiden sich signifikant voneinander ($F(1{,}232) = 48{,}2$, $p < 0{,}01$, partielles $\eta^2 = 0{,}17$). Außerdem zeigen die Fehlerraten (Anzahl der Fehler in Relation zur Anzahl der Items) der beiden Testversionen, dass in der analogen Testung bedeutend weniger Fehler gemacht wurden als in der digitalen Testung (analog: 0,039; digital: 0,061; $F(1{,}232) = 41{,}8$, $p < 0{,}01$, partielles $\eta^2 = 0{,}15$).

Wie in Tab. 6 dargestellt, zeigten von den 42 Kindern, die in der analogen Version T-Werte ≤40 erzielten, 31 Kinder auch in der digitalen Testversion T-Werte ≤40. Umgekehrt wurden in der analogen Testung 191 Kinder mit einem T-Wert >40 identifiziert, von denen 168 Kinder auch in der digitalen Testung T-Werte >40 erzielten. Die Trefferquote der Minderleistungen bei analoger Testung durch die Leistungen bei digitaler Testung beträgt 85,4 %. Der RATZ-Index liegt bei 65,90 %.

Tab. 6 Übereinstimmung der Identifikation durchschnittlicher und unterdurchschnittlicher Testleistungen

			Analog		
			$T \leq 40$	$T > 40$	Total
Digital	VSL	$T \leq 40$	3 (5,1 %)	56 (94,9 %)	59 (100 %)
		$T > 40$	2 (1,1 %)	175 (98,9 %)	177 (100 %)
		Total	5	231	236
	WLLP-R	$T \leq 40$	31 (57,4 %)	23 (42,6 %)	54 (100 %)
		$T > 40$	11 (6,1 %)	168 (93,9 %)	179 (100 %)
		Total	42	191	233
	WRT 3+	$T \leq 40$	3 (14,3 %)	18 (85,7 %)	21 (100 %)
		$T > 40$	0 (0,0 %)	47 (100 %)	47 (100 %)
		Total	3	65	68
	WRT 4+	$T \leq 40$	21 (50,0 %)	21 (50,0 %)	42 (100 %)
		$T > 40$	1 (0,8 %)	122 (99,2 %)	123 (100 %)
		Total	22	143	165

Zwischen den Testleistungen der digitalen Version der WLLP-R und den Deutschnoten besteht ein mittlerer Zusammenhang (3. Klasse: $N = 1653$, $r = –0{,}390$, $p < 0{,}01$; 4. Klasse: $N = 1916$, $r = –0{,}383$, $p < 0{,}01$). Ein ebenfalls mittlerer Zusammenhang besteht zwischen den Testleistungen der analogen Version und den Deutschnoten (3. Klasse: $N = 67$, $r = –0{,}53$, $p < 0{,}01$; 4. Klasse: $N = 163$, $r = –0{,}42$, $p < 0{,}01$).

Die Rohwertdifferenzen zwischen der digitalen und der analogen Testung von Kindern mit eigenem ToS unterscheiden sich nicht bedeutsam von den Leistungsdifferenzen der Kinder ohne eigenes ToS (Tab. 5). In der entsprechenden Varianzanalyse mit Messwiederholung finden sich keinerlei signifikante Effekte.

3.3 WRT 3+ und WRT 4+

Für 54 Kinder wurden die WRT-Daten als unplausibel kategorisiert. Die im Folgenden berichteten Ergebnisse der digitalen Testleistungen beim WRT 3+ und WRT 4+ basieren somit auf einer Stichprobe von 3859 Kindern (3. Klasse: $N = 1782$, 4. Klasse: $N = 2077$). Davon zeigten 3198 Kinder (82,9 %; 3. Klasse: 1478, 82,9 %; 4. Klasse: 1720, 82,8 %) Leistungen im mindestens durchschnittlichen (T-Wert >40) und 661 Kinder (17,1 %) im unterdurchschnittlichen Bereich.

Für den WRT 3+ und WRT 4+ bestätigen die Ergebnisse der Regressionsanalysen, dass das Zeitintervall zwischen den Testungen vernachlässigt werden kann. Außerdem bestätigen die Ergebnisse, dass die Zuordnung „mindestens durchschnittliche“ und „unterdurchschnittliche“ Testleistungen über beide Testformate vergleichbar ausfällt, siehe Tab. 4.

Von den Kindern, die beide Versionen bearbeitet haben, wurden die Daten der digitalen Testung von einem Kind als unplausibel kategorisiert und von der Auswertung ausgeschlossen. Für weitere drei Kinder fehlten die Daten der analogen Testung des WRT. Die Analysestichprobe besteht somit aus 233 Kindern, von denen 170

(73,0 %) eine mindestens durchschnittliche und 63 (27,0 %) eine unterdurchschnittliche Leistung in der digitalen Testung gezeigt hatten.

