Tabletbasierter Fehleridentifikationstest zur ökonomischen und validen Erfassung von Rechtschreibfähigkeiten in der Grundschule

Darius Endlich, Wolfgang Lenhard, Peter Marx & Tobias Richter

Lehrstuhl für Psychologie IV, Universität Würzburg

Manuskript zur Publikation angenommen in der Zeitschrift
***Lernen und Lernstörungen* (1/2021)**

Autorenhinweis

Die hier dargestellte Forschung wurde vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projekts "Lernstörungen – Onlineplattform für Diagnostik und Intervention (LONDI)" gefördert.

Parallel zur vorliegenden Arbeit streben die Autoren eine weitere Veröffentlichung mit unterschiedlichem inhaltlichen Fokus an, deren Auswertungen sich teilweise auf den gleichen Datensatz beziehen.

Korrespondierender Autor:
Dr. Darius Endlich, Universität Würzburg, Lehrstuhl für Psychologie IV, Röntgenring 10, 97070 Würzburg, darius.endlich@uni-wuerzburg.de

Zusammenfassung

Einleitung

In der vorliegenden Arbeit wurde das Potenzial von tabletbasierten Fehleridentifikationstests zur Erfassung der Rechtschreibleistung in der Grundschule untersucht. Studien aus dem englischen Sprachraum belegen hohe Zusammenhänge zwischen Leistungen in klassischen Diktaten und Leistungen in Aufgaben, in denen Rechtschreibfehler in präsentierten Texten zu identifizieren sind (Fehleridentifikationstests). Im deutschen Sprachraum hingegen liegen ähnliche Untersuchungen bisher nur für die Sekundarstufe vor.

Methode

Die vorliegende Arbeit untersuchte die produktive und rezeptive Rechtschreibleistung von 144 Schülerinnen und Schülern der Jahrgangsstufen 2 bis 4.

Ergebnisse

Es konnten hohe Zusammenhänge für die deutsche Primarstufe nachgewiesen werden (r = .69 bis r = .82), wobei als neuer Aspekt auch die Effizienz der Fehleridentifikation betrachtet wurde. Die Zusammenhänge stiegen über die Jahrgangsstufen hinweg an. Während Kinder Fehler, die Verstöße gegen die phonologische Wortform beinhalten, bereits in Jahrgangsstufe 2 zuverlässig erkannten, wurden Verstöße gegen Rechtschreibregeln erst im Verlauf der Grundschulzeit zuverlässiger erkannt. Um die diagnostisch besonders relevante Gruppe rechtschreibschwacher Kinder zu identifizieren, erwiesen sich jedoch auch Fehler mit Verstößen gegen die phonologische Wortform als bedeutsam.

Diskussion

Computergestützte Fehleridentifikationstests stellen somit ein valides und ökonomisches Instrument zur Erfassung von Rechtschreibkompetenzen in der Primarstufe dar.

Schlüsselwörter: Computergestütztes Testen, Diktat, Fehleridentifikation, Grundschule, Rechtschreibung

Abstract

Introduction

The potential of tablet-based error identification tests for the assessment of spelling performance in elementary school was investigated. Studies from the English-speaking world show high correlations between performance in classical dictations and performance in tasks in which spelling errors in presented texts are to be identified (error identification tests). In the German-speaking world, on the other hand, similar studies have so far only been available for the level of secondary school.

Method

The present study investigated the active and passive component of spelling of 144 students in Grade 2 to 4.

Results

High correlations were found for German primary level (r = .69 to r = .82), whereby the efficiency of error identification was also considered as a new aspect. The correlations increased across grades. While children reliably identified errors that involve violations of phonetic fidelity as early as in the Grade 2, violations of spelling rules were only detected more reliably in the course of primary school. In order to identify the diagnostically particularly relevant group of children with poor spelling skills, however, errors involving violations of phonetic fidelity also proved to be relevant.

Discussion

Computer-aided error identification tests thus represent a valid and economic instrument for measuring spelling skills at the primary level.

Keywords: computer-based assessment, dictation, error identification, primary school, spelling

Tabletbasierter Fehleridentifikationstest zur ökonomischen und validen Erfassung von Rechtschreibfähigkeiten in der Primarstufe

Einleitung

Onlineinstrumente zur Leistungsmessung in der Schule sind in den vergangenen Jahren immer wichtiger geworden. Während für zwei der drei Kernkompetenzen in der Grundschule – Lesen und Rechnen – bereits eine Reihe wissenschaftlich fundierter, computerbasierter Testverfahren vorliegen (z.B. Kuhn, Schwenk, Raddatz, Dobel & Holling, 2017; Lenhard, Lenhard & Schneider, 2017; Richter, Naumann, Isberner, Neeb & Knoepke, 2017), fehlen derartige Verfahren zur Erfassung von Rechtschreibkompetenzen. Dieser Umstand ist darauf zurückzuführen, dass eine zuverlässige computerbasierte Erhebung produktiver Rechtschreibleistungen in Form von Diktaten oder Formen des freien Schreibens einen geübten Umgang mit der Tastatur voraussetzt, der gerade von jüngeren Grundschulkindern noch nicht erwartet werden kann. Computergestützte Versionen vorliegender analoger Testverfahren würden daher mit großer Wahrscheinlichkeit neben der Rechtschreibkompetenz auch die Vertrautheit im Umgang mit der Tastatur miterheben. Eine Lösung dieses Dilemmas könnte im Einsatz von Fehleridentifikationstests liegen (Richter, Lenhard, Marx & Endlich, 2018). Im Gegensatz zu Diktaten, die eine produktive Anwendung kognitiv repräsentierter Rechtschreibregeln sowie einen Abruf von Wortformen oder Morphemen aus dem mentalen Lexikon erfordern (produktive Rechtschreibung), sind für die Lösung von Fehleridentifikationsaufgaben (rezeptive Rechtschreibung) Rekognitionsprozesse und ein Abgleich von Graphemfolgen mit Graphem-Phonem-Korrespondenzregeln, Rechtschreibregeln und orthographischen Wortformen bzw. Morphemen im mentalen Lexikon erforderlich. Es handelt sich somit um unterschiedliche kognitive Anforderungen, die gleichwohl auf dieselbe Wissensbasis zugreifen. So berichten mehrere Studien aus dem englischen Sprachraum von hohen Korrelationen zwischen Tests der rezeptiven und produktiven Rechtschreibung in der Primarstufe (z. B. Howard et al., 2017). Im deutschen Sprachraum liegen ähnliche Befunde allerdings bisher nur für die Sekundarstufe vor. Die vorliegende Arbeit soll diese Lücke schließen, indem sie das Potenzial von computerbasierten Fehleridentifikationstests zur Erfassung der Rechtschreibleistung in der Grundschule untersucht.

Anhand von Theorien zum Lesen und Rechtschreiben wird zunächst herausgestellt, dass beiden Prozessen dasselbe orthografische Wissen zugrunde liegt. Nach einem Überblick über vorliegende Studien aus dem angelsächsischen Raum, die einen hohen Zusammenhang

von rezeptiver und produktiver Rechtschreibleistung sowohl in der Primar- als auch in der Sekundarstufe belegen, werden bereits vorhandene Fehleridentifikationstests für die Sekundarstufe im deutschen Sprachraum dargestellt. Sodann wird dafür argumentiert, dass ein ökonomisches Verfahren zur Erfassung der Rechtschreibleistung auch in der Primarstufe von großer Relevanz wäre.

Zwei-Wege-Modelle schriftsprachlicher Leistungen

Zwei-Wege-Modelle schriftsprachlicher Leistungen weisen darauf hin, dass sich sowohl die Fähigkeit des Lesens als auch die des Rechtschreibens auf die gleiche kognitive Architektur stützen. Das Zwei-Wege-Modell des Lesens (Coltheart, 1978) wie auch das Zwei-Wege-Modell des Rechtschreibens (Houghton & Zorzi, 2003; Romani, Olson & Di Betta, 2005, S. 432) postulieren jeweils eine direkte und eine indirekte Route auf dem Weg zum flüssigen Lesen bzw. dem korrekten Schreiben von Wörtern. Beim Lesen wird die direkte Route aktiviert, wenn sich das zu lesende Wort bereits im orthografischen Lexikon befindet und leicht zugänglich ist. Bei unbekannten (oder auch seltenen) Wörtern läuft der Prozess des Lesens über die nicht lexikalische Route, bei der anhand des Graphem-Phonem-Regelsystems Buchstabe für Buchstabe kodiert werden muss (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Ähnliche Prozesse werden für das Rechtschreiben angenommen: Während bei bekannten Wörtern der direkte Weg vom phonologischen Lexikon (Ort der Speicherung des Wortklangs) über das semantische Lexikon (Speicherung der Bedeutung des Wortes) hin zum orthografischen Lexikon (Speicherung der Schreibung des Wortes) aktiviert wird, muss bei unbekannten Wörtern auf die nicht lexikalische Route zurückgegriffen werden, bei der – umgekehrt zum Leseprozess – Phoneme in Grapheme rekodiert werden. Die indirekte Route ist beim Schreiben wie beim Lesen kognitiv aufwendig und fehleranfällig. Beim Schreiben ergibt sich zusätzlich das Problem, dass eine Vielzahl von Realisationen eines Wortes zwar die phonologische Form korrekt abbilden, jedoch meist nur eine Variante orthografisch gültig ist. Die deutsche Sprache ist also besonders hinsichtlich des Schreibens nicht zuverlässig lautgetreu (Landerl & Wimmer, 2008; für eine Klärung der Begrifflichkeit der Lauttreue, siehe Corvacho del Toro & Hoffmann-Erz, 2014).

Rechtschreibkompetenz und die Effizienz zugrunde liegender kognitiver Prozesse

Bei guten Rechtschreiber(inne)n gewinnt im Laufe des Schriftspracherwerbs die direkte Route zunehmend an Bedeutung. Schreibungen der meisten Wörter können somit automatisch und unter Einsatz nur sehr geringer kognitiver Ressourcen abgerufen werden.

Auf diese Weise bleiben kognitive Ressourcen für ressourcenintensive Prozesse wie Planungs- und Revisionsprozesse verfügbar, die für das gute Schreiben wichtig sind (vgl. Perfetti & Hart, 2002). Eine effiziente Fehleridentifikation – d.h. eine korrekte und zudem im Vergleich zur Referenzgruppe schnelle und somit ressourcenschonende Identifikation fehlerhaft verschriftlichter Wörter – kann nur mit gut zugänglichen orthografischen Repräsentationen gelingen. Während geübte Schreiberinnen und Schreiber auf gute, umfassende und leicht zugängliche orthografische Repräsentationen als Teil ihres mentalen Lexikons zurückgreifen können, müssen ungeübte Schreiberinnen und Schreiber häufiger die indirekte Route nutzen. Somit sollten sich Kinder mit guten Rechtschreibkompetenzen neben der Fähigkeit zu einer zuverlässigen Fehleridentifikation insbesondere auch durch eine effiziente Fehleridentifikationsleistung auszeichnen.

