

Walter, Jürgen; Clausen-Suhr, Kristina

Entwicklung und Evaluation eines Verfahrens zur Prognose von zukünftigen Schwierigkeiten beim Schriftspracherwerb in der Grundschule

Empirische Sonderpädagogik 14 (2022) 1, S. 79-99



Quellenangabe/ Reference:

Walter, Jürgen; Clausen-Suhr, Kristina: Entwicklung und Evaluation eines Verfahrens zur Prognose von zukünftigen Schwierigkeiten beim Schriftspracherwerb in der Grundschule - In: Empirische Sonderpädagogik 14 (2022) 1, S. 79-99 - URN: urn:nbn:de:0111-pedocs-255318 - DOI: 10.25656/01:25531

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-255318>

<https://doi.org/10.25656/01:25531>

in Kooperation mit / in cooperation with:

Pabst Science Publishers <https://www.psychologie-aktuell.com/journale/empirische-sonderpaedagogik.html>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by-nc/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen und das Werk bzw. den Inhalt nicht für kommerzielle Zwecke verwenden.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-Licence: <http://creativecommons.org/licenses/by-nc/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor. You are not allowed to make commercial use of the work, provided that the work or its contents are not used for commercial purposes.

By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

peDOCS

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Informationszentrum (IZ) Bildung

E-Mail: pedocs@dipf.de

Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Empirische Sonderpädagogik, 2022, Nr. 1, S. 79-99
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Entwicklung und Evaluation eines Verfahrens zur Prognose von zukünftigen Schwierigkeiten beim Schriftspracherwerb in der Grundschule

Jürgen Walter & Kristina Clausen-Suhr

Europa-Universität Flensburg

Zusammenfassung

Einen wichtigen Baustein im Rahmen der Prävention von Problemen beim Schriftspracherwerb stellen Screening-Verfahren dar. Die meisten Instrumente zur Prognose solcher Schwächen im Grundschulalter sind als Einzelverfahren konzipiert. Die damit verbundene extrem hohe Zeit- und Personalintensität erschwert den flächendeckenden Einsatz in der Praxis deutlich. Im vorliegenden Beitrag werden die Entwicklung und Evaluation eines Filter-Screenings an einer Stichprobe von $N = 173$ Kindern der ersten Klassenstufe beschrieben, das auf Gruppenebene durchgeführt werden kann. Auf der Basis der logistischen Regressionsanalyse konnte ein durch eine Kreuzvalidierung abgesichertes Drei-Variablen-Prognosemodell (nonverbaler IQ, Satzverstehen, Phonemsynthese) identifiziert werden, das sehr gute AUC-Werte (bis zu $> .90$) und vor dem Hintergrund eines optimalen Prognose-Cutoffs gute Sensitivitäts- und Spezifitätswerte (91.2%; 84.9%) bei einem RATZ-Index = 87.4% aufweist.

Schlüsselwörter: Prävention, Screening, Schriftspracherwerb, Grundschüler, Lese- und Rechtschreibprobleme

Development and evaluation of a screening procedure estimating the risk of future literacy problems in children when entering primary school

Screening methods play an important role in preventing problems with the acquisition of written language. Most of these instruments applied in children when entering primary school are designed as individual procedures. The associated extremely high time and personnel intensity makes it much more difficult to use them across the board in practice. This paper describes the development and evaluation of a filter screening-instrument on a sample of $N = 173$ first graders, which can be carried out at group level. On the basis of a logistic regression analysis, a three-variable prognosis model (nonverbal IQ, sentence comprehension, phoneme synthesis), secured by a cross-validation, could be developed, which has very good AUC-values (up to $> .90$) and, against the background of an optimal prognosis cutoff, good sensitivity and specificity values (91.2%; 84.9%) with a RATZ index = 87.4%.

Keywords: prevention, screening, literacy acquisition, primary school, reading and spelling problems

Wie die Entwicklungen der letzten Jahre gezeigt haben, stehen Lehrerinnen und Lehrer vor dem Hintergrund der stattfindenden Inklusion und damit einer immer heterogener werdenden Schülerschaft in der Grundschule im Rahmen der Schuleingangsuntersuchung u.a. vor der Herausforderung, gerade im Bereich des Schriftspracherwerbs (Lesen und Rechtschreiben) passende Präventionsmaßnahmen zu initiieren (Kretschmann, 2007; Tröster, 2009). Präventive Maßnahmen besitzen in diesem Kontext insofern einen hohen Stellenwert, als Schriftsprachkompetenz eine Schlüsselkompetenz für die spätere schulische und berufliche Laufbahn sowie gesellschaftliche Teilhabe darstellt. Probleme in diesem Bereich manifestieren sich zu einem frühen Zeitpunkt und bleiben häufig bis ins Jugend- und Erwachsenenalter erhalten (Gallagher et al., 2000; Klicpera & Schabmann, 1993; Klicpera et al., 2006; Kohn et al., 2013; Landerl & Wimmer, 2008).