Die Mittelwerte und Standardabweichungen für die Rohwerte und T-Werte beider Testversionen sind den Tab. 2 und 3 zu entnehmen. Zwischen den Testrohwerten der beiden Testversionen ist der statistische Zusammenhang jeweils bedeutsam (WRT 3+: $r = 0{,}74$, $p < 0{,}01$; WRT 4+: $r = 0{,}86$, $p < 0{,}01$), liegen allerdings etwas unter den im Manual berichteten Retest-Reliabilitäten (3. Klasse: r = 0,93; 4. Klasse: r = 0,94). Die Korrelationen der T-Werte sind vergleichbar (WRT 3+: $r = 0{,}73$, $p < 0{,}01$; WRT 4+: $r = 0{,}83$, $p < 0{,}01$). Mit höheren Werten bei der analogen Testung unterscheiden sich die Testrohwerte jedoch bedeutsam voneinander (WRT 3+: $F(1{,}67) = 19{,}3$, $p < 0{,}01$, partielles $\eta^2 = 0{,}22$; WRT 4+: $F(1{,}164) = 26{,}3$, $p < 0{,}01$, partielles $\eta^2 = 0{,}14$). Für die 3. Klasse liegt die Rohwertdifferenz der beiden Testformate etwas höher, als es auf Basis der Normwerte des Testmanuals zu erwarten wäre. Für die 4. Klasse ist dieser Vergleich aufgrund unzureichender Normwerte für die 5 Klasse nicht möglich gewesen (beobachtet: 3. Klasse RD = 4,5; erwartet: 3. Klasse RD = 2,2; Details siehe Onlinematerial 1).

Die T-Werte der analogen Testung liegen höher als die der digitalen Testung (WRT 3+: $F(1{,}67) = 53{,}5$, $p < 0{,}01$, partielles $\eta^2 = 0{,}44$; WRT 4+: $F(1{,}164) = 85{,}2$, $p < 0{,}01$, partielles $\eta^2 = 0{,}34$).

Alle Kinder, die in der analogen Testung des WRT 3+ Leistungen im unterdurchschnittlichen Bereich erzielten, schnitten auch in der digitalen Testung unterdurchschnittlich ab (Tab. 6). Von 65 Kindern, deren Testleistung in der analogen Testung im mindestens durchschnittlichen Bereich lag, hatten 47 Kinder in der digitalen Testung ebenfalls mindestens durchschnittliche Testergebnisse. Die Trefferquote der Minderleistungen beim analog erfassten WRT 3+ durch die digitale Testleistung beträgt 73,5 %. Der RATZ-Index liegt bei 100 %.

In der analogen Testung des WRT 4+ lagen die Leistungen von 22 Kindern im unterdurchschnittlichen Bereich. 21 dieser Kinder erzielten auch in der digitalen Testung Leistungen im unterdurchschnittlichen Bereich. Von den 143 Kindern, deren analog erfasste Testleistung im mindestens durchschnittlichen Bereich lag, erzielten 122 Kinder auch in der digitalen Testung mindestens durchschnittliche Testwerte (Tab. 6). Die Trefferquote liegt somit bei 86,7 %. Der RATZ-Index beträgt 93,9 %.

Für das digitale Testformat des WRT 3+ und des WRT 4+ zeigt sich ein mittlerer Zusammenhang zur Deutschnote (WRT 3+: $N = 1782$, $r = -0{,}374$, $p < 0{,}01$; WRT 4+: $N = 2077$, $r = -0{,}466$, $p < 0{,}01$). Ein ebenfalls mittlerer Zusammenhang besteht zwischen den Testleistungen des analogen Testformats und den Deutschnoten (3. Klasse: $N = 66$, $r = -0{,}53$, $p < 0{,}01$; 4. Klasse: $N = 164$, $r = -0{,}58$, $p < 0{,}01$).

Die Rohwertdifferenz der beiden Testversionen unterscheidet sich nicht zwischen Kindern mit und ohne eigenem ToS, sowohl für den WRT 3+ als auch den WRT 4+ (Tab. 5). Die Varianzanalyse mit Messwiederholung ergab keinen statistisch bedeutsamen Interaktionseffekt (WRT 3+: $F(1{,}66) = 1{,}7$, $p = 0{,}17$; WRT 4+: $F(1{,}163) = 0{,}3$, $p = 0{,}59$).

Die Differenzen der Itemschwierigkeiten der beiden Testformate liegen beim WRT 3+ zwischen –0,17 und 0,33 (analoge Testung: $p_M = 0{,}65$; digitale Testung: $p_M = 0{,}57$; Differenz $p_M = 0{,}08$). Für den WRT 4+ liegen die Differenzen der Itemschwierigkeiten zwischen –0,08 und 0,21 (analoge Testung: $p_M = 0{,}68$; digitale Tes-

Tab. 7 Mittlere unspezifische Itemschwierigkeiten des WRT 3+ und 4+ für die beiden Testformate

		Items mit Großschreibung	Items mit Kleinschreibung
	N Items	*M (SD)*	*M (SD)*
WRT 3+			
Digital	55	0,49 (0,15)	0,63 (0,19)
Analog	55	0,62 (0,18)	0,68 (0,18)
WRT 4+			
Digital	60	0,57 (0,14)	0,66 (0,16)
Analog	60	0,63 (0,17)	0,67 (0,15)

Tab. 8 Mittelwerte (M) und Standardabweichungen (*SD*) der Prozentsätze an GK-Fehler im WRT 3+ und WRT 4+ für die beiden Testformate

		Digitale Testung	Analoge Testung
	N Items	*M (SD)*	*M (SD)*
WRT 3+			
GK-Fehler bei Items mit Kleinschreibung	33	3,70 (5,8)	6,48 (9,3)
GK-Fehler bei Items mit Großschreibung	22	29,0 (19,6)	16,9 (21,5)
GK-Fehler gesamt	55	13,8 (18,0)	10,6 (16,0)
WRT 4+			
GK-Fehler bei Items mit Großschreibung	40	1,5 (2,3)	3,1 (3,9)
GK-Fehler bei Items mit Kleinschreibung	20	25,2 (19,4)	20,9 (18,5)
GK-Fehler gesamt	60	9,4 (15,9)	9,0 (13,9)

GK Groß-/Kleinschreibung

tung: $p_M = 0{,}63$; Differenz $p_M = 0{,}05$; Itemkennwerte siehe Tab. E1 und E2 im Onlinematerial 2).