Fehleridentifikationstests zur Erfassung von Rechtschreibkompetenzen

Im englischsprachigen Raum werden in der Schule Verfahren zur Erfassung der rezeptiven Rechtschreibkompetenz häufig zur Überprüfung der allgemeinen Rechtschreibkompetenz eingesetzt (Howard et al., 2017). Der Einsatz von Testverfahren zur Fehleridentifikation (engl. error detection) und Fehlerkorrektur (engl. error correction) fußt auf zahlreichen Studien, die den Zusammenhang mit Leistungen im Diktat (engl. dictation) herausstellen. Bereits zu Beginn des 20. Jahrhunderts wurden in der englischsprachigen Literatur Verfahren zur rezeptiven Rechtschreibkompetenz erörtert (Guiler, 1929). Allen und Ager (1965) konnten zeigen, dass verschiedene Maße der produktiven und rezeptiven Rechtschreibung auf demselben Faktor luden und hoch miteinander korreliert waren. Freyberg (1970) verglich die Rechtschreibleistungen von Schülerinnen und Schülern einer neuseeländischen Mittelschule in einem Diktat sowie in einem Fehleridentifikationstest mit den Leistungen in einer Probe ihres freien Schreibens. Er fand hohe Korrelationen zwischen beiden Variablen und der freien Verschriftung, wobei die Leistung im Diktat leicht höher mit der Fehleranzahl im freien Schreiben korrelierte (r = .72) als die Leistung im Fehleridentifikationstest (r = .68). Der höchste Zusammenhang jedoch ergab sich zwischen der Leistung im Diktat und im Fehleridentifikationstest (r = .85). Eine Untersuchung von Croft (1982) an 9- und 10-jährigen Kindern erlaubt einen genaueren Vergleich verschiedener Maße der Rechtschreibkompetenz, da in allen untersuchten Variablen dieselben Wörter zum Einsatz kamen. Es wurden drei Maße der Rechtschreibleistung mit der Leistung im freien Schreiben korreliert: Neben einem Diktat und einem Fehlerkorrekturtest wurde auch ein Fehleridentifikationstest in Form eines Multiple-Choice-Tests dargeboten. Croft (1982)

berichtet ähnlich hohe Korrelationen zwischen dem Kriterium einerseits und dem Diktat (r = .73), dem Fehlerkorrekturtest (r = .73) und dem Fehleridentifikationstest (r = .66) andererseits. Die höchsten Korrelationen jedoch fanden sich auch hier, wie schon bei Freyberg (1970), für die verschiedenen Maße der Rechtschreibkompetenz untereinander – jenseits des freien Schreibens: Die Korrelationen bewegten sich zwischen .79 (Fehlerkorrektur und Fehleridentifikation), .82 (Diktat und Fehlerkorrektur) und .84 (Diktat und Fehleridentifikation). Somit fanden sich keine höheren Zusammenhänge, wenn zwei Leistungen der produktiven Rechtschreibung miteinander verglichen wurden, als wenn ein Maß der produktiven und ein Maß der rezeptiven Rechtschreibung miteinander korreliert wurden.

Ähnlich hohe Zusammenhänge zwischen Verfahren zur Fehleridentifikation und Diktaten wurden auch für Grundschulkinder nachgewiesen. So sieht Allred (1984) als Ergebnis seiner Studie den Einsatz von Fehleridentifikationstests in den Jahrgangsstufen 1 bis 6 zur Überprüfung von Rechtschreibkompetenzen im Allgemeinen als gerechtfertigt an: In allen Jahrgangsstufen zeigten sich durchgängig hohe Korrelationen zwischen Diktat und Fehleridentifikation, sowohl in unterschiedlich leistungsstarken Schulklassen als auch in getrennten Analysen nach Geschlechtern. Darüber hinaus konnte er nachweisen, dass nur Kinder der jüngeren Jahrgangsstufen in der rezeptiven Rechtschreibleistung signifikant besser abschnitten als in der produktiven Rechtschreibleistung, während sich für Kinder der Jahrgangsstufe 6 dieser Unterschied nicht nachwiesen ließ. Dies könnte darauf hindeuten, dass sich die passiven und aktiven Rechtschreibleistungen mit zunehmendem Alter der Schülerinnen und Schüler anpassen.

Bei einem Vergleich verschiedener Maße der Rechtschreibleistung in den Jahrgangsstufen 2 bis 5 fand Westwood (1999) die höchsten Zusammenhänge zwischen Diktat und Fehleridentifikation bzw. zwischen Diktat und Fehlerkorrektur (jeweils r = .90). Alle erhobenen Maße korrelierten geringer mit der Leistung im freien Schreiben: Diktat (r = .79), Fehleridentifikationstest (r = .77), Fehlerkorrektur (r = .79). Auch diese Ergebnisse unterstreichen, dass verschiedene Maße der produktiven Rechtschreibung in der Regel nicht höher miteinander assoziiert sind als ein Maß der produktiven und ein Maß der rezeptiven Rechtschreibung.

Fehleridentifikationstests im deutschsprachigen Raum

Einer Übertragung und Nutzbarmachung der konsistenten internationalen Studienergebnisse für die Rechtschreibdiagnostik im deutschsprachigen Raum steht der

Umstand entgegen, dass hier Fehleridentifikationstests zur Erfassung von Rechtschreibkompetenzen bislang kaum Beachtung geschenkt wurde. In der schulischen Praxis und in der Individualdiagnostik sind hier weiterhin Diktate zur Erfassung der Rechtschreibkompetenzen weitverbreitet – trotz teilweise massiver Kritik unter anderem an der Objektivität dieser Erhebungsmethode (vgl. Brinkmann, 2004; Fix, 2004). Erst seit etwa einem Jahrzehnt wird der Zusammenhang zwischen Maßen der produktiven und rezeptiven Rechtschreibleistung auch im deutschsprachigen Raum untersucht. So existieren mittlerweile für die Sekundarstufe zwei standardisierte Testverfahren (R-FIT 5-6+; M. Schneider, Martinez Méndez & Hasselhorn, 2014; R-FIT 9-10; Lenhart, Segerer, Marx & Schneider, im Druck). Bei beiden Verfahren handelt es sich um reine Fehleridentifikationstests: Aufgabe der Schülerinnen und Schüler ist es, in einem vorgelegten Text Fehler in einem Wort mit einem senkrechten Strich auf dem jeweiligen Buchstaben zu markieren. Mit Korrelationen von .81 bis .83 mit einem standardisierten Fließdiktat weist der R-FIT 5-6+ eine sehr gute konvergente Validität aus. Auch die interne Konsistenz ist in beiden Jahrgangsstufen überzeugend (Cronbachs $\alpha = .89$ bis $\alpha = .91$). Lenhart, Marx, Segerer & Schneider (2019) berichten einen annähernd perfekten Zusammenhang in der Gesamtstichprobe zwischen dem Fehleridentifikationstest und einem Lückendiktat ($r = .88$) sowie einen ebenfalls sehr guten Reliabilitätskennwert (Cronbachs $\alpha = .91$). Im Hinblick auf die diagnostisch besonders relevante Prognose rechtschreibschwacher Kinder ergaben sich im R-FIT 9-10 vergleichbare Werte wie im R-FIT 5-6+ (R-FIT 9-10: Sensitivität = 68%; R-FIT 5-6+: Sensitivität = 69%). Der relative Anstieg der Trefferquote gegenüber der Zufallstrefferquote (RATZ; vgl. P. Marx & Lenhard, 2010) ist im R-FIT 5-6+ als gut (RATZ = 61%), im R-FIT 9-10 sogar als sehr gut zu bewerten (RATZ = 76%).

Bei allen in Deutschland verfügbaren Verfahren zur Fehleridentifikation handelt es sich um Papier- und Bleistifttests für die Sekundarstufe. Dabei wäre gerade vor dem Hintergrund der Leistungsstabilität von Rechtschreibkompetenzen im Verlauf der Schulkarriere (Klicpera, Schabmann & Gasteiger-Klicpera, 1993; W. Schneider, 2008, 2009) ein ökonomisches Verfahren zur Erfassung der Rechtschreibkompetenz bereits in der Primarstufe wünschenswert: Ausgehend von einer frühzeitigen Identifikation von Leistungsrückständen könnten dann entsprechende Fördermaßnahmen eingeleitet werden, die sich gerade in der Primarstufe als sehr wirksam erwiesen haben (Berger, 2010). Den ohnehin schon bestehenden ökonomischen Vorteil eines Fehleridentifikationstests gegenüber einem Diktat würde eine computerbasierte Umsetzung nochmals erheblich steigern.

Fragestellungen und Hypothesen

Unsere Analysen beziehen sich auf die Validität des Fehleridentifikationstests zur Erfassung der Rechtschreibkompetenz (Fragestellung 1), auf die Unterscheidung von Fehlerprofilen in der Rechtschreibentwicklung (Fragestellung 2) und auf die korrekte Identifikation rechtschreibschwacher Kinder (Fragestellung 3). Dabei wird jeweils nicht nur die Antwortrichtigkeit, sondern auch die Effizienz betrachtet, mit der die Grundschulkinder Rechtschreibfehler identifizieren können (für ein ähnliches diagnostisches Konzept im Bereich des Lesens s. Richter et al., 2017).

Bezüglich der ersten Fragestellung zur Validität erwarten wir in den untersuchten Jahrgangsstufen 2 bis 4 substanzielle Zusammenhänge zwischen der Leistung im Lückendiktat (produktive Rechtschreibung) und der Leistung im Fehleridentifikationstest (rezeptive Rechtschreibung) (Hypothese 1a). Wir erwarten zudem eine systematische Zunahme des Zusammenhangs über die Jahrgangsstufen hinweg, da im Zuge der sukzessiven Herausbildung eines modalitätsunabhängigen Orthografiemoduls modalitätsspezifische Aspekte an Bedeutung verlieren sollten (Hypothese 1b). Und schließlich sollte sich der Zusammenhang auch im Extremgruppenvergleich zeigen. Rechtschreibschwache Kinder (definiert anhand einer unterdurchschnittlichen Leistung im DERET) sollten dementsprechend im Fehleridentifikationstest weniger akkurat und weniger effizient arbeiten als Kinder mit einer mindestens durchschnittlichen Rechtschreibkompetenz (Hypothese 1c).