Zur Planung und Implementierung von Fördermaßnahmen sollte also spätestens beim Schuleintritt der Risiko-Status der Kinder eines Einschulungsjahrgangs bekannt sein. Um diesen festzustellen, können Screening-Verfahren auch im Sinne von Filter-Screenings (Tröster, 2009, S. 69) zum Einsatz kommen, die nicht nur das Kriterium der prognostischen Validität als zentrales Kriterium der Schulbereitschaftsdiagnostik erfüllen (Schneider & Hasselhorn, 2018), sondern sich auch zeitökonomisch und flächendeckend zum Einsatz bringen lassen. Letzteres wird durch vorwiegend als Einzelverfahren verfügbare Screenings deutlich erschwert (z.B. den *Rundgang durch Hörhäuser*, Martschinke et al., 2001; das *Bielefelder Screening* (BISC), Jansen et al., 2002; das *Heidelberger Auditive Screening* (HASE), Schöler & Brunner, 2008; Roos

et al., 2007; das *BASIC-Preschool*, Daseking & Petermann, 2009; den *Würzburger Vorschultest* (WVT), Endlich et al., 2016 oder das *LRS-Screening*, Endlich et al., 2019). Nur wenige Instrumente sind bislang in der Gruppe einsetzbar (z.B. *PB-LRS*, Barth & Gomm, 2014; *TEPHOBE*, Mayer & Motsch, 2014).

Um hier Abhilfe zu schaffen, ist es notwendig, mehr Gruppenverfahren zu entwickeln, die trotzdem gute klassifikatorische Eigenschaften besitzen. Die Entwicklung und Evaluation eines Filter-Screenings bezüglich der frühzeitigen Identifikation von Kindern, die im Bereich des Schriftspracherwerbs (Lese-Rechtschreibkompetenz) am Anfang der Grundschulzeit ein deutliches Risiko tragen, sollen im vorliegenden Beitrag beschrieben werden.

Schriftsprachkompetenz stellt ein komplexes Geflecht unterschiedlicher Fähigkeiten dar (Wortlesen, Leseflüssigkeit, Textverständnis, Rechtschreibung). Auch wenn Gruppen mit isolierten Defiziten in diesen Bereichen existieren (Fischbach et al., 2013) hat die vorliegende Untersuchung aus sonderpädagogischer Sicht Kinder mit umfangreichen und schwerwiegenden Defiziten im Blick, so dass der Risikostatus im Bereich der Schriftsprachkompetenz hier inhaltlich als Minderleistung im Lesen *und* im Rechtschreiben operationalisiert wird. Außerdem weisen Grube und Hasselhorn (2006) auf eine gewisse Globalität der Schulleistungen hin und zeigen, dass sich signifikante und darüber hinaus sowohl substanzielle zeitgleiche Korrelationen zwischen Lese- und Rechtschreibkompetenz (WLLP/DRT) innerhalb eines Schuljahres als auch zeitversetzte Zusammenhänge zwischen den Schuljahren ergeben. Die Interkorrelation von Lese- und Rechtschreibleistungen variiert zwischen $r = .42$ und $r = .79$ und liegt im

Schnitt bei $r = .58$. Dies deutet zumindest während der frühen Grundschulzeit erstens auf ein gewisses Bedingungsverhältnis hin und zweitens auf mögliche Parallelen bezüglich ihres Zustandekommens (Ahmed et al., 2014).

Die besondere Herausforderung für die Autoren bestand nun darin, theoretisch und empirisch bewährte Prädiktoren in Aufgabenformate zu kleiden, die Kindern in Gruppen präsentiert werden können, gleichzeitig aber den üblichen Gütekriterien genügen. Ziel dabei ist, ein sparsames und prognostisch valides Set von theoretisch und empirisch begründeten Prädiktoren zu identifizieren.

Prognostisch relevante Kompetenzen für den Schriftspracherwerb

Im Großen und Ganzen scheint sich die Befundlage bezüglich prognostisch relevanter Kompetenzen für den Lese- und Rechtschreiberwerb recht gut durch das von Ennemoser et al. (2012) entwickelte Pfadmodell für das Lesen (Abbildung 1) darstellen und

zusammenfassen zu lassen. Wie ersichtlich, liefern im Wesentlichen drei mehr oder weniger komplexe Bereiche unterschiedlicher Basiskompetenzen substanzielle Beiträge zur Vorhersage verschiedener Aspekte der Lesekompetenz (Lesegeschwindigkeit und Textverständnis) während der Grundschulzeit.