In der digitalen Testung sind großgeschriebene Zielitems deutlich schwieriger im Vergleich zu kleingeschriebenen Zielitems (WRT 3+: $t(53) = 2{,}94$, $p < 0{,}01$, $d = 0{,}82$; WRT 4+: $t(58) = 2{,}15$, $p < 0{,}05$, $d = 0{,}60$) während in der analogen Testung kein bedeutsamer Unterschied zwischen groß- und kleingeschriebenen Testitems besteht (WRT 3+: $t(53) = 1{,}26$, $p = 0{,}215$, $d = 0{,}33$; WRT 4+: $t(58) = 1{,}69$, $p = 0{,}096$, $d = 0{,}25$; vgl. Tab. 7). Die Häufigkeit von Groß-/Kleinschreibungsfehlern (GK-Fehler) liegt für den WRT 3+ bei der digitalen Testung zwischen 0 und 85,3 % und bei der analogen Testung zwischen 0 und 86,8 %. Die Häufigkeit der GK-Fehler ist beim WRT 3+ höher beim digitalen im Vergleich zum analogen Testformat ($t(54) = 2{,}4$, $p < 0{,}05$, $d = 0{,}32$). Für den WRT 4+ liegt die Häufigkeit von GK-Fehlern bei der digitalen Testung zwischen 0 und 55,4 % und bei der analogen Testung zwischen 0 und 41,8 %. Zwischen den beiden Testformaten des WRT 4+ gibt es keinen bedeutsamen Unterschied in der Häufigkeit der GK-Fehler ($t(59) = 0{,}5$, $p = 0{,}60$, $d = 0{,}07$). Für kleingeschriebene Zielitems ist die GK-Fehlerrate in der digitalen Testung geringer als in der analogen Testung (WRT 3+: $t(32) = -3{,}2$; $p < 0{,}01$, $d = 0{,}55$; WRT 4+: $t(39) = -4{,}7$; $p < 0{,}01$, $d = 0{,}74$), während für großgeschriebene Zielitems die GK-Fehlerrate in der analogen Testung niedriger ist als in der digitalen Testung (WRT 3+: $t(21) = 6{,}3$, $p < 0{,}01$, $d = 1{,}33$; WRT 4+: $t(19) = 2{,}9$, $p < 0{,}05$, $d = 0{,}64$; vgl. Tab. 8).

4 Diskussion

Die Überprüfung der Übertragbarkeit zweier Lese- und eines Rechtschreibtests von einem analogen (Papier-Bleistift-Format) in ein digitales Format stand im Fokus der vorliegenden Studie. Neben der Übereinstimmung der beiden Testformate wurde auch untersucht, ob jeweils die gleichen Kinder als solche mit unterdurchschnittlicher Lese- und Rechtschreibleistungen identifiziert werden und ob die Vertrautheit im Umgang mit Tablets oder Smartphones die Testleistung im digitalen Format beeinflussen.

4.1 Vergleichbarkeit der Testleistungen zwischen den Testformaten

Für alle Tests (VSL, WLLP-R, WRT 3+ und WRT 4+) konnte ein zufriedenstellender Zusammenhang der Testleistungen bei den beiden Testformen (analog und digital) gefunden werden, wobei die Rohwerte und T-Werte der analogen Testung für alle Tests bedeutsam über den Werten der digitalen Testung liegen. Ein bemerkenswert hoher Unterschied zeigt sich zwischen dem analogen und digitalen T-Werte der VSL, wobei die erzielten Leistungen in der analogen Testung mehr als eine Standartabweichung über denen der digitalen Testung liegen. Da die Streuung des Zeitintervalls zwischen der digitalen und analogen Testung sehr hoch ist, können die Rohwertunterschiede einen Lerneffekt aufgrund von Testwiederholung oder auch auf einen Entwicklungseffekt zurückgeführt werden. Die Effektstärken in der Dekodiergeschwindigkeit und dem Leseverständnis der aktuellen Studie liegen über den in der Literatur berichteten Vergleichswerten für das Konstrukt Lesen (Hill et al. 2008), was einem größeren Leistungszuwachs entspricht als aufgrund von Entwicklungseffekten zu erwarten. Dieser Vergleich ist allerdings aufgrund von Unterschieden in der Sprache und dem Schulsystem (die zum Vergleich herangezogene Studie fand in den USA statt) mit Einschränkungen verbunden. Aus diesem Grund sind die deutschsprachigen Normdaten der Testverfahren ebenfalls zum Vergleich herangezogen. Aus ihnen lassen sich grobe Richtwerte für Entwicklungseffekte ableiten (siehe Onlinematerial 1). Zusammenfassend führt ein Vergleich mit diesen Richtwerten zum Ergebnis, dass die Rohwertdifferenz im Leseverständnis (VSL) für die Klassenstufe 3 und 4 als auch in der Dekodiergeschwindigkeit (WLLP-R) für die Klassenstufe 3 in etwa dem zu erwartenden Entwicklungseffekt des jeweiligen Testverfahrens entsprechen. Für die Rechtschreibleistung in Klasse 3 (WRT 3+) liegt die Rohwertdifferenz über dem zu erwartenden Entwicklungseffekt. Schwierigkeiten bzw. geringe Erfahrung beim Bedienen einer Tastatur sind eine naheliegende Erklärung für dieses Ergebnis. Für die Dekodiergeschwindigkeit und Rechtschreibleistung in Klassenstufe 4 konnte dieser Vergleich aufgrund unzureichender Normwerte nicht bzw. nur eingeschränkt durchgeführt werden. Richtwerte für Testwiederholungseffekte konnten nicht abgeleitet werden.