Bezüglich der zweiten Fragestellung zur Unterscheidung von Fehlerprofilen gehen wir von der Annahme aus, dass Kinder am Beginn des Schriftspracherwerbs zunächst mit der Festigung der phonologischen Basis der orthografischen Kompetenzen beschäftigt sind, wohingegen im Laufe der Jahrgangsstufe 1 und insbesondere ab Jahrgangsstufe 2 orthografische Regeln zunehmend relevanter werden. Der Entwicklungslogik des Schriftspracherwerbs entsprechend sollte also die Identifikation von Fehlern mit Verstößen gegen Rechtschreibregeln grundsätzlich schwieriger sein als die Identifikation von Fehlern mit Verstößen gegen die phonologische Wortform (Hypothese 2a). Da angenommen wird, dass die zuverlässige Identifikation orthografischer Fehler erst im Verlauf der zweiten Jahrgangsstufe vielen Schülerinnen und Schülern gelingen wird, sollte sich zudem eine Interaktion der Jahrgangsstufe mit der Fehlerkategorie zeigen (Hypothese 2b).

Im Zusammenhang mit der dritten Fragestellung soll für den diagnostisch relevanten unteren Leistungsbereich (in der vorliegenden Stichprobe $PR \leq 16$ bzw. T-Wert ≤ 40)

geprüft werden, inwiefern eine zuverlässige Identifikation rechtschreibschwacher Kinder anhand der Leistung im Fehleridentifikationstest möglich ist. Da es sich bei dieser letzten Thematik nicht um eine Hypothesenprüfung im eigentlichen Sinn handelt, werden relevante Testgütekriterien anhand von ROC-Analysen exploriert und deskriptiv beschrieben.

Methode

Stichprobe und Vorgehen

Die Gesamtstichprobe bestand aus 144 Schülerinnen und Schülern (48.6% männlich) aus 8 bayerischen Grundschulklassen. An der Untersuchung nahmen 37 Kinder in Jahrgangsstufe 2, 52 Kinder in Jahrgangsstufe 3 und 55 Kinder in Jahrgangsstufe 4 teil. Insgesamt erhielten 172 Eltern im Vorfeld ein Elternanschreiben mit Informationen über die Studie, die Teilnahmequote entsprach damit 83.7%. Da die Studie vollständig anonymisiert durchgeführt wurde, wurden außer dem Geschlecht keine weiteren Informationen über die teilnehmenden Kinder erhoben. Der Anteil von Jungen und Mädchen lag in allen Jahrgangsstufen annähernd bei 50% und es lagen keine signifikanten Verteilungsunterschiede vor.

Die Datenerhebung erfolgte 2.5 Monate nach Schuljahresbeginn im Dezember 2018. Geleitet wurden die Untersuchungen durch zuvor geschulte studentische Hilfskräfte. Im Gruppensetting absolvierten die Schülerinnen und Schüler nach einer kurzen allgemeinen Instruktion durch die Testleiterinnen zunächst den Fehleridentifikationstest am Tablet. Im Anschluss wurden die Subtests Wortverständnis und Satzverständnis des *Leseverständnistests für Erst- bis Siebtklässler* (ELFE II; Lenhard et al., 2017) durchgeführt. Abschließend bearbeiteten die Schülerinnen und Schüler das Lückendiktat des DERET als Paper-Pencil-Test. Die Untersuchungen waren innerhalb einer Schulstunde abgeschlossen.

Eingesetzte Messinstrumente

Die produktive Rechtschreibleistung wurde anhand von Lückendiktaten aus dem Deutschen Rechtschreibtest (DERET) erhoben (jeweils Formen A und B). Da die Datenerhebung etwa zwei Monate nach Schuljahresbeginn stattfand, die Texte des DERET jedoch für das Ende des jeweiligen Schuljahres ausgelegt sind, kamen folgende Versionen zum Einsatz: In Jahrgangsstufe 2 die Testform für die Jahrgangsstufe 1 des *Deutschen Rechtschreibtests für das erste und zweite Schuljahr* (DERET 1-2+; Stock & Schneider, 2008a), in Jahrgangsstufe 3 die Testform für die Jahrgangsstufe 2 des DERET 1-2+ und in Jahrgangsstufe 4 die Testform für die Jahrgangsstufe 3 des *Deutschen Rechtschreibtests für das dritte und vierte Schuljahr* (DERET 3-4+; Stock & Schneider, 2008b). Aufgabe der

Schülerinnen und Schüler war es, die fehlenden Wörter in diktierten Sätzen zu ergänzen: insgesamt 12 in Jahrgangsstufe 2, 24 in Jahrgangsstufe 3 und 28 in Jahrgangsstufe 4. Die Testdurchführung nahm etwa 15 Minuten in Anspruch. Der Testrohwert entsprach der Anzahl der korrekt geschriebenen Wörter.

Als weiteres Maß der Rechtschreibkompetenzen wurde ein neu entwickelter, computerbasierter Fehleridentifikationstest durchgeführt, der auf Tablets präsentiert wurde (10.1 Zoll). Aufgabe der Schülerinnen und Schüler war es, fehlerhafte Wörter in einem Fließtext zu identifizieren. Hierfür wurden insgesamt drei Texte präsentiert (Jahrgangsstufe 2: 237 Wörter; Jahrgangsstufen 3 und 4: 284 Wörter), von denen sich der erste und dritte Text im Hinblick auf Länge und zu identifizierende Fehler zwischen den Jahrgangsstufen unterschieden. Nachdem die Geschichten zunächst jeweils einmal am Stück über Kopfhörer auditiv präsentiert worden waren, wurden die Sätze anschließend einzeln visuell dargeboten. Durch Antippen wurden die zuvor in schwarzer Schrift dargestellten Wörter blau gefärbt. Wurde ein Wort versehentlich markiert, konnte dies durch ein zweites Antippen wieder rückgängig gemacht werden. Gewertet wurde die letztgültige Markierung. Für die vorliegende Arbeit wurden lediglich die ersten 8 von 10 Sätzen des zweiten Texts („Ein Besuch im Zoo“) betrachtet, da diese für die Jahrgangsstufen 2 bis 4 identisch präsentiert wurden. Der relevante Abschnitt besteht somit aus acht Sätzen mit 61 Wörtern. In jeden Satz sind ein bis drei Fehler eingebaut, die von den Kindern erkannt werden müssen. Bezogen auf die acht Sätze handelt es sich um 17 Fehlschreibungen, die in der folgenden Analyse betrachtet werden. Da korrekt geschriebene Wörter von den Kindern nur extrem selten als Fehlschreibung markiert wurden, werden diese nicht weiter berücksichtigt. Im Hinblick auf die Antwortakkuratheit der 17 zu identifizierenden Fehler, die auf Itemebene erfasst wurde (fehlerhaftes Wort markiert vs. nicht markiert), ergab sich trotz der geringen Anzahl an zu identifizierender Fehler eine interne Konsistenz, die als gut zu bewerten ist (McDonalds ω = .81; vgl. McNeish, 2018).

Die Konstruktion des Stimulusmaterials ist an den Fehlertypen orientiert, die in bereits vorhandenen Fehleridentifikationstests (z.B. M. Schneider et al., 2014) zur Anwendung kamen. Hierzu gehörten unter anderem Auslassungen von Buchstaben, Ersetzungen durch falsche Buchstaben, überflüssige Buchstaben und Fehler in der Groß- und Kleinschreibung. Das Textmaterial wurde mit dem Grundwortschatz von Kindern abgeglichen. Alle Zielwörter waren im Basiswortschatz von Grundschulkindern enthalten. Ein Abgleich erfolgte mit der Onlinedatenbank childLex (Schroeder, Würzner, Heister,

Geyken & Kliegl, 2014). Darüber hinaus konnten die Fehlertypen in zwei übergeordnete Kategorien unterteilt werden: „Verstöße gegen die phonologische Wortform“ (9 Items) und „Verstöße gegen Rechtschreibregeln“ (8 Items). Da in der vorliegenden Arbeit nur eine geringe Anzahl an Zielwörtern untersucht wurde, wurden die Zielwörter nicht in feinere Unterkategorien aufgeschlüsselt, sondern ausschließlich diesen beiden Kategorien zugeordnet. Jeder Satz konnte bis zu zwei Fehler enthalten.

Datenanalyse

Für jedes zu identifizierende Zielwort lagen Antwortakkuratheit sowie Reaktionszeit vor. Um beide Informationen im Antwortverhalten von Kindern bestmöglich zu nutzen, wurde zudem ein integrierter Gesamtrohwert berechnet (Formel 1).

$$\text{EFF} = \frac{1}{k}\sum_{j=1}^{k}\left(Antwortakkuratheit(j) \cdot \frac{m_{RT\_j}}{ln(RT_j)}\right) \qquad (1)$$

Formel 1. Formel zur Berechnung des Effizienzwertes im Fehleridentifikationstest. Die Antwortakkuratheit bezeichnet, ob eine Person Wort j korrekt als Fehler eingestuft hat (0 = falsch, 1 = richtig), m_{RT_j} bezeichnet den Durchschnitt der ln-transformierten Bearbeitungsdauer in ms der Referenzgruppe bei Aufgabe j und $ln(RT_j)$ die ln-transformierte, benötigte Arbeitszeit der Person bei dieser Aufgabe.

Der auf diese Weise ermittelte Gesamtwert EFF im Fehleridentifikationstest nimmt demnach üblicherweise einen Wert zwischen 0 und 1 an. Ein Wert von 0 resultiert nur dann, wenn keines der 17 Zielworte identifiziert wurde, während ein Wert von 1 bedeutet, dass entweder alle 17 Zielworte korrekt erkannt wurden, und das zudem in einer Geschwindigkeit, die dem Durchschnitt der Referenzgruppe entspricht, oder aber dass wenige nicht identifizierte Zielworte durch eine besonders effiziente Arbeitsweise kompensiert wurden.

Um eine Zunahme der Höhe der Korrelationen zwischen Lückendiktat und Fehleridentifikationstest über die Jahrgangsstufen hinweg zu prüfen, wurden statistische Vergleiche der Korrelationskoeffizienten aus unabhängigen Stichproben berechnet (vgl. Eid, Gollwitzer & Schmitt, 2017).