Erstens spielt die phonologische Informationsverarbeitung eine Rolle (Wagner & Torgesen, 1987). Diese spezifische Vorläuferfähigkeit besteht aus den Teilaspekten phonologische Bewusstheit (Fähigkeit, lautsprachliche Analyse- und Synthese zu realisieren), phonologisches Rekodieren im Arbeitsgedächtnis (Fähigkeit, Lautfolgen im Arbeitsgedächtnis bereit zu halten) sowie das phonologische Rekodieren beim Zugriff auf das semantische Lexikon (Fähigkeit, möglichst schnell lautliche Informationen aus dem Langzeitgedächtnis abzurufen = Benennungsgeschwindigkeit).

So kommen Ennemoser et al. (2012) im Rahmen ihrer Längsschnittstudie (vom letzten Kindergartenjahr bis zum Ende der Grundschulzeit) zu dem Ergebnis, dass die

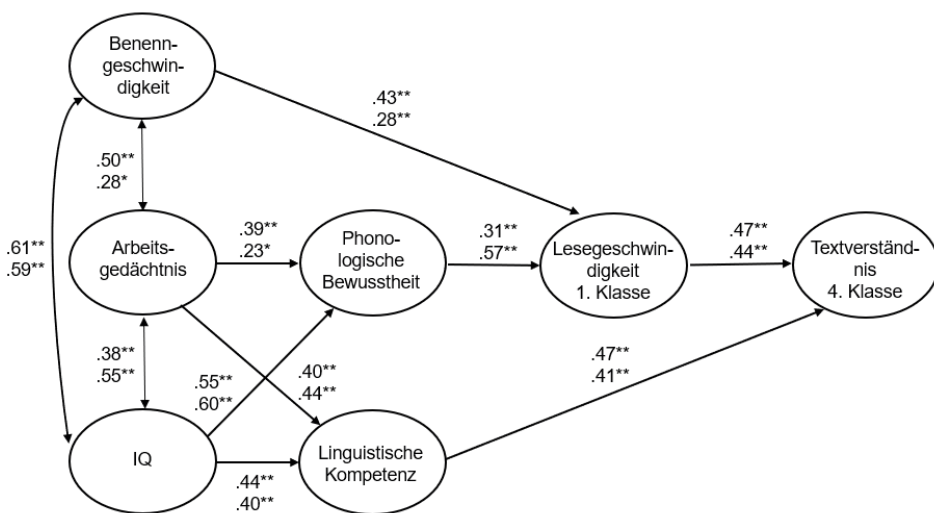


Abbildung 1. Strukturgleichungsmodell zur Vorhersage von Lesegeschwindigkeit und Textverständnis (obere Pfadkoeffizienten: Studie 1; untere Pfadkoeffizienten: Studie 2; * $p < .05$; ** $p < .01$), geändert nach Ennemoser et al., 2012, S. 63

Lesegeschwindigkeit während der gesamten Grundschulzeit am besten durch die Benennungsgeschwindigkeit (Mayer, 2018) und während des ersten Schuljahres darüber hinaus auch noch durch die phonologische Bewusstheit (Alliterations- und Anlautaufgaben, Phonemsynthese, Reimaufgaben, Silbensegmentierung, Laut-zu-Wort-Aufgaben) vorhergesagt werden kann.

Das Textverständnis wird in Übereinstimmung mit dem Simple-View-of-Reading Ansatz (Gough & Tunmer, 1986; Hoover & Gough, 1990) während der Grundschulzeit durch einen zweiten Faktor, nämlich vor allem durch die linguistische Kompetenz (Wortschatz, Korrektur semantischer Inkonsistenzen, Verstehen von Sätzen) prognostiziert. Bezüglich der Rechtschreibkompetenz (hier nicht dargestellt) stellte sich als dritter Faktor fast durchgängig der nichtsprachliche IQ (CFT-1) zusammen mit der phonologischen Bewusstheit und in den Klassenstufen drei und vier auch die linguistische Kompetenz als prognoserelevant heraus.

Die prädiktive Relevanz der in Abbildung 1 dargestellten Variablenkomplexe findet in internationalen Untersuchungen im Wesentlichen ihre Bestätigung.

So existiert eine Reihe eindeutiger Hinweise darauf, dass das *Arbeitsgedächtnis*, nämlich die Fähigkeit des kurzfristigen Behaltens und Manipulierens von Informationen, zusammen mit *domänen-spezifischem Vorwissen* bei Kindern am Anfang der Grundschulzeit, einen größeren prädiktiven Einfluss auf spätere schulische Leistungen (basales Lesen, Leseverständnis sowie Rechtschreiben) besitzt als die fluide und kristalline *Intelligenz* (Alloway, 2009; Alloway & Alloway, 2010; Peng & Fuchs, 2016; Sassu & Roebers, 2016).