Die Übereinstimmung bei der Identifikation unterdurchschnittlicher Leistungen in der Dekodiergeschwindigkeit (WLLP-R) ist eher moderat. Von den Kindern, die im analogen Testformat einen T-Wert ≤40 erzielten ($N = 42$), erzielte ca. jedes vierte Kind (26 %, $N = 11$) in der vorhergehenden digitalen Testung T-Werte über 40. Das überrascht, da Lerneffekte durch Testwiederholung und mögliche Entwicklungszu-

wächse ein umgekehrtes Ergebnis haben erwarten lassen. Für 88 % der Kinder mit durchschnittlichen Testleistungen im analogen Format zeigten sich auch in der digitalen Version durchschnittliche Testleitungen (168 von 191). Der RATZ-Index von 65,9 % verdeutlicht, dass die Gesamttrefferquote gegenüber der Zufallstrefferquote zufriedenstellend ist. Die Testsituation der digitalen Testung stellt eine mögliche Erklärung für die schlechte Übereinstimmung dar. Die Kinder wurden bei der Durchführung nicht durch eine neutrale Testleitung begleitet, was die Zuhilfenahme anderer Personen (Eltern oder ältere Geschwister) ermöglicht und wiederum zu einer Überschätzung der tatsächlichen Leistung führen kann. Solche, auch unter dem Begriff „faking good" erforschten Täuschungen treten in begleiteten Testsituationen seltener auf und unterstreicht die Relevanz kontrollierter Testsituationen (Steger et al. 2018). Die spielerische Rahmengeschichte, anhand derer die Kinder durch die App geleitet wurden, wurde zwar explizit minimalistisch konzipiert und Motivationselemente wurden lediglich zwischen den Testverfahren eingesetzt, dennoch liefern sie weitere Erklärungsmöglichkeiten für die schlechte Übereinstimmung, da der Einsatz motivierender Elemente einen positiven Effekt auf die Testleistung haben kann. Die Rahmengeschichte wurde eingesetzt, um die Kinder nach einem abgeschlossenen Test zu motivieren, die weiteren Aufgaben des Tages zu bearbeiten und nach einem abgeschlossenen Tag am nächsten Tag erneut die App zu öffnen, um die weiteren Aufgaben zu bearbeiten. Dieses Ziel wurde erreicht, wie die überdurchschnittlich gute *retention-rate* (Prozentsatz an Appnutzern, die nach einer bestimmten Zeitperiode weiterhin die App nutzen) von 85 % belegt. Auswertungen weltweiter App-Nutzungsdaten ergeben, dass an Tag 7 nach der Installation *retention-rates* von durchschnittlich 6,3 % (im Bereich Bildung) bis zu 19,8 % (im Bereich Nachrichten) erreicht werden (Statista 2020).

Für die VSL sowie für den WRT 3+ und 4+ zeigt sich bei der Identifikation unterdurchschnittlicher Leistungen, dass in der analogen Testung deutlich weniger Kinder unterdurchschnittlich abschnitten als in der digitalen Testung. Anders als bei der WLLP-R lässt sich dieser Unterschied nicht durch Zuhilfenahme anderer Personen oder Hilfsmittel erklären. Vielmehr kann vermutet werden, dass diese Unterschiede durch Lerneffekte (aufgrund von Testwiederholung) oder Entwicklungszuwächse (aufgrund der teils großen Zeitspanne zwischen den Testungen) erklärt werden können. Für den WRT 3+ konnte ein Effekt der Zeitspanne zwischen den Testungen auch statistisch belegt werden. Möglicherweise ist es bei der digitalen Testung aber auch zu Schwierigkeiten bei der Bedienung gekommen, die zu schlechteren Testergebnissen geführt haben. Unabhängig von den hier präsentierten Ergebnissen, muss an dieser Stelle deutlich darauf hingewiesen werden, dass Entscheidungen über individuelle Förderbedarfe grundsätzlich nicht auf einzelnen Testergebnissen beruhen sollten (Heimlich et al. 2013).

Die Kriteriumsvalidität der digitalen Testversionen von VSL, WLLP-R, WRT 3+ und WRT 4+ entsprechen den Anforderungen an die Testgütekriterien und sind vergleichbar mit den Werten der traditionellen analogen Testung.