Zur Testung der Hypothesen 1c, 2a und 2b wurden linear-gemischte Modelle mit der Effizienz als abhängiger Variable sowie generalisiert-gemischte Modelle mit Logit-Link mit der Antwortrichtigkeit als abhängiger Variable geschätzt. Als Prädiktoren wurden auf

Itemebene die Fehlerkategorie (dummykodiert; Referenzkategorie „Verstöße gegen die phonologische Wortform“ versus „Verstöße gegen Rechtschreibregeln“) und auf Personenebene die Jahrgangsstufe und die Gruppenzuordnung nach dem DERET (dummykodiert; Referenzkategorie „Kontrollgruppe“ versus „rechtschreibschwach“) in das Modell aufgenommen. Die Jahrgangsstufe wurde um die mittlere Jahrgangsstufe zentriert, was einem Wert von 3 entsprach. Es wurde jeweils ein Modell mit gekreuzten Zufallseffekten (random intercepts) für die Versuchspersonen und Items geschätzt. Die Modelle hatten die folgende Form:

$$\begin{aligned} EFF = {} & b_{ij} + b_{1j}\textit{(Fehlerkategorie)}_{ij} + \\ & b_{i1}\textit{(Gruppe nach DERET)}_{ij} + b_{i2}\textit{(Klassenstufe)}_{ij} + \\ & b_{12}\textit{(Fehlerkategorie X Klassenstufe)}_{ij} + r_{ij} \end{aligned} \quad (2a)$$

$b_{ij} = g_{00} + u_{0j} + v_{i0}$ Zufallskoeffizient, u_{0j}: Personenparameter, v_{i0}: Itemparameter

$b_{1j} = g_{10}$ } Feste Koeffizienten: Effekte von Itemeigenschaften

$b_{i1} = g_{01}$ } Feste Koeffizienten: Effekte von Personeneigenschaften

$b_{i2} = g_{02}$ }

$b_{12} = g_{12}$ } Feste Koeffizienten: Effekte von Interaktionstermen

$$\begin{aligned} \text{logit}(\textit{Antwortakkuratheit}_{ij}) = {} & b_{ij} + b_{1j}\textit{(Fehlerkategorie)}_{ij} + \\ & b_{i1}\textit{(Gruppe nach DERET)}_{ij} + b_{i2}\textit{(Klassenstufe)}_{ij} + \\ & b_{12}\textit{(Fehlerkategorie X Klassenstufe)}_{ij} \end{aligned} \quad (2b)$$

$b_{ij} = g_{00} + u_{0j} + v_{i0}$ Zufallskoeffizient, u_{0j}: Personenparameter, v_{i0}: Itemparameter

$b_{1j} = g_{10}$ } Feste Koeffizienten: Effekte von Itemeigenschaften

$b_{i1} = g_{01}$ } Feste Koeffizienten: Effekte von Personeneigenschaften

$b_{i2} = g_{02}$ }

$b_{12} = g_{12}$ } Feste Koeffizienten: Effekte von Interaktionstermen

Formel 2. Formeln für die linear-gemischten Modelle für die Effizienz EFF (2a) sowie der generalisierten linear-gemischten Modelle für die Antwortakkuratheit (2b) mit gekreuzten Zufallseffekten.

Alle Modelle wurden mit dem Softwarepaket lme4 (Bates, Maechler, Bolker & Walker, 2020; Version 1.1-23) für R (Version 4.0.2) geschätzt. Für die Signifikanztests kam lmerTest (Kuznetsova, Brockhoff & Christensen, 2020; Version 3.1-2) zum Einsatz. Zur Parameterschätzung wurde die Maximum-Likelihood-Methode (ML) verwendet. Für alle Signifikanztests wurde eine Typ-I-Irrtumswahrscheinlichkeit von .05 festgesetzt; die gerichteten Hypothesen wurden einseitig getestet.

Ergebnisse

Deskriptive Statistiken

Die erzielten Rohwertpunkte im Lückendiktat sowie erzielte Rohwertpunkte und durchschnittliche Effizienzwerte im Fehleridentifikationstest sind in Tabelle 1 dargestellt. Die maximal möglichen Rohwertpunkte im Lückendiktat unterscheiden sich zwischen den Jahrgangsstufen, aber die Rohwertpunkte (Summe der Antwortakkuratheit) und Effizienzscores im Fehleridentifikationstest können direkt verglichen werden. In Jahrgangsstufe 2 identifizierten Kinder durchschnittlich nur etwa die Hälfte der präsentierten Fehler (9 von 17). Dagegen waren es in Jahrgangsstufe 4 bereits über zwei Drittel (12 von 17). Ähnliche Leistungssteigerungen zeigen sich auch im Hinblick auf die Effizienz der Arbeitsweise ($EFF_{Jahrgangsstufe2} = 0.51$; $EFF_{Jahrgangsstufe3} = 0.58$; $EFF_{Jahrgangsstufe4} = 0.72$). Die verbesserte Leistung ist dabei insbesondere auf eine zuverlässigere und effizientere Identifikation von Fehlern zurückzuführen, die Verstöße gegen Rechtschreibregeln beinhalten: Einem durchschnittlichen Effizienzwert von 0.19 in Jahrgangsstufe 2 steht ein Wert von 0.53 in Jahrgangsstufe 4 gegenüber. In Jahrgangsstufe 2 wird durchschnittlich nur etwa einer von fünf präsentierten Fehlern der Kategorie „Verstöße gegen Rechtschreibregeln“ identifiziert (1.54 von 8; 19%), während es in Jahrgangsstufe 4 durchschnittlich mehr als die Hälfte sind (4.15 von 8; 52%). Nichtsdestotrotz gibt es in allen drei untersuchten Jahrgangsstufen Kinder, die nicht einen einzigen präsentierten Fehler dieser Kategorie identifizieren. Im Hinblick auf Fehler der Kategorie „Verstöße gegen die phonologische Wortform“ zeigt sich hingegen, dass bereits in Jahrgangsstufe 2 alle Kinder in der Lage sind, einen wesentlichen Anteil an Fehlern effizient zu identifizieren (*Min*(EFF) = 0.52).

– Tabelle 1 hier einfügen –

– Abbildung 1 hier einfügen –

Abbildung 1 zeigt die Entwicklung der effizienten Identifikation von Fehlern für vier Leistungsgruppen im DERET (Quartile). Fehler der Kategorie „Verstöße gegen die

phonologische Wortform“ werden von den Schülerinnen und Schülern mit guten Rechtschreibkompetenzen (= oberes Quartil im DERET, Q4) in allen Jahrgangsstufen zuverlässig und effizient erkannt, sodass hier Deckeneffekte keine weitere Verbesserung erlauben (s. Abb. 1c). Demgegenüber ist im Hinblick auf die Gesamtleistung (s. Abb. 1a) und die Identifikation von Fehlern, welche Verstöße gegen Rechtschreibregeln beinhalten (s. Abb. 1b), eine Leistungssteigerung in allen Leistungsgruppen zu beobachten. Besonders deutlich zeichnet sich die Entwicklung für die beiden Gruppen der Schülerinnen und Schüler mit überdurchschnittlichen Rechtschreibkompetenzen ab (Quartile 3 und 4 im DERET), während die schwächsten 25% weiter zurückfallen. Erst in Jahrgangsstufe 4 arbeiten Schülerinnen und Schüler mit schwachen Rechtschreibleistungen im Lückendiktat (Quartil 1 im DERET; EFF = 0.54) durchschnittlich vergleichbar effizient im Fehleridentifikationstest wie Kinder mit unauffälligen Rechtschreibleistungen bereits in der Jahrgangsstufe 2 (EFF = 0.57).

Im Hinblick auf die Identifikation von leistungsschwachen Schülerinnern und Schülern (Q1) erscheint darüber hinaus gegen Ende der Grundschulzeit die Fehlerkategorie „Verstöße gegen die phonologische Wortform“ als aussichtsreich. Während Kindern der oberen drei Quartile das Erkennen von phonologischen Fehlern effizient gelingt und sie sich daher kaum unterscheiden, fällt diese Leistung der Gruppe des unteren Quartils weiterhin schwer.

Zusammenhänge zwischen Lückendiktat und Fehleridentifikationstest

Für die Gesamtstichprobe zeigte sich im Sinne von Hypothese 1a eine hohe Korrelation zwischen den jeweils jahrgangsstufenweise *z*-transformierten Ergebnissen im Lückendiktat und im Fehleridentifikationstest (r = .745; vgl. Abb. 2). Die Zusammenhänge zwischen den Rohwerten im Lückendiktat und im Fehleridentifikationstest stiegen deskriptiv über die Jahrgangsstufen hinweg an, von .693 in der zweiten Jahrgangsstufe über .705 in der 3. Jahrgangsstufe hin zu .818 in der vierten Jahrgangsstufe. Keiner der Unterschiede zwischen den Jahrgangsstufen war allerdings statistisch signifikant, auch wenn sich Tendenzen abzeichneten für Unterschiede zwischen Jahrgangsstufe 2 und 4 (z = -1.35, p = .089) sowie zwischen Jahrgangsstufe 3 und 4 (z = -1.37, p = .085). Hypothese 1b konnte also nicht gestützt werden.

– Tabelle 2 hier einfügen –

Leistung der Fehleridentifikation nach Fehlerkategorie, Jahrgangsstufe und Leistung im Lückendiktat

In Tabelle 2 sind die Parameterschätzungen der festen Effekte und der Varianzkomponenten der Modelle für die Antwortrichtigkeit und die Effizienz über die drei untersuchten Jahrgangsstufen hinweg dargestellt. Wie in Hypothese 1c angenommen, identifizierten rechtschreibschwache Kinder signifikant weniger Fehler im Fehleridentifikationstest als Kinder mit zumindest durchschnittlich ausgebildeter Rechtschreibkompetenz, $\beta = -1.24$, $z = -5.45$, $p < .001$, und sie arbeiteten zudem weniger effizient, $\beta = -0.18$, $t(143) = -5.85$, $p < .001$. Darüber hinaus wurden, wie in Hypothese 2a angenommen, Fehler der Kategorie „Verstöße gegen die phonologische Wortform" signifikant häufiger erkannt, $\beta = -3.05$, $z = -5.32$, $p < .001$, und effizienter bearbeitet als Fehler der Kategorie „Verstöße gegen Rechtschreibregeln", $\beta = -0.48$, $t(17) = -5.66$, $p < .001$. Auch die in Hypothese 2b angenommene Interaktion zwischen Jahrgangsstufe und Fehlerkategorie war signifikant, und zwar sowohl im Modell für die Antwortrichtigkeit, $\beta = 0.82$, $z = 5.39$, $p < .001$, als auch im Modell für die Effizienz, $\beta = 0.12$, $t(2288) = 6.62$, $p < .001$. Abbildung 2 bildet die Unterschiede in geschätzten Wahrscheinlichkeiten für eine korrekte Antwort und Unterschiede in der geschätzten Effizienz ab und illustriert damit die signifikanten Interaktionseffekte für die Antwortrichtigkeit (s. Abb. 2a) und die Effizienz (s. Abb. 2b): Während Grundschulkinder zu Beginn der Jahrgangsstufe 2 nur sehr wenige Fehler der Kategorie „Verstöße gegen Rechtschreibregeln" identifizierten, gelang ihnen diese Leistung im Verlauf der Grundschule zunehmend besser. Da demgegenüber bereits in Jahrgangsstufe 2 Fehler mit Verstößen gegen die phonologische Wortform zuverlässig und effizient erkannt wurden, konnte hier keine bedeutsame Leistungssteigerung stattfinden.