Im Zusammenhang mit Untersuchungen zur Prognose von späteren Lese- und Rechtschreibproblemen spielt die *phonologische Bewusstheit* seit langem eine wichtige Rolle (Jansen et al., 2002; Martschinke et al., 2001; Schatschneider et al., 2004). Dies ist deswegen der Fall, weil vor allem im

englischsprachigen Raum gewonnene Befunde die präventive Wirkung einschlägiger Förderprogramme hervorheben (Ehri et al., 2001). Inzwischen überwiegen jedoch für den deutschsprachigen Raum deutlich weniger stark ausgeprägte Befunde sowohl bezüglich der Prognosekraft (Gorecki & Land-erl, 2015; Marx & Weber, 2006) als auch die Interventionseffekte betreffend (Fischer & Pfof, 2015).

Vor diesem Hintergrund sind in den letzten Jahren *linguistische Kompetenzen* in den Fokus zur Prognose von Schriftsprachkompetenz gerückt (Catts & Hogan, 2003; Roos et al., 2007; Schöler & Brunner, 2008). Der theoretische und empirische Hintergrund hierfür wird durch den Simple-View-of-Reading Ansatz geliefert (Gough & Tunmer, 1986). Dieser Theorie liegt die Annahme zugrunde, dass die gesamte Varianz der Lesefähigkeit (R) sowohl durch die elementare Dekodier- bzw. Wortlesefähigkeit (D) als auch durch sprachliche Verstehensprozesse (L), vor allem durch das Hörverstehen, aufgeklärt wird.

Bei der Überprüfung der Frage, ob nicht breitere sprachliche Kompetenzen (Wortschatz, Grammatik) den Erwerb von Schriftsprache vorhersagen können, zeigte sich bei Goldammer et al. (2010), dass allein das Satzgedächtnis (Nachsprechen von Sätzen im SETK 3-5; Grimm, 2001) von fünfjährigen Kindern ($N = 53$) eine herausragende Rolle bei der Prognose von Schriftsprachkompetenzen (Lesen und Rechtschreiben) im Alter von acht Jahren spielt. Die Varianz der Leistungen beim Nachsprechen von Sätzen (Satzgedächtnis) konnte wiederum zu 43% durch die auditive Merkspanne (Nachsprechen von Kunstwörtern) und darüber hinaus zu 8% durch den Wortschatz aufgeklärt werden. Die Autoren kommen zu der Schlussfolgerung, „dass eine ausschließliche Betrachtung der phonologischen Informationsverarbeitung keine hinreichend gute Prognose des Schriftspracherwerbs erlaubt ...“ (Goldammer et al., 2010, S. 54–55).

Schließlich zeigt eine Reihe von Untersuchungen, dass sich durch die *nonverbale Intelligenz* (IQ) Unterschiede in der Entwicklung schriftsprachlicher Kompetenzen vorhersagen lassen. In diesem Sinne konnten Peng et al. (2019) meta-analytisch vor dem Hintergrund von 680 Studien eine moderate Beziehung zwischen dem nonverbalen IQ und der Dekodierfähigkeit von $r = .29$, 95% CI [.27, .31] bzw. dem IQ und dem Leseverstehen von $r = .49$, 95% CI [.45, .52] feststellen. Des Weiteren wurde ermittelt, dass sich der Zusammenhang mit zunehmendem Alter erhöht und sich Intelligenz und Schulleistung (Lesen/Mathematik) im Laufe der Zeit gegenseitig beeinflussen. Schon allein dieser letzte Befund belegt die Sinnhaftigkeit des Heranziehens von sowohl domänenspezifischen als auch allgemein-kognitiven Prädiktorvariablen.

Als ein Beispiel für eine solche Vorgehensweise ermittelten Hayiou-Thomas et al. (2006) auf der Basis einer multiplen Regressionsanalyse, welche allgemeinsprachlichen, sprachlich-phonologischen und nichtsprachlichen Variablen, die im Alter von 4;06 Jahren erhoben wurden, die Lesekompetenz (Index aus Wort- und Nonwort-Lesekompetenz und Lehrereinschätzung) im Alter von sieben Jahren am besten vorhersagen konnten. Drei sprachliche (domänenspezifische) Variablen lieferten einen von allen anderen Prädiktorvariablen unabhängigen signifikanten Beitrag, nämlich die expressive grammatische Kompetenz ($\beta = .14$), die verbale Flüssigkeit ($\beta = .14$) sowie das phonologische Arbeitsgedächtnis ($\beta = .12$). Ein mindestens ebenso großes prognostisches Gewicht wie die jeweils einzelnen sprachlichen Variablen besaß aber auch die nonverbale Kompetenz ($\beta = .19$). Zu ähnlichen Befunden kommen auch Tiu et al. (2003).