4.2 Testfairness

Bei der WLLP-R konnte kein Unterschied in der Testleistungsdifferenz zwischen Kindern mit und ohne eigenem ToS gefunden werden. Die einfache und intuitive Bedienung der digitalen Version der WLLP-R (Itemauswahl durch Antippen via Touchscreen und direkte Präsentation des nächsten Items) ist eine mögliche Erklärung dafür und spricht, bei einer kontrollierten Testsituation, für eine hohe Objektivität des digitalen Testformates. Für die beiden Rechtschreibtests konnte, trotz der systematischen Unterschiede bei der Groß-/Kleinschreibung, ebenfalls kein statistisch bedeutsamer Effekt für den Besitz eines eigenen Tablets oder Smartphones auf die Testleistungsdifferenz zwischen den beiden Versionen gefunden werden. Dies könnte dadurch erklärt werden, dass Kinder im Grundschulalter, auch wenn sie ein eigenes ToS besitzen, nur wenig Erfahrung im Schreiben mit Tastaturen haben.

Bei der VSL zeigten Kinder ohne eigenes ToS höhere Testleistungen in der analogen gegenüber der digitalen Testung, während Kinder mit eigenem ToS vergleichbare Testleistungen in den beiden Testformaten zeigten. Möglicherweise konnten stärkere Leseverständnisleistungen der Kinder ohne ToS in der digitalen Version der VSL aufgrund von Bedienungsschwierigkeiten nicht abgebildet werden.

4.3 Testformatspezifische Fehlerschwerpunkte

Die Itemschwierigkeiten des WRT 3+/4+ weisen im Mittelwert eine zufriedenstellende Übereinstimmung mit geringfügig höheren Werten im analogen Testformat auf. Beim WRT 3+ gibt es 7 Items („beginnen“, „vorbereitet“, „kämpfte“, „Käfig“, „Lehrerin“, „Wette“, „Kohlen“) die Differenzen ≥0,20 aufweisen, während dies beim WRT 4+ nur auf ein Item („traurig“) zutrifft. Diese Items weisen nicht auf testformatspezifische Fehlerkategorien hin. Der Vergleich der Itemschwierigkeiten großgeschriebener und kleingeschriebener Zielitems zeigt allerdings, dass es testformatspezifische Unterschiede der Itemschwierigkeiten gibt. Während in der digitalen Testung weniger großgeschriebene Zielitems richtig gelöst werden als kleingeschriebene Zielitems, ergibt sich in der analogen Testung kein Unterschied.

Dies sagt jedoch noch nichts darüber aus, ob in der digitalen im Vergleich zur analogen Testung tatsächlich mehr Groß-/Kleinschreibungsfehler (GK-Fehler) auftreten. Der Vergleich der GK-Fehlerraten zeigt, dass bei großgeschriebenen Zielitems in der digitalen Testung sowohl im WRT 3+ als auch im WRT 4+ mehr GK-Fehler auftreten als in der analogen Testung. Für den WRT 4+ fällt dieser Unterschied deutlich geringer aus als beim WRT 3+. Dies hat möglicherweise damit zu tun, dass Kinder der dritten Klassenstufe weniger Erfahrungen im Umgang mit Tastaturen haben, als Kinder der 4. Klassenstufe. Für kleingeschriebene Zielitems zeigt sich ein umgekehrtes Muster. In der analogen Testung treten mehr GK-Fehler auf als in der digitalen Testung. Dies gilt sowohl für den WRT 3+ als auch für den WRT 4+. Insgesamt verdeutlichen die Ergebnisse, dass Items in der digitalen Testung häufiger kleingeschrieben werden. Neben einer mangelnden Erfahrung im Umgang mit einer Tastatur könnte auch die veränderte digitale Schreibkultur, die eine häufige Missachtung der Groß-/Kleinschreibung beinhaltet, eine mögliche Erklärung für die Ergebnisse sein. Ob die Missachtung der Groß-Kleinschreibung in einem der üb-

lichen nicht formellen digitalen Kontexte tatsächlich in einen Zusammenhang mit der Groß-Kleinschreibung einer formalen Leistungstestung gebracht werden kann, bleibt allerdings klärungsbedürftig. So steigt beispielsweise bei Erwachsenen das orthografische Normbewusstsein in der mobilen schriftbasierten Kommunikation mit dem Grad der Formalität an (Morel und Natale 2019). Welchen Effekt etwa die Nutzung von Messenger-Diensten auf die digital erfasste Rechtschreibleistung bei Kindern haben, sollte in zukünftigen Studien untersucht werden. Dies ist für die Bewertung digital erfasster Rechtschreibleistungen von zentraler Bedeutung. Dabei gilt es auch die starke Dynamik im Mediennutzungsverhalten von Kindern zu beachten. Während unter den 8–9 Jahre alten Kindern 36 % regelmäßig WhatsApp nutzen, sind es bei den 10–11 Jahre alten Kindern bereits 73 % (Medienpädagogischer Forschungsverbund Südwest 2018).