– Abbildung 2 hier einfügen –

Identifikation von rechtschreibschwachen Kindern

Abschließend wurde der Frage nachgegangen, wie zuverlässig rechtschreibschwache Kinder – identifiziert anhand einer schwachen Leistung im Lückendiktat (1 *SD* unterhalb des Mittelwerts der jeweiligen Jahrgangsstufe) – mithilfe des Fehleridentifikationstests erkannt werden. In Abbildung 3 wird ersichtlich, dass sich der Zusammenhang der Leistungen im Lückendiktat und im Fehleridentifikationstest anhand von drei Gruppen von Kindern beschreiben lässt:

(1) Kinder, die hinreichend gute Leistung im Kriterium ($z > -1$ im DERET) bei einer über dem Durchschnitt liegenden Leistung im Prädiktor ($z > 0$ im FIT) erzielen ("Richtig Negative").
(2) Kinder, die hinreichend gute Leistung im Kriterium ($z > -1$ im DERET) bei einer unter dem Durchschnitt liegenden Leistung im Prädiktor ($z < 0$ im FIT) erzielen ("Falsch Positive").
(3) Kinder, die schwache Leistung im Kriterium ($z < -1$ im DERET) bei einer unter dem Durchschnitt liegenden Leistung im Prädiktor ($z < 0$ im FIT) erzielen („Richtig Positive“).

Eine Gruppe mit falsch negativem Ergebnis fehlt hingegen nahezu vollständig (4). Somit existieren so gut wie keine Schülerinnen und Schüler, die effizient im Fehleridentifikationstest arbeiten, obwohl sie eine schwache Rechtschreibleistung im Lückendiktat vorweisen.

– Abbildung 3 hier einfügen –

Abbildung 4 stellt die geschätzte Wahrscheinlichkeit einer korrekten Antwort nach Jahrgangsstufen und Gruppeneinteilung für unterschiedliche Fehlerkategorien dar. Dabei wird deutlich, dass für die Identifikation rechtschreibschwacher Kinder (DERET-*T*-Wert ≤ 40) in verschiedenen Jahrgangsstufen unterschiedliche Fehlerarten besonders relevant sind. Während sich Fehler der Kategorie „Verstöße gegen Rechtschreibregeln“ für die Jahrgangsstufen 3 und 4 als ein gutes Unterscheidungskriterium darstellen, gilt das für Jahrgangsstufe 2 nicht (s. Abb. 4a und 4b). Stattdessen unterscheiden sich Schülerinnen und Schüler mit schwachen Rechtschreibleistungen hier in der Fähigkeit, Fehler der Kategorie „Verstöße gegen die phonologische Wortform“ zu identifizieren (s. Abb. 4c).

– Abbildung 4 hier einfügen –

Schließlich verdeutlichen ROC-Kurven, dass bei einer Spezifität (Anteil der richtig vorhergesagten Unauffälligen) von etwa 80% in allen untersuchten Jahrgangsstufen eine *Sensitivität* (Anteil der durch das Screening korrekt entdeckten Problemkinder) von über 80% erreicht wird (s. Abb. 5). Tabelle 3 stellt die Güteindizes für die Vorhersage unterdurchschnittlicher Rechtschreibleistung im Lückendiktat durch die Effizienz im Fehleridentifikationstest für die Jahrgangsstufen 2 bis 4 dar. Bei einer Spezifität von 84% werden drei Viertel der rechtschreibschwachen Kinder durch den Fehleridentifikationstest korrekt identifiziert. Der RATZ-Index von 68% deutet auf eine gute Klassifikation hin (H. Marx, 1992).

– Tabelle 3 hier einfügen –

– Abbildung 5 hier einfügen –

Diskussion

Ziel der vorliegenden Arbeit war es, das Potenzial eines tabletbasierten Fehleridentifikationstests zur Erfassung der Rechtschreibleistungen in der Grundschule zu untersuchen. Hohe Korrelationen zwischen der Leistung im Lückendiktat und effizienter Arbeitsweise im Fehleridentifikationstest bereits ab Beginn der Jahrgangsstufe 2 unterstreichen, dass die beiden Fähigkeiten eng miteinander verknüpft sind. Die Ergebnisse fügen sich nahtlos ein in Befunde aus dem englischen Sprachraum (Allred, 1984; Frisbie & Cantor, 1995; Westwood, 1999) sowie dem deutschen Sprachraum für den Sekundarbereich (Lenhart et al., 2019; M. Schneider et al., 2014). Auch wenn die gefundenen Korrelationen nicht als perfekt zu bewerten sind, so bewegen sie sich mit einer Größenordnung von .7 bis .8 doch ein einem Bereich, welcher etwa der kriterienbezogenen Validität von standardisierten Diktaten entspricht und der Retestreliabilität von produktiven Rechtschreibtests entspricht. Die Übereinstimmung zwischen DERET und anderen standardisierten Tests zur Erfassung von Rechtschreibleistung wird für den DERET 1-2+ mit $r = .63$ bis $r = .82$ angegeben, für den DERET 3-4+ mit $r = .64$ bis $r = .83$. Darüber hinaus existieren kaum Ausreißerwerte (vgl. Abb. 2): So gibt es nur vereinzelt Kinder, die bei einer zumindest durchschnittlichen Leistung in einem der beiden Testverfahren (z-Wert > 0) im anderen Testverfahren nur eine unterdurchschnittliche Leistung erzielen (z-Wert < -1). Die geringe Anzahl an falsch negativen Kindern ist im Hinblick auf den Einsatz des Fehleridentifikationstests als Screeningverfahren besonders ermutigend: Rechtschreibschwachen Kindern (definiert anhand ihrer Leistung im Lückendiktat) ist es nahezu unmöglich, im Fehleridentifikationstest ausreichend effizient zu arbeiten. Neben der Erklärung, dass diesen Kindern die korrekte Schreibweise oftmals nicht bekannt ist, ist es auch vorstellbar, dass sie auf dem Weg zu ihrem Urteil mehr Zeit investieren müssen, da ihnen die direkte Route aufgrund fehlender Repräsentationen im orthografischen Lexikon versperrt ist (vgl. Houghton & Zorzi, 2003).

In Anbetracht der Einsatzmöglichkeit von Fehleridentifikationstests als Screeningverfahren finden sich weitere ermutigende Befunde: Die in dieser Studie berichteten Güteindizes zur Vorhersage von rechtschreibschwachen Kindern anhand ihrer Leistung im Fehleridentifikationstest sind als solide (SE = 75%, SP = 84%, RATZ = 68%), bei Inkaufnahme einer geringeren Spezifität sogar als sehr gut einzuschätzen: Wird als

Prädiktor das untere Leistungsdrittel gewählt, so ergibt sich bei einer Spezifität von 79% bereits eine Sensitivität von 86% (RATZ = 79%). Auf diese Weise würde nur noch ein kleiner Prozentsatz rechtschreibschwacher Kinder übersehen werden, was genau der Intention eines Screeningverfahrens entspricht. Es stellt sich angesichts dieser Ergebnisse sogar die Frage, ob nicht der Fehleridentifikationstest und der daraus abgeleitete Effizienzkennwert eine validere Erfassung der Rechtschreibkompetenz ermöglicht, da er nicht nur die Genauigkeit abbildet, sondern auch die Geschwindigkeit, mit der eine Person auf orthografische Muster zurückgreift. Analog zur Theorie der lexikalischen Qualität (Perfetti & Hart, 2002) für Leseprozesse könnte im Einklang mit dem Zwei-Wege-Modell des Rechtschreibens (Houghton & Zorzi, 2003) davon ausgegangen werden, dass umso mehr Ressourcen für Prozesse auf hierarchiehöheren Verarbeitungsebenen zur Verfügung stehen, je automatisierter der Zugriff auf das orthografische Lexikon erfolgt. Geübte Schreiberinnen und Schreiber sind demnach jene Kinder, die ein fehlerhaft verschriftlichtes Wort auf Anhieb zuverlässig als solches identifizieren, ohne ihr Wissen über Rechtschreibregeln konsultieren zu müssen. In der Folge könnte sich in Alltagssituationen, in denen häufig unter Zeitdruck geschrieben werden muss, der EFF-Wert als prädiktiver erweisen als etwa Ergebnisse aus einem Diktat.

Darüber hinaus lohnt ein detaillierterer Blick in Fehleridentifikationskompetenzen nach Fehlerkategorien. So sollte die Kompetenz, Wortfehler mit Verstößen gegen die phonologische Wortform zuverlässig zu erkennen, spätestens am Ende der Grundschulzeit solide ausgebildet sein. Tatsächlich erreichten alle Kinder mit einem Prozentrang über 25 im DERET durchschnittliche Effizienzwerte in dieser Kategorie von über EFF = .90 (s. Abb. 1c). Demgegenüber haben Kinder, welche diese vermeintlich sehr leichten Aufgaben nicht ausreichend zuverlässig lösen können, mit einer sehr hohen Wahrscheinlichkeit tatsächlich Schwierigkeiten im Rechtschreiben: 14 von 15 Kinder mit einem DERET *T*-Wert ≤ 40 erzielten einen niedrigen Effizienz-Wert (SE = 93%), während dies nur auf 6 von 40 Kindern mit einem *T*-Wert größer als 40 zutrifft (SP = 85%). Bei einer isolierten Betrachtung der Fehlerkategorie „Verstöße gegen die phonologische Wortform“ am Ende der Jahrgangsstufe 4 ergeben sich demnach herausragende Güteindizes für die Vorhersage von Leistungsschwächen (RATZ = 90%).

Limitationen

Im Hinblick auf die Generalisierbarkeit der Ergebnisse der vorliegenden Studie sind jedoch einige Einschränkungen zu beachten. So beruht die Studie auf einer relativ kleinen

Stichprobe von Schülerinnen und Schülern, die darüber hinaus aus nur einem Bundesland stammt. Curriculare Unterschiede etwa im Hinblick auf den Beginn der Einführung von Rechtschreibregeln dürften zumindest zu einer zeitlichen Verschiebung der hier berichteten Übereinstimmungen in den jeweiligen Jahrgangsstufen führen. Da Arbeiten am Computer mit einem Risiko von Rapid Guessing einhergeht, sollten zukünftige Analysen darauf abzielen, nicht instruktionskonformes Antwortverhalten zu identifizieren. Zudem ist unklar, welchen Einfluss die tabletgestützte Darbietung hat, da bislang die Effekte von Darbietungsmedien bislang fast ausschließlich für den Vergleich Papierfassung versus computerbasierte Darbietung verfügbar sind.