Die bisherige Befunddarstellung zur Identifikation prognostisch relevanter Kompetenzen für den Schriftspracherwerb basiert methodologisch auf korrelativen bzw. linear-regressiven Analysemethoden, die jeweils das gesamte Messwertespektrum

der Kriteriumsvariablen einbeziehen. Da es bei der Entwicklung eines Filter-Screenings jedoch vor allem darum geht, Risiko-Kinder am unteren Ende der Messwerteskala zu identifizieren (Fischbach et al., 2013), bildet ein korrelativ-regressiver Ansatz den Zusammenhang zwischen Prädiktoren und Kriterium möglicherweise unscharf ab.

An dem Beispiel von Fritz et al. (2013, S. 34) wird deutlich, dass selbst eine Korrelation von $r = .69$ zwischen Prädiktor und Kriterium auf klassifikatorisch-prognostischer Ebene nur zu einer Sensitivität von 33.3% und einem unzulänglichen RATZ-Wert von 24.5% führen kann.

Aus diesem Grund ist es notwendig, Befunde zu betrachten, die auf der Basis eines klassifikatorischen Ansatzes (Tröster, 2009) zustande gekommen sind. Catts et al. (2001) testeten Kinder im Vorschulalter mithilfe einer breit angelegten Batterie von sozialen, allgemein-kognitiven (nonverbaler IQ) und sprachspezifisch-linguistischen Basiskompetenzen (expressiver und rezeptiver Wortschatz, Satzimitation, Grammatik, verbaler Ausdruck) inklusive der phonologischen Bewusstheit (Phoneme-/Silben weglassen) sowie der Benennungsgeschwindigkeit. Kriterium war das Leseverstehen in der 2. Klasse. Auf der Basis einer schrittweise angelegten logistischen Regressionsanalyse wurde ein Variablenset mit hoher Prognosegüte (Sensitivität = 73.5%) identifiziert, das aus Kompetenzen wie Sätze nachsprechen, phonologischer Bewusstheit (Silben/Phoneme weglassen), Buchstabenidentifikation, schnelles Benennen von Gegenständen sowie dem Bildungsniveau der Mutter bestand.

Abgesehen von der soziologischen Komponente erfuhr dieses klassifikatorische Grundmodell in weiteren Untersuchungen im Wesentlichen seine Bestätigung (AUC-Werte zwischen .85 und .92), wenn ein Mix aus phonologischen und linguistisch-sprachspezifischen Variablen integriert war (Catts et al., 2015, 2016).

Ableitungen und Hypothesen für die vorliegende Untersuchung

Vor dem Hintergrund der Zielsetzung, nämlich ein zeitökonomisch einsetzbares, prognostisch valides sowie in Gruppen durchführbares Filter-Screening (Tröster, 2009) zu entwickeln, liegt es nahe, auf der Basis der beschriebenen bewährten Basiskompetenzen (Abbildung 1) jeweils ein für Gruppen präsentierbares Aufgabenformat zu entwickeln. Dazu sollte ein zunächst sehr breites Tableau von zehn Prädiktorvariablen (Untertests) herangezogen werden. Dieses umfasst die Phonem-Synthese, die Phonem-Identifikation (am Anfang und am Ende eines Wortes), das Arbeitsgedächtnis, den Wortschatz, das Satzverstehen (zwei Varianten) sowie drei Untertests aus dem CFT (Klassifikation, Ähnlichkeiten, Matrizen). Aufgrund der Tatsache, dass sich die Benennungsgeschwindigkeit sinnvollerweise nur im Einzelverfahren erheben lässt, wurde diese Variable nicht herangezogen. Die zu überprüfende Hypothese 1 ist nun folgende: Aus einem breiten Set von im Prinzip bewährten Prädiktorvariablen, die bei Kindern zum Einschulungszeitpunkt erfasst werden, lässt sich per Datenreduktion mithilfe einer schrittweisen logistischen Regressionsanalyse eine überschaubare Anzahl von Prädiktoren identifizieren, die eine statistisch signifikante und praktisch brauchbare Klassifikation von Risiko- und Nicht-Risiko-Kindern bezüglich ihrer Schriftsprachkompetenz nach ca. einem Jahr erlaubt.

Daraus ergibt sich die Notwendigkeit der Überprüfung der Hypothese 2: Das identifizierte Prognosemodell erweist sich vor dem Hintergrund einer Kreuzvalidierung an unabhängigen Gruppen als stabil.