4.4 Limitationen der Studie

Bei der Interpretation der Ergebnisse der vorgelegten Studie sind methodische Einschränkungen zu berücksichtigen. Eine zentrale Einschränkung ist, dass alle teilnehmenden Kinder die zwei Testformate in der gleichen Reihenfolge bearbeitet haben (die digitale vor der analogen Testung). Mögliche Wiederholungseffekte auf die Leistungsdifferenz zwischen den beiden Testformaten konnten deshalb nicht kontrolliert werden. Außerdem variieren die Zeitspannen zwischen digitaler und analoger Testung erheblich (19 bis 237 Tage). Einerseits könnten für Kinder mit einer großen Zeitspanne zwischen den beiden Testungen Entwicklungseffekte einen Einfluss auf die Testleistungsdifferenzen gehabt haben; andererseits könnten Lerneffekte in diesen Fällen kleiner ausfallen. Mit einer Vorabanalyse wurde der Effekt der Zeitspanne zwischen den Testungen auf die Rohwertdifferenz zwischen den Testungen überprüft und für die meisten der untersuchten Testverfahren ausgeschlossen, nicht allerdings für den WRT 4+.

Möglicherweise lassen sich unter Rückgriff auf die Analysemöglichkeiten der probabilistischen Testtheorie die Testwiederholungseffekte gezielter untersuchen. Im Rahmen dieser Studie entschieden wir uns allerdings gegen diese Ansätze, weil die hier untersuchten einschlägigen Testverfahren zur Erfassung schriftsprachlicher Leistungen im Grundschulalter im Rahmen der klassischen Testtheorie konstruiert wurden.

Eine weitere Einschränkung ergibt sich aus der unbeobachteten Testsituation der digitalen Testung. Mögliche Fehlerquellen während der Testung konnten hier nicht durch eine Testleitung beobachtet und dokumentiert bzw. abgewendet werden. Neben der Zuhilfenahme anderer Personen oder Hilfsmittel kann es auch zu Störungen und Ablenkungen durch andere Personen gekommen sein. Auch wenn wir Ergebnisse mit großen Abweichungen ausgeschlossen haben, können diese Fehlerquellen erheblichen Einfluss auf die Testleistung haben. Eine individuelle Leistungsbestimmung sollte deshalb in der individuellen Diagnostik ausschließlich in beobachteten Testsituationen stattfinden.

Einschränkungen gibt es auch bei der Frage, ob Routinen im Umgang mit Tablets oder Smartphones einen Einfluss auf die Testleistungen haben. Ob die Kinder routiniert mit ToS umgehen, wurde im Rahmen der Studie nicht getestet. Erfasst wurde,

ob die Kinder ein eigenes ToS besitzen. Da der Besitz der Geräte mit einer regelmäßigen Bedienung dieser verbunden ist, kann zumindest davon ausgegangen werden, dass diese Kinder routinierter als andere Gleichaltrige mit ToS umgehen. Der Einfluss von Routinen im Umgang mit Tablets und Smartphones auf digital erfasste Lese- und Rechtschreibleistungen sollte in zukünftigen Studien systematisch untersucht werden, da sich hieraus wichtige Implikationen für die Leistungsdiagnostik ergeben können.

Als eine weitere Einschränkung muss berücksichtigt werden, dass es sich um keine repräsentative Stichprobe handelt, da Kinder mit unterdurchschnittlichen Lese- und/oder Rechtschreibleistungen sowie Jungen überrepräsentiert sind. Zusätzlich kann vermutet werden, dass sich in Hessen gerade Eltern mit Klärungsbedarf bezüglich der schulischen Leistungen ihrer Kinder für die analoge Testung angemeldet haben, was auch die Überrepräsentation von Kindern der 4. Klassenstufe erklären könnte, da der anstehende Übertritt bei Eltern das Bewusstsein für Klärungsbedarfe schärfen dürfte. Für die VSL muss als zusätzliche Limitation berücksichtigt werden, dass aufgrund eines technischen Fehlers die Antworten von 4 der 50 Items aus der analogen Version nicht abgespeichert wurden. Beim Vergleich der Rohwertdaten wurden für die analoge Testung diese 4 Items bei den Analysen ausgeschlossen. Zur Ermittlung der T-Werte für die analoge Testung wurden jedoch alle Items berücksichtigt.

Zukünftige Studien zur Untersuchung von Testmoduseffekten sollten die Planungsschwächen der vorliegenden Untersuchung für die hier bearbeiteten Fragen vermeiden. Besonders sollten dabei folgende Aspekte berücksichtigt werden: I) Um Testmoduseffekte systematisch zu untersuchen ist ein „counterbalanced Design" zu realisieren, so dass Testmoduseffekte und Lerneffekte bzw. Entwicklungseffekte voneinander abgrenzbar sind; II) bekannte Einflussfaktoren, wie z. B. das verwendete Endgerät, sollten systematisch und umfangreich erfasst oder kontrolliert werden; III) digitale Verfahren sollten sowohl für die Testdurchführung als auch den Datenexport vorab umfangreich, mit allen Teilprozessen, pilotiert werden.

4.5 Fazit und Ausblick

Zusammenfassend zeigen die Ergebnisse, dass die digitalen Testversionen valide zu sein scheinen und dass beide Testformate dasselbe Konstrukt erfassen. Aufgrund der fehlenden Äquivalenz der Mittelwerte kann allerdings nicht von einer Parallelität der Tests gesprochen werden. Sowohl die signifikant höheren Werte der analogen Testung als auch die nicht zufriedenstellende Übereinstimmung bei der Identifizierung unterdurchschnittlicher Testleistungen, sind vermutlich auf Wiederholungs- und Entwicklungseffekte zurückzuführen. Möglicherweise spielten auch Bedienungsschwierigkeiten eine Rolle, aber das konnte aufgrund des Studiendesigns nicht überprüft werden.