Relevanz für die Praxis

Trotz dieser Einschränkungen ist auf Grundlage der hier berichteten Befunde die tabletbasierte Erhebung von Fehleridentifikationskompetenzen als gute Möglichkeit einzuschätzen, die Rechtschreibkompetenzen von Schülerinnen und Schülern der Jahrgangsstufen 2 bis 4 zuverlässig zu erfassen. Gegenüber der klassischen Vorgehensweise mit Fließ- oder Lückendiktaten bieten computergestützte Fehleridentifikationstests einen überzeugenden ökonomischen Vorteil. Hinzu kommt die Auswertungsmöglichkeit von Effizienzwerten, die wir in der vorliegenden Arbeit im Kontext von Rechtschreibleistungen erstmalig beschrieben haben, und die im Einklang mit wichtigen Theorien zum Schriftspracherwerb künftig möglicherweise Informationen über die die Fähigkeit der Rechtschreibkompetenz liefern können, die über die diagnostische Information von diktatbasierten Rechtschreibtests hinausgehen.

Literatur

Allen, D. & Ager, J. (1965). A factor analytic study of the ability to spell. *Educational and Psychological Measurement, 25*(1), 153–161. doi:10.1177/001316446502500117

Allred, R. A. (1984). Comparison of proofreading-type standardized spelling tests and written spelling test scores. *The Journal of Educational Research, 77*(5), 298–303. doi:10.1080/00220671.1984.10885544

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2020). *lme4: Linear mixed-effects models using 'Eigen' and S4. R package version 1.1-23*. http://CRAN.R-project.org/package=lme4.

Berger, N. (2010). *Mehr als nur ein Wort: Zur Diagnostik und Förderung von Grundschulkindern mit schwachen Rechtschreibleistungen im Rahmen des Regelunterrichts*. München: utzverlag.

Brinkmann, E. (2004). Schreiben nach Diktat oder selbstständig Rechtschreibung lernen? *Grundschule, 36*(1), 11–13.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). San Diego, CA: Academic Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256. doi:10.1037/0033-295X.108.1.204

Corvacho del Toro, I. & Hoffmann-Erz, R. (2014). Was ist lautgetreu? Zur Notwendigkeit einer begrifflichen Differenzierung. In K. Siekmann (Hrsg.), *Theorie, Empirie und Praxis effektiver Rechtschreibdiagnostik* (S. 29–40). Tübingen: Stauffenburg.

Croft, A. C. (1982). Do spelling tests measure the ability to spell? *Educational and Psychological Measurement, 42*(3), 715–723. doi:10.1177/001316448204200301

Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.

Fix, M. (2004). Funktionen des Diktats und Diktatkritik. *Grundschule, 36*(1), 8–10.

Freyberg, P. S. (1970). The concurrent validity of two types of spelling tests. *British Journal of Educational Psychology, 40*(1), 68–71. doi:10.1111/j.2044-8279.1970.tb02100.x

Frisbie, D. A. & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement, 32*(1), 55–78. doi:10.1111/j.1745-3984.1995.tb00456.x

Guiler, W. S. (1929). Validation of methods of testing spelling. *Journal of Educational Research, 20*(3), 181–189. doi:10.1080/00220671.1929.10879981

Houghton, G. & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology, 20*(2), 115–162. doi:10.1080/02643290242000871

Howard, S. J., Burianová, H., Calleia, A., Fynes-Clinton, S., Kervin, L. & Bokosmaty, S. (2017). The method of educational assessment affects children's neural processing and performance: Behavioural and fMRI Evidence. *npj Science of Learning, 2*(10), 1–7. doi:10.1038/s41539-017-0010-9

Klicpera, C., Schabmann, A. & Gasteiger-Klicpera, B. (1993). Lesen- und Schreibenlernen während der Pflichtschulzeit: Eine Längsschnittuntersuchung über die Häufigkeit und Stabilität von Lese- und Rechtschreibschwierigkeiten in einem Wiener Schulbezirk. *Zeitschrift für Kinder- und Jugendpsychiatrie, 21*(4), 214–225.

Kuhn, J.-T., Schwenk, C., Raddatz, J., Dobel, C. & Holling, H. (2017). *CODY-Mathetest für die 2.-4. Klasse (CODY-M 2-4)*. Düsseldorf: Kaasa health.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2020). *lmerTest: Tests for random and fixed effects for Linear Mixed Effect Models (Lmer objects of lme4 package). R Package version 3.1-2.* http://CRAN.R-project.org/package=lmerTest

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

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

Lenhart, J., Marx, P., Segerer, R. & Schneider, W. (2019). Rechtschreibung ohne Schreiben: Messen Fehleridentifikation und Diktat dasselbe? *Diagnostica, 65*(4), 216–227. doi:10.1026/0012-1924/a000229

Lenhart, J., Segerer, R., Marx, P. & Schneider, W. (im Druck). *Fehleridentifikationstest – Rechtschreibung für neunte und zehnte Klassen (R-FIT 9-10)*. Göttingen: Hogrefe.

Marx, H. (1992). Methodische und inhaltliche Argumente für und wider eine frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten. *Diagnostica, 38*(3), 249–268.

Marx, P. & Lenhard, W. (2010). Diagnostische Merkmale von Screeningverfahren. In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen*. Tests und Trends, N. F. (Bd. 9, S. 68–84). Göttingen: Hogrefe.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. doi:10.1037/met0000144

Perfetti, C. & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam, Niederlande: John Benjamin.

Richter, T., Lenhard, W., Marx, P. & Endlich, D. (2018). Konzeption eines Online-Screenings für Lernstörungen. *Lernen und Lernstörungen, 7*(4), 203–207. doi:10.1024/2235-0977/a000237

Richter, T., Naumann, J., Isberner, M.-B., Neeb, Y. & Knoepke, J. (2017). *ProDi-L: Prozessbasierte Diagnostik von Lesefähigkeiten bei Grundschulkindern*. Göttingen: Hogrefe.

Romani, C., Olson, A. & Di Betta, A. M. (2005). Spelling disorders. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 431–447). Oxford, UK: Blackwell.

Schneider, M., Martinez Méndez, R. & Hasselhorn, M. (2014). *Fehleridentifikationstest – Rechtschreibung für fünfte und sechste Klassen (R-FIT 5-6+)*. Göttingen: Hogrefe.

Schneider, W. (2008). *Entwicklung von der Kindheit bis zum Erwachsenenalter: Befunde der Münchner Längsschnittstudie LOGIK*. Weinheim: Beltz.

Schneider, W. (2009). The development of reading and spelling. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20-year longitudinal study* (pp. 199–220). New York, NY: Psychology Press.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (2014). childLex: A lexical database of German read by children. *Behavior Research Methods, 47*(4), 1085–1094. doi:10.3758/s13428-014-0528-1

Stock, C. & Schneider, W. (2008a). *Deutscher Rechtschreibtest für das erste und zweite Schuljahr (DERET 1-2+)*. Göttingen: Hogrefe.

Stock, C. & Schneider, W. (2008b). *Deutscher Rechtschreibtest für das dritte und vierte Schuljahr (DERET 3-4+)*. Göttingen: Hogrefe.

Westwood, P. (1999). The correlation between results from different types of spelling test and children's spelling ability when writing. *Australian Journal of Learning Disabilities, 4*(1), 31–36. doi:10.1080/19404159909546584

Legende

Abbildung 1. Entwicklung der effizienten Identifikation von Fehlern nach Leistungsgruppen im DERET (Quartile Q1–Q4) für den Gesamtscore (a), Fehler der Kategorie „Verstöße gegen Rechtschreibregeln“ (b) sowie Fehler der Kategorie „Verstöße gegen die phonologische Wortform“ (c) in den Jahrgangsstufen 2 bis 4 (Fehlerbalken sind Standardfehler).

Abbildung 2. Zusammenhang zwischen durchschnittlicher Effizienz im Fehleridentifikationstest und Lückendiktat in den Jahrgangsstufen 2 bis 4.

Abbildung 3. Interaktion der Fehlerart mit der Jahrgangsstufe für (a) die geschätzte Wahrscheinlichkeit einer korrekten Antwort und (b) die geschätzte Effizienz (Fehlerbalken sind Standardfehler).

Abbildung 4. Geschätzte Wahrscheinlichkeit einer korrekten Antwort nach Jahrgangsstufen und Gruppeneinteilung für (a) den Gesamtscore, (b) Fehler der Kategorie „Verstöße gegen Rechtschreibregeln“ sowie (c) Fehler der Kategorie „Verstöße gegen die phonologische Wortform“ (Fehlerbalken sind Standardfehler).

Abbildung 5. ROC-Kurven für die Vorhersage von rechtschreibschwachen Kindern (*T*-Wert ≤ 40 im Lückendiktat) anhand der Effizienz im Fehleridentifikationstest in den Jahrgangsstufen 2 (a), 3 (b) und 4 (c).

Tabelle 1

Mittelwerte, Standardabweichungen, Minima, Maxima und Korrelationen der Testverfahren zur Rechtschreibung nach Jahrgangsstufen

		M	*SD*	*Min*	*Max*	1	2	3	4
Jahrgangsstufe 2	1 Lückendiktat (Anzahl korrekter Wörter)	6.38	2.94	1	12	-			
	2 Akkuratheit im FIT (Anzahl identifizierter Fehler)	9.00	2.03	6	13	.648***	-		
	3 Effizienz im FIT (EFF)	0.51	0.12	0.33	0.75	.693***	.989***	-	
	4 Effizienz im FIT (Verstöße gegen die Lauttreue)	0.81	0.13	0.52	1.01	.549***	.829***	.834***	-
	5 Effizienz im FIT (Verstöße gegen Rechtschreibregeln)	0.19	0.15	0.00	0.51	.603***	.825***	.838***	.400*
Jahrgangsstufe 3	1 Lückendiktat (Anzahl korrekter Wörter)	12.79	4.37	4	21	-			
	2 Akkuratheit im FIT (Anzahl identifizierter Fehler)	10.00	2.73	4	15	.678***	-		
	3 Effizienz im FIT (EFF)	0.58	0.16	0.24	0.89	.705***	.993***	-	
	4 Effizienz im FIT (Verstöße gegen die Lauttreue)	0.80	0.18	0.33	1.06	.582***	.828***	.832***	-
	5 Effizienz im FIT (Verstöße gegen Rechtschreibregeln)	0.34	0.21	0.00	0.75	.609***	.849***	.857***	.427**
Jahrgangsstufe 4	1 Lückendiktat (Anzahl korrekter Wörter)	14.00	6.03	3	25	-			
	2 Akkuratheit im FIT (Anzahl identifizierter Fehler)	11.96	2.82	5	17	.800***	-		
	3 Effizienz im FIT (EFF)	0.72	0.18	0.31	1.02	.818***	.993***	-	
	4 Effizienz im FIT (Verstöße gegen die Lauttreue)	0.90	0.14	0.48	1.11	.582***	.734***	.761***	-
	5 Effizienz im FIT (Verstöße gegen Rechtschreibregeln)	0.53	0.28	0.00	1.09	.781***	.929***	.923***	.453**

Anmerkungen. FIT = Fehleridentifikationstest. Unterschiedliche Versionen des Lückendiktats (Jahrgangsstufe 2: max. 12 Punkte; Jahrgangsstufe 3: max. 24 Punkte; Jahrgangsstufe 4: max. 28 Punkte). EFF = Durchschnittliche Effizienz im Fehleridentifikationstest (vgl. Formel 1).
* $p < .05$, ** $p < .01$, *** $p < .001$.