In der Regel wird der Ausprägungsgrad von Prädiktorvariablen (Intelligenz, Aspekte der phonologischen Bewusstheit, Arbeitsgedächtnis etc.) zur Vorhersage zukünftiger Schulleistungen mithilfe von Skalen auf der Basis des Latent-Trait-Modells erfasst. Denkbar wäre aber auch speziell für den Schulleistungsbereich, dass ein formati-

ves Modell (Bühner, 2011) zum Einsatz kommt. Es gibt nämlich aus anderen Bereichen (Persönlichkeit, Intelligenz, Mathematik) empirische Hinweise darauf, dass das Heranziehen von Einzelitems (formatives Messmodell) im Rahmen der Vorhersage individueller Unterschiede zu einer besseren Prognosequalität führen kann (Mazzocco & Thompson, 2005; Möttus & Rozgonjuk, 2021; Schroeders et al., 2020; Seeboth & Möttus, 2018). Daraus ergibt sich ein Interesse an der Überprüfung der Hypothese 3: Auf der Basis eines formativen Messmodells lassen sich bessere prognostische Resultate erzielen als durch ein Latent-Trait-Modell.

Da es bei der Entwicklung eines Filter-Screenings vor allem darum geht, Kinder am unteren Ende der Messwerteskala frühzeitig zu identifizieren, sollte bei der Überprüfung der Hypothesen ein entsprechender Schwellenwert (Cutoff) zugrunde gelegt werden. Analog zu vergleichbaren Untersuchungen macht es Sinn, Kinder mit mehr als einer SD ($PR \leq 16$) unterhalb des Mittelwertes im Lesen und Rechtschreiben als gefährdet zu definieren (Catts et al., 2016). Da aber sowohl liberalere Definitionen ($PR \leq 25$, Fletcher et al., 1994) als auch strengere ($1.2 SD \triangleq PR \leq 10$; Badian et al., 1990) existieren, sollen die formulierten Hypothesen auf Basis der drei unterschiedlichen Cutoffs überprüft werden.

Methode

Datenerhebung und Stichprobenbeschreibung

Für die Evaluation des ins Auge gefassten Screening-Instruments (Arbeitstitel: Flensburger Schulspiel, FleSch) konnten in einer ersten Welle (Oktober/November 2019) in 12 Grundschulen die Prädiktor-Daten von $N = 258$ Kindern im Alter zwischen 73 und 105 Monaten ($M = 85.3$; $SD = 5.5$) zu Beginn der 1. Klasse erhoben werden. Dabei handelt es sich um $N = 135$ Jungen (52.3%) und $N = 123$ Mädchen (47.7%). Insgesamt

besaßen $N = 21$ (8.1%) einen Migrationshintergrund (zu Hause wird neben Deutsch eine andere Sprache gesprochen) und $N = 7$ Kinder (2.7%) wiesen einen offiziell festgestellten sonderpädagogischen Förderbedarf auf. In dieser Stichprobe verweigerten $N = 4$ Kinder (1.6%) die weitere Teilnahme, so dass sich schließlich eine Stichprobe von $N = 254$ Kindern ergab.

Die Kriteriumsmessungen mit dem ELFE-II (Lenhard et al., 2017) sowie dem DRT-1 (Müller, 2004) erfolgten im September/Oktober 2020, also am Anfang der 2. Klassenstufe. Insgesamt erstreckt sich der Prognosezeitraum damit über ca. ein Jahr, an dessen Ende sich für die Lesekompetenz (ELFE-Gesamt) ein $T = 41.6$ ($SD = 10.5$) und für das Rechtschreiben im DRT-1 (Wortfehler) ein $T = 47.7$ ($SD = 8.5$) ergab, und zwar auf der Basis von jeweils $N = 184$ bzw. $N = 185$ Kindern.

Bei der vorliegenden Stichprobe zeigt sich dabei für Kinder mit schwachen Leistungen im Lesen und Rechtschreiben zu Beginn der 2. Klasse (Cutoff von $PR \leq 10$) eine Prävalenz von 14.3%, für denjenigen von $PR \leq 16$ eine von 20.3% und bezüglich des Cutoff von $PR \leq 25$ eine Rate von 29.7%.

Vorgehensweise

In Rahmen der Erhebung der Prädiktorvariablen in Gruppen von 10 bis 15 Kindern wurde diese in eine Rahmenhandlung eingebettet. Jeder Untertest wird in einem situativen Bezug durch verschiedene Protagonisten eingeführt. Um die Motivation und Aufmerksamkeit der Kinder über den Zeitraum der Aufgabendurchführung aufrechtzuerhalten, wurden immer wieder musische Elemente und Bewegungsaufgaben integriert. Zur Unterstützung der Aufgaben-Fokussierung aller Kinder und zur Vermeidung möglicher Störfaktoren wurde die Bearbeitung jedes Untertests im Sinne eines kontinuierlichen Kontingenzmanagements verstärkt.