Trotz der genannten Einschränkungen haben die Ergebnisse der vorliegenden Studie eine hohe praktische Relevanz, zeigen sie doch auf, dass eine Erfassung der Dekodiergeschwindigkeit, des Leseverständnisses und der Rechtschreibleistung mit einer mobilen App im Grundschulalter grundsätzlich möglich ist. Allerdings dürfen Unterschiede in der Darstellung und Handhabung keinesfalls unterschätzt werden.

Dies spielt insbesondere eine zentrale Rolle, wenn ein flexibles Wechseln zwischen den Testformaten angestrebt wird um damit individuell auf den Bedarf der Kinder und auf die Situation zu reagieren. Die Äquivalenz der Testformate ist hierfür eine grundlegende Voraussetzung.

Den Einfluss eines routinierten Umgangs mit Tablets und Smartphones auf digital erfasste Lese- und Rechtschreibleistungen gilt es weiter zu untersuchen. Aufgrund der unterschiedlichen Darstellung und Handhabung sollte die Äquivalenz unterschiedlicher Testformate stets überprüft werden und bei Bedarf gesonderte Normen bereitgestellt und verwendet werden. Dennoch bietet die digitale Leistungserfassung einen klaren ökonomischen Vorteil gegenüber der traditionellen analogen Testung. Die individuelle Beobachtung und Einschätzung durch eine Fachkraft wird sie aus heutiger Sicht jedoch nicht ersetzen können.

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s11618-022-01068-1) enthalten.

Danksagung Wir danken den teilnehmenden Kindern und deren Eltern sowie der Meister Cody GmbH für die Unterstützung des Forschungsprojekts. Frau Grunwald danken wir für die Unterstützung bei der Datenerhebung sowie Herrn Haberstroh und Herrn Wörner für die statistische Beratung.

Förderung Diese Studie wurde durch das Bundesministerium für Bildung und Forschung finanziert.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt J. Rothe, L. Visser, R. Görgen, J. Kalmar, G. Schulte-Körne und M. Hasselhorn geben an, dass kein Interessenkonflikt besteht.

Literatur

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, *9*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, *6*(9), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x.

Berninger, V. W., Nagy, W., Tanimoto, S., Thompson, R., & Abbott, R. D. (2015). Computer instruction in handwriting, spelling, and composing for students with specific learning disabilities in grades 4–9. *Computers & education*, *81*, 154–168.

Birkel, P. (2007a). *Weingartener Grundwortschatz Rechtschreib-Test für dritte und vierte Klassen (WRT 3+). Manual*. Göttingen: Hogrefe.

Birkel, P. (2007b). *Weingartener Grundwortschatz Rechtschreib-Test für vierte und fünfte Klassen:(WRT 4+) Manual*. Göttingen: Hogrefe.
Clinton, V. (2019). Reading from paper compared to screens: a systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325.
Cockburn, A., Ahlström, D., & Gutwin, C. (2012). Understanding performance in touch selections: tap, drag and radial pointing drag with finger, stylus and mouse. *International Journal of Human-Computer Studies*, *70*(3), 218–233.
Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology*, *77*(2), 479–492.
Davis, L.L., Janiszewska, I., Schwarts, R., & Holland, L. (2016). NAPLAN device effects study. https://nap.edu.au/docs/default-source/default-document-library/naplan-online-device-effect-study.pdf?sfvrsn=2. Zugegriffen: 20. Juli 2021.
Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: a meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38.
Eickelmann, B., & Gerick, J. (2020). Lernen mit digitalen Medien. Die Deutsche Schule. *Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis*, 153–162. https://doi.org/10.31244/9783830992318.
Elliott, L.J., Ljubijanac, M., & Wieczorek, D. (2019). The effect of screen size on reading speed: a comparison of three screens to print. In *International Conference on Applied Human Factors and Ergonomics* (S. 103–109). Cham: Springer.
Findlater, L., Froehlich, J.E., Fattal, K., Wobbrock, J.O., & Dastyar, T. (2013). Age-related differences in performance with touchscreens compared to traditional mouse input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 343–346). : ACM.
Geiger, M., & Wilhelm, O. (2018). Methodische Abwägungen zur Validität onlinebasierter Leistungsprüfungen. *Lernen und Lernstörungen*, *7*(4), 215–218.
Gerth, S., Klassert, A., Dolk, T., Fliesser, M., Fischer, M.H., Nottbusch, G., & Festman, J. (2016). Is handwriting performance affected by the writing surface? Comparing preschoolers', second graders', and adults' writing performance on a tablet vs. Paper. *Frontiers in psychology*, *7*, 1308.
Goldhammer, F., Harrison, S., Bürger, S., Kröhne, U., Lüdtke, O., Robitzsch, A., Köller, O., Heine, J.-H., & Mang, J. (2019). Vertiefende Analysen zur Umstellung des Modus von Papier auf Computer. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018: Grundbildung im internationalen Vergleich* (S. 163–186). Münster: Waxmann.
Hatlevik, O., Ottestad, G., & Throndsen, I. (2015). Predictors of digital competence in 7th grade: a multilevel analysis. *Journal of Computer Assisted Learning*, *31*(3), 220–231. https://doi.org/10.1111/jcal.12065.
Heimlich, U., Lutz, S., & Wilfert de Icaza, K. (2013). *Ratgeber Förderdiagnostik. Feststellung des sonderpädagogischen Förderbedarfs im Förderschwerpunkt Lernen*. Hamburg: Persen.
Hill, C.J., Bloom, H.S., Black, A.R., & Lipsey, M.W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, *2*(3), 172–177.
Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., & Ziefle, M. (2011). Investigating paper vs. screen in real-life hospital workflows: performance contradicts perceived superiority of paper in the user experience. *International Journal of Human-Computer Studies*, *69*(9), 563–570.
Horkay, N., Bennett, R.E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), n2.
Irmert, N., & Trutwin, E. (2020). The different effects of digital devices on students' motivation – evidence from the United States. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9030897&fileOId=9030898. Master thesis, Lund University. Zugegriffen: 19. Juli 2021.
Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (1999). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC)*. Göttingen: Hogrefe.
Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420.
Klemm, K., & Zorn, D. (2019). *Steigende Schülerzahlen im Primarbereich: Lehrkräftemangel deutlich stärker als von der KMK erwartet*. Gütersloh: Bertelsmann Stiftung.
Kong, Y., Seo, Y.S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: a meta-analysis. *Computers & Education*, *123*, 138–149.
Kuhn, J.T., & Schwenk, C. (2018). Onlinebasierte Diagnostik mathematischer Kompetenzen: Möglichkeiten und Grenzen. *Lernen und Lernstörungen*, *7*(4), 231–235.