Tabelle 2.

Feste Effekte und Varianzkomponenten im generalisiert linear-gemischten Modell für die Antwortrichtigkeit (logit-transformiert) sowie im linear-gemischten Modell für die Effizienz

Parameter	Antwortrichtigkeit	Effizienz
	β (*SE*)	β (*SE*)
	Feste Effekte	
Konstante	2.421 (0.41)***	0.867 (0.06)***
Fehlerkategorie	-3.048 (0.57)***	-0.479 (0.08)***
Gruppe	-1.237 (0.23)***	-0.177 (0.03)***
Fehlerkategorie X Jahrgangsstufe	0.820 (0.15)***	-0.122 (0.02)***
Jahrgangsstufe	0.196 (0.14)	0.042 (0.02)*
	Varianzkomponenten	
Versuchspersonen	0.697	0.013
Items	1.309	0.029

Anmerkungen. Die Jahrgangsstufe ist zentriert um den Wert 3. Fehlerkategorie dummykodiert (Verstöße gegen die Lauttreue = 0, Verstöße gegen Rechtschreibregeln = 1). Gruppe dummykodiert (Kontrollgruppe mit T > 40 im DERET = 0, Rechtschreibschwach T ≤ 40 im DERET = 1). * $p < .05$, *** $p < .001$ (zweiseitig).

$N = 2\,448$.

Tabelle 3.

Vorhersage unterdurchschnittlicher Rechtschreibleistung im Lückendiktat durch die Effizienz im Fehleridentifikationstest

		Kriterium: Lückendiktat	
		T-Wert ≤ 40	*T*-Wert > 40
Prädiktor: Effizienz in der Fehleridentifikation	PR ≤ 25	21[a]	18[b]
	PR > 25	7[c]	98[d]
Güteindizes	Sensitivität		75%
	Spezifität		84%
	Positiver prädiktiver Wert		54%
	RATZ-Index		68%

Anmerkungen: Die *T*-Werte wurden anhand der vorliegenden Stichprobe ermittelt.

[a]richtig positiv; [b]falsch positiv; [c]falsch negativ; [d]richtig negativ.

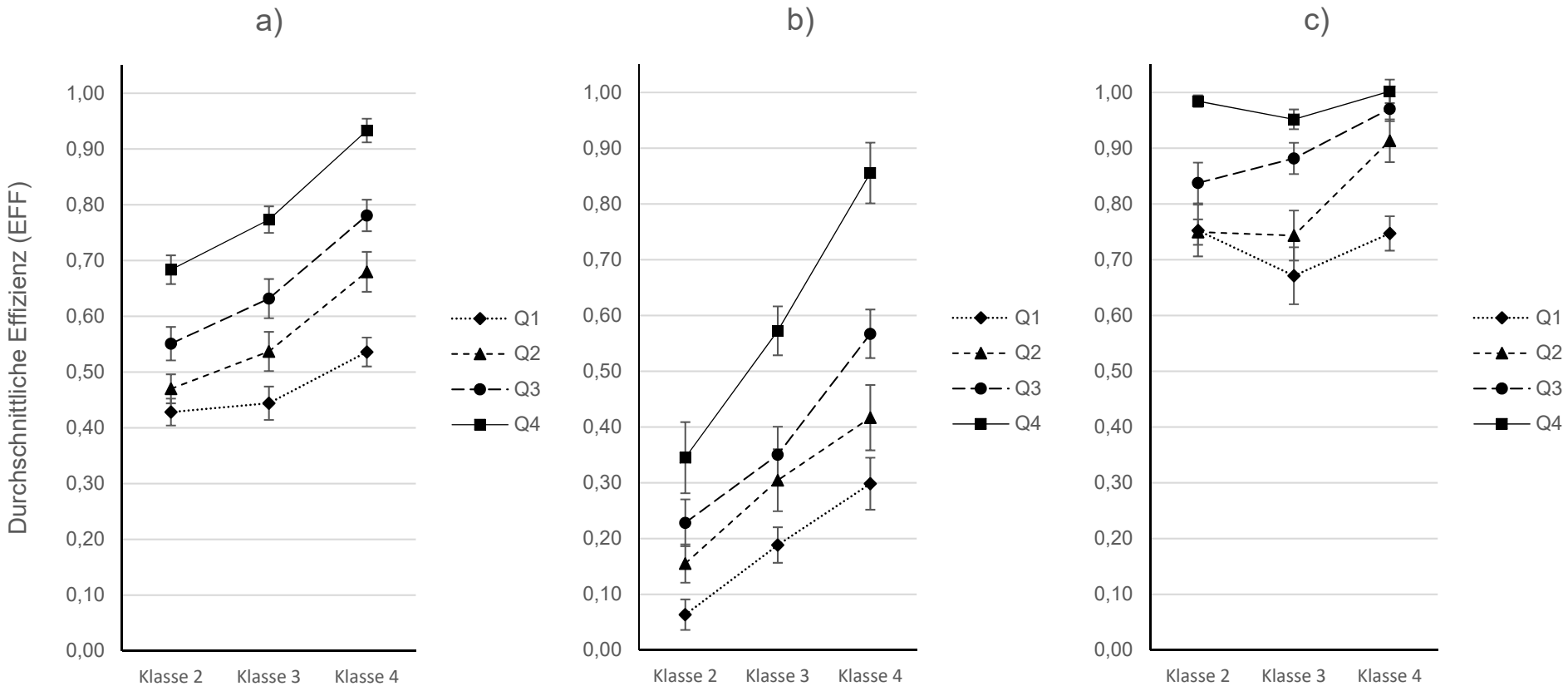

Abbildung 1. Entwicklung der effizienten Identifikation von Fehlern nach Leistungsgruppen im DERET (Quartile Q1–Q4) für den Gesamtscore (a), Fehler der Kategorie „Verstöße gegen Rechtschreibregeln“ (b) sowie Fehler der Kategorie „Verstöße gegen die Lauttreue“ (c) in den Jahrgangsstufen 2 bis 4 (Fehlerbalken sind Standardfehler).

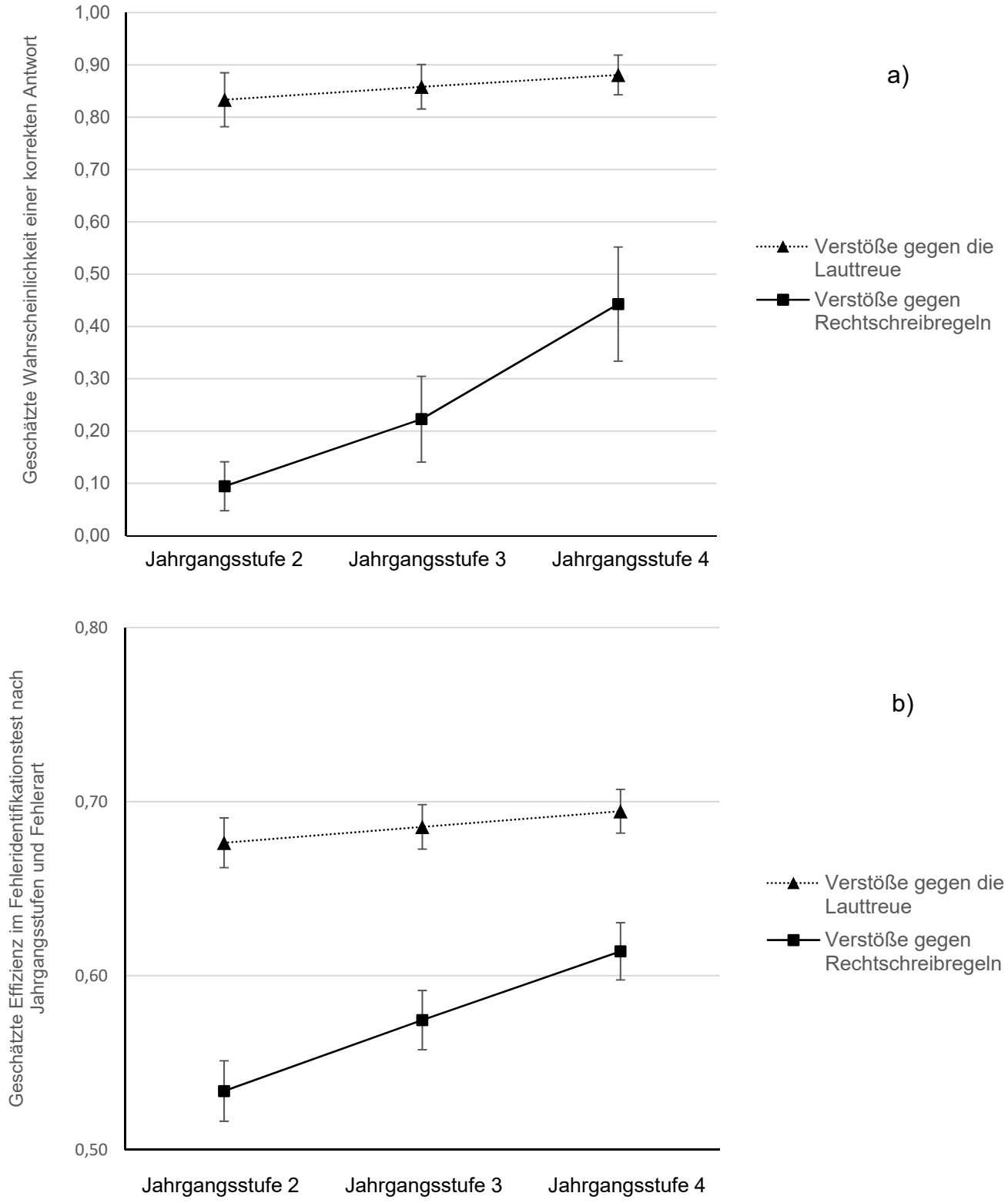

Abbildung 2. Interaktion der Fehlerart mit der Jahrgangsstufe für (a) die geschätzte Wahrscheinlichkeit einer korrekten Antwort und (b) die geschätzte Effizienz (Fehlerbalken sind Standardfehler).