Um die Aufgabenpräsentation gruppentauglich zu gestalten sowie eine gute Durchführungsobjektivität zu gewährleis-

ten, wurden die Aufgaben mithilfe einer Präsentations-Software standardisiert umgesetzt. Die Steuerung der Bearbeitung aller Untertests beinhaltete auch eine zeitliche Grenze, deren Beginn und Ende durch akustische Signale angezeigt wurde. Zu jedem Untertest bearbeiteten die Kinder die Aufgaben in den jeweils zugehörigen Testheften. Während der Input also in der Regel computerbasiert erfolgte, lagen die Testergebnisse grundsätzlich in Papierform vor. Jeder Untertest wurde mit zwei Übungsaufgaben eingeführt. Die Durchführung erfolgte durch zuvor geschulte Studierende der Sonderpädagogik.

Operationalisierung der Prädiktor- und Kriteriumsvariablen

Zur Erfassung der nonverbalen *Intelligenz* (IQ) wurden die drei Untertests Klassifikation, Ähnlichkeiten und Matrizen des CFT-1 (Weiß & Osterland, 1997) mit ihren jeweils 12 Items herangezogen. Die interne Konsistenz der drei Skalen liegt auf der Basis der vorliegenden Stichprobe bei $\alpha = .71$, $\alpha = .80$ und $\alpha = .83$ ($N = 248$).

Der Untertest zur Erfassung der Kapazität des *Arbeitsgedächtnisses* (AG, 12 Items) wurde von den Autoren komplett neu konzipiert. Die Kinder müssen sich vier über den Lautsprecher genannte Gegenstände in der richtigen Reihenfolge merken, um diese hinterher in ihrem Testheft als die richtige Bildreihe aus einer Auswahl abgebildeter Gegenstände zu identifizieren. Die Durchführungszeit beträgt ca. 10 Minuten ($\alpha = .77$, $N = 241$).

Um den rezeptiven *Wortschatz* der Kinder (WS) zu erfassen, wurden 15 Items aus dem WWT 6-10 (Glück, 2007) in aufsteigendem Schwierigkeitsgrad verwendet. Aus einer Auswahl von vier Bildern/Situationen bzw. Gefühlen soll dasjenige Item (Wort-Bild-Zuordnung) ausgewählt werden, das lautsprachlich vorgegeben wurde. Die zeitliche Taktung der Aufgaben betrug 10 Sekunden, die benötigte Bearbeitungszeit ca. 17 Minuten ($\alpha = .71$, $N = 250$).

Für die Operationalisierung der *Phonem-Synthese* (PS, 20 Items) werden die Benennungen von Gegenständen laut für laut vorgesprochen. Aufgabe ist es dann, aus vier abgebildeten Möglichkeiten die richtige Lösung zu finden. Die Bearbeitungszeit beträgt 11 Sekunden pro Item. Für den gesamten Untertest werden ca. 12 Minuten benötigt ($\alpha = .83$, $N = 247$).

Die Operationalisierung der *Phonem-Identifikation* am Anfang (PIA, acht Items) erfolgt in Form eines Spiels: „Ich sehe was, was du nicht siehst und da hört man ganz vorne /m/. Was ist das wohl?“ Die Kinder müssen den passenden Gegenstand im Testheft ankreuzen. Die zeitliche Taktung der Aufgaben liegt bei 20 Sekunden pro Item. Analog wird mit den Endphonem verfahren. Die Bearbeitungszeit beträgt ca. 12 Minuten (PIA: $\alpha = .81$, $N = 244$; PIE: $\alpha = .69$, $N = 245$).

Die Präsentation des Untertests *Satzverständnis* (SVS, acht Items) besteht in der richtigen Zuordnung eines vorgesprochenen Satzes zu einem passenden Bild aus einer Auswahl von vier Varianten. Die Sätze sind im Schwierigkeitsgrad ansteigend und berücksichtigen zunehmend komplexe Satzgefüge, Negationen und präpositionale Bezüge. Die zeitliche Taktung der Aufgaben beträgt 15 Sekunden pro Item, die benötigte Zeit zusammen mit dem 2. Teil, dem nachfolgend beschriebenen SVQ, beträgt ca. 15-20 Minuten ($\alpha = .71$, $N = 243$).

Der zweite Untertest zum *Satzverständnis* basiert auf „Quatschsätzen“ (SVQ, 16 Items): „Ein guter Detektiv merkt schnell, ob Leute Quatsch erzählen oder nicht. Ihr sollt mir jetzt dabei helfen zu merken, ob jemand Quatsch erzählt oder nicht. Ihr hört nun einen Satz und müsst merken, ob das ein Quatschsatz ist oder nicht.“ Die Kinder haben dann Sätze wie „Die Bäuerin pflückte Erdbeeren von den Bäumen“ oder „In Afrika ist es sehr heiß“ auf Richtigkeit zu überprüfen und in ihrem Testheft als richtig oder falsch zu markieren ($\alpha = .81$, $N = 243$).