Lauth-Lebens, M., & Lauth, G. W. (2020). Motivationale Einflüsse auf exekutive Funktionen bei Aufmerksamkeitsdefizit-/Hyperaktivitätsstörungen (ADHS). *Lernen und Lernstörungen*, *9*(2), 111–125. https://doi.org/10.1024/2235-0977/a000284.

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

Lindberg, S., Hasselhorn, M., & Lonnemann, J. (2018). Förderrelevante Diagnostik bei Lernstörungen. *Lernen und Lernstörungen*, *7*(4), 197–201.

Marx, P., & Lenhard, W. (2010). Diagnostische Merkmale von Screeningverfahren. In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen*. Göttingen: Hogrefe.

Medienpädagogischer Forschungsverbund Südwest (Hrsg.). (2018). *JIM-Studie 2018: Jugend, Information, (Multi-)Media; Basisuntersuchung zum Medienumgang 12- bis 19jähriger*. Stuttgart: mpfs.

Morel, E., & Natale, S. (2019). Orthographie in WhatsApp & Co: Eine Untersuchung zum Normbewusstsein in der mobilen schriftbasierten Kommunikation. *Networx*, *85*, 1–20.

Noyes, J., & Garland, K. (2003). VDT versus paper-based text: reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, *31*, 411–423.

Scherer, R., & Siddiq, F. (2019). The relation between students' socioeconomic status and ICT literacy: findings from a meta-analysis. *Computers & Education*, *138*, 13–32. https://doi.org/10.1016/j.compedu.2019.04.011.

Schneider, W., Blanke, I., Faust, V. & Küspert, P. (2011). *WLLP-R. Würzburger Leise Leseprobe - Revision*. Göttingen: Hogrefe.

Schneider, N., Wilkes, J., Grandt, M., & Schlick, C. M. (2008). Investigation of input devices for the age-differentiated design of human-computer interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(2), 144–148.

Schulte-Körne, G., Lonnemann, J., Lindberg, S., & Hasselhorn, M. (2018). Neue Wege in der Diagnostik und Förderung bei schulischen Entwicklungsstörungen. *Lernen und Lernstörungen*, *7*(4), 195–196.

Shin, J., Bulut, O., & Gierl, M. J. (2020). The effect of the most-attractive-distractor location on multiple-choice item difficulty. *The Journal of Experimental Education*, *88*(4), 643–659.

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*, *27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007.

Statista Mobile app user retention rate worldwide 2020, by vertical. https://www.statista.com/statistics/259329/ios-and-android-app-user-retention-rate/. Zugegriffen: 3. Aug. 2021.

Steffener, J., Rana, Z., Dancey, S., Chang, Y. Y., Hernandez, F. R., & Guy, C. (2020). Screen position effects on task performance in a delayed match to sample task. *Acta Psychologica*, *208*, 103123.

Steger, D., Schroeders, U., & Gnambs, T. (2018). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000494.

Tumbas, P., Sakal, M., Pavlicevic, V., & Rakovic, L. (2019). Digital competencies in business informat-ics curriculum innovation. In L. Gómez Chova, A. López Martínez & I. Candel Torres (Hrsg.), *INTED Proceedings, INTED2019 Proceedings* (S. 9655–9664). https://doi.org/10.21125/inted.2019.2400.

Visser, L., Kalmar, J., Linkersdörfer, J., Görgen, R., Rothe, J., Hasselhorn, M., & Schulte-Körne, G. (2020). Comorbidities between specific learning disorders and psychopathology in elementary school children in Germany. *Frontiers in Psychiatry, 11*, 292. https://doi.org/10.3389/fpsyt.2020.00292.

Walter, J. (2013). Lernfortschrittsdiagnostik Lesen (LDL) und Verlaufsdiagnostik sinnerfassenden Lesens (VSL): Zwei Verfahren als Instrumente einer formativ orientierten Lesediagnostik. In *Lernverlaufsdiagnostik* (S. 165–201).

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* (S. 354–359). Boston: American Statistical Association.