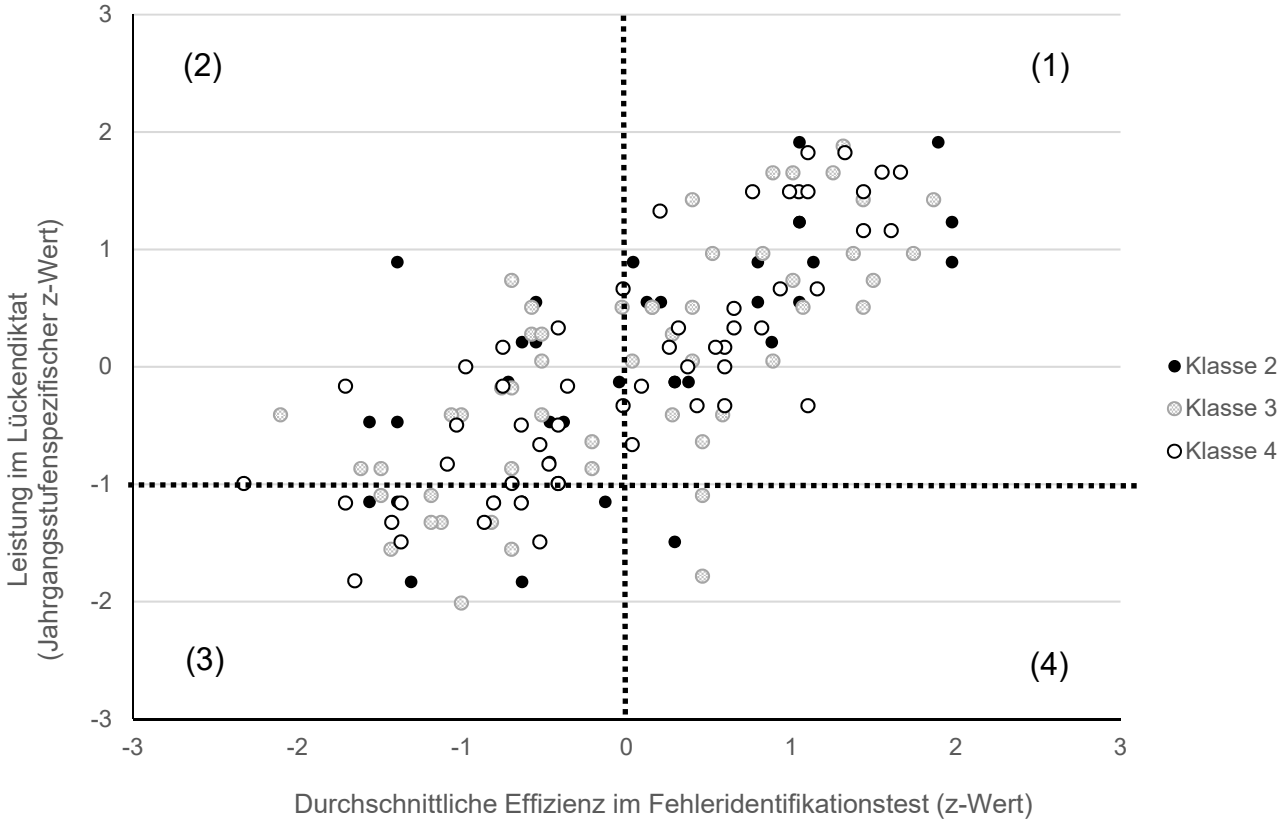

Abbildung 3. Zusammenhang zwischen durchschnittlicher Effizienz im Fehleridentifikationstest und Lückendiktat in den Jahrgangsstufen 2 bis 4.

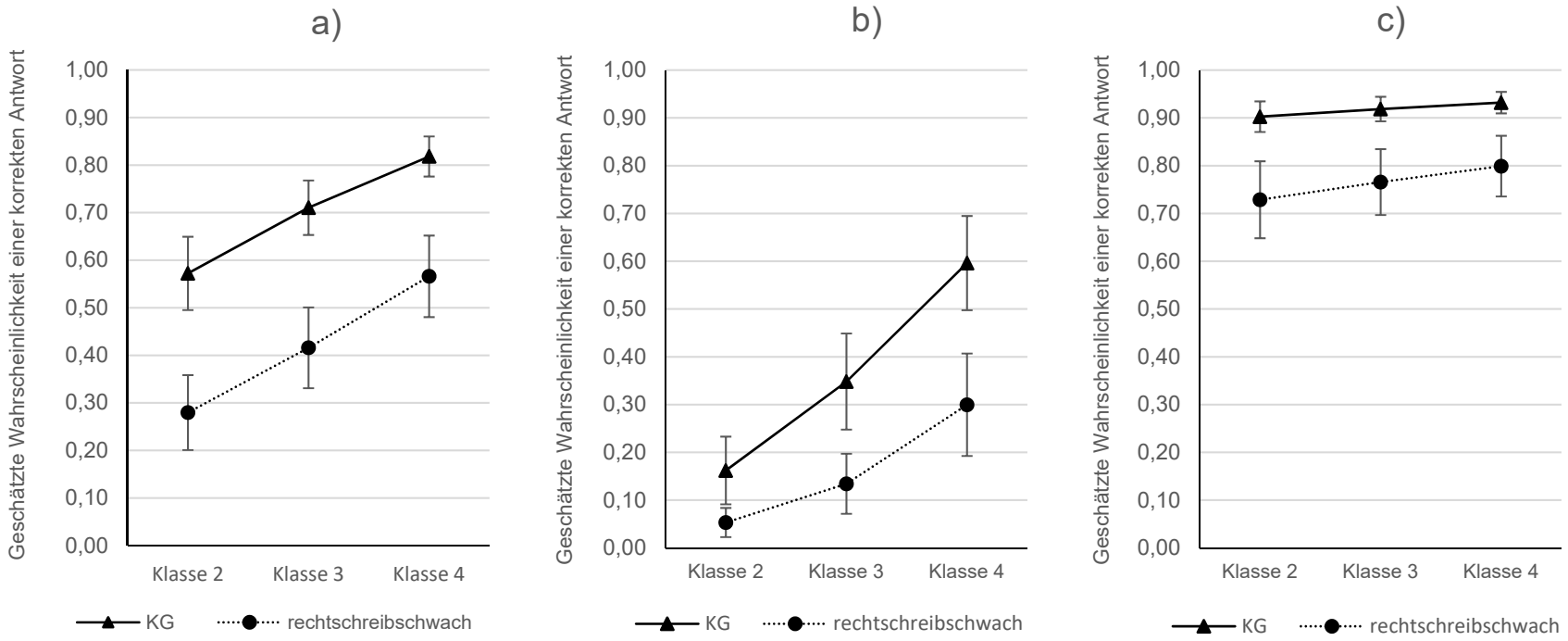

Abbildung 4. Geschätzte Wahrscheinlichkeit einer korrekten Antwort nach Jahrgangsstufen und Gruppeneinteilung für (a) den Gesamtscore, (b) Fehler der Kategorie „Verstöße gegen Rechtschreibregeln“ sowie (c) Fehler der Kategorie „Verstöße gegen die Lauttreue“ (Fehlerbalken sind Standardfehler).

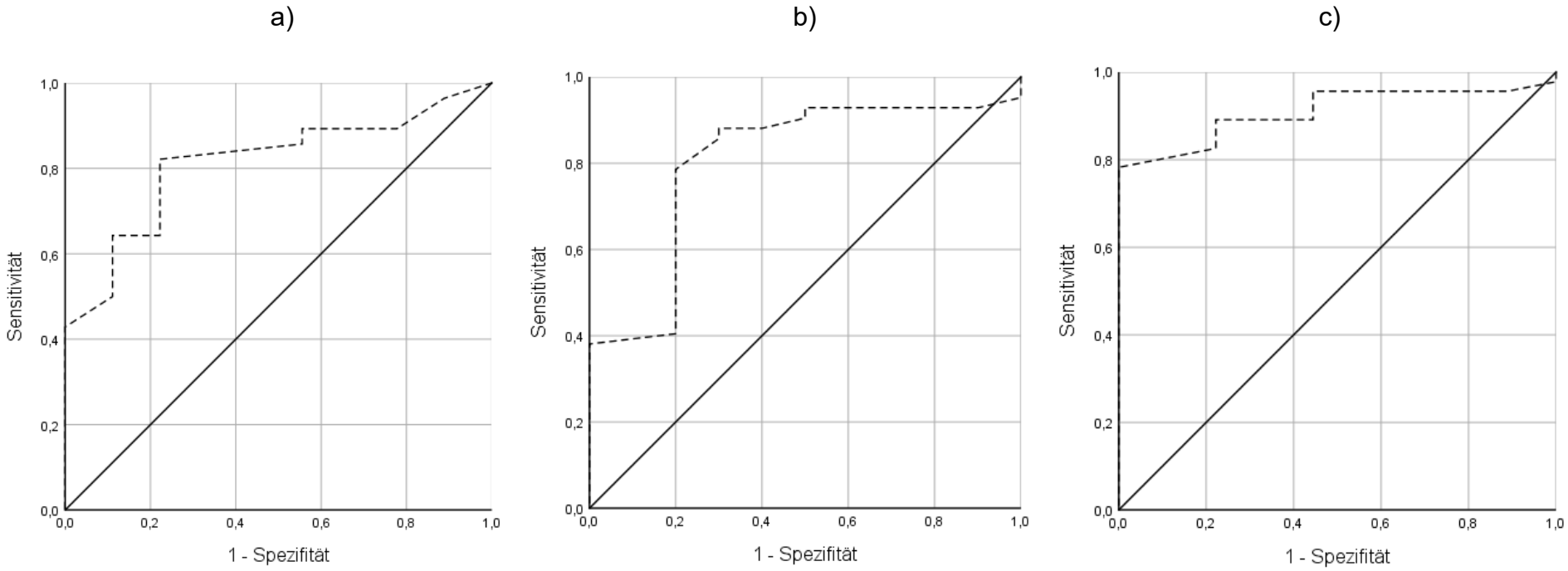

Abbildung 5. ROC-Kurven für die Vorhersage von rechtschreibschwachen Kindern (*T*-Wert ≤ 40 im Lückendiktat) anhand der Effizienz im Fehleridentifikationstest in den Jahrgangsstufen 2 (a), 3 (b) und 4 (c).