Im Zusammenhang mit den Kriteriums-messungen wurde die *Lesekompetenz* durch den ELFE-II erfasst (Lenhard et al., 2017). Das Verfahren erfasst neben dem Leseverständnis sowohl die Leseflüssigkeit als auch die Lesegenauigkeit auf der Wort-, Satz- und Textebene. Die Teilergebnisse werden zu einem Gesamtergebnis verrechnet. Damit wird hier die Lesekompetenz im Sinne eines breiten Fähigkeitsspektrums erfasst, zu dem auch das Sprachverstehen sowie das schlussfolgernde Denken gehören. Die Odd-Even-Split-Half-Reliabilität für das Gesamtergebnis liegt sie bei $r = .96$ für die hier verwendete Papierform. Nach 30 Tagen zeigt sich eine Retestreliabilität von $r = .93$ für das Gesamtergebnis. Für den Test ergibt sich eine mittlere kriteriumsbezogene Validität auf der Basis der Korrelation mit einem anderen standardisierten Lesetest (SLS 2-9) von $r = .77$. Die Übereinstimmung mit dem Lehrerurteil bezüglich der Leseleistung liegt bei $r = .70$.

Die *Rechtschreibkompetenz* wurde durch den DRT-1 erfasst (Müller, 2004). Das Verfahren ist konzipiert als differenziertes System der Früherfassung von Rechtschreibstörungen. Die 30 Testwörter sind in zwei Geschichten eingekleidet, die der TL zunächst vorliest. Als Testwörter werden nur einfache, lauttreue Wörter gewählt. Für die im vorliegenden Zusammenhang herangezogene quantitative Rechtschreibleistung (quantitative Gesamtleistung) ergeben sich folgende Reliabilitätsbefunde: Gemäß der Halbierungsmethode ein $r = .95$ und für die Paralleltestmethode ein $r = .89$. Die Korrelationen zur Bestimmung der Validität mit dem Lehrerurteil liegen im Mittel bei $r = .81$.

Forschungsstatistische Verarbeitung der Daten

Zur Überprüfung der Hypothese 1, ob sich per Datenreduktion eine überschaubare Anzahl von Prädiktoren identifizieren lässt, die eine statistisch signifikante und praktisch bedeutsame Klassifikation ermöglicht,

wurde eine schrittweise logistische Regression (Backhaus et al., 2018; Rudolf & Müller, 2012) mit den zehn weiter oben aufgeführten Variablen als Start-Set unter der Bedingung „Vorwärts schrittweise“ gerechnet. Ein solche Vorgehensweise beinhaltet einen Test auf Aufnahme einer Variablen, der auf der Signifikanz der Scorestatistik beruht ($PIN = 0.05$) und einem Test auf Abschluss ($POUT = 0.10$).

Die logistische Regression ermöglicht im Gegensatz zur klassischen Regressionsanalyse eine Vorhersage für den Fall eines dichotomen Kriteriums (Rudolf & Müller, 2012) so wie es häufig im Zusammenhang mit der Evaluation von Screening-Verfahren zur Anwendung kommt (Catts et al., 2001, 2015, 2016). Anders ausgedrückt: Bei der logistischen Regression wird der Einfluss des Sets der erklärenden und prognose-relevanten (unabhängigen) Variablen auf die Wahrscheinlichkeit geschätzt, dass die abhängige Variable den Wert 1 ($p [y=1]$) annimmt. Damit kann für jedes Kind eine Risiko-Wahrscheinlichkeit berechnet werden. Liegt diese oberhalb eines definierten Schwellenwertes (Risiko-Cutoffs), erfolgt eine Zuordnung in die Risiko-Gruppe.

Vor diesem Hintergrund ist auch die Berechnung der Prognose-Güte möglich, indem man für alle Probanden den empirisch beobachteten Status der Entwicklung der Schriftsprachkompetenz (Störung/keine Störung = Statusvariable) am Ende der ersten Klasse mit den anhand der berechneten (vorhergesagten) Risiko-Wahrscheinlichkeiten zum Einschulungszeitpunkt (Testvariable = Prädiktor-Variable als individuelle Risiko-Wahrscheinlichkeit) in einer Vierfeldertabelle in Beziehung setzt.

Um das Ziel einer maximalen Trennung zwischen Risiko- und Nicht-Risiko-Kindern zu erreichen, wird inzwischen sehr häufig auf der Basis einer so genannten ROC-Analyse (Receiver Operating Characteristic; Tröster, 2009, S. 121; Youngstrom, 2014) der beste Risiko-Cutoff des Prädiktors ermittelt.

Dies passiert, indem jeweils für alle möglichen Trennwerte (Cutoffs) des Prädiktors der Anteil korrekt vorhergesagter Merkmalsträger (Richtig-Positive; Sensitivität) auf der Y-Achse und der Anteil von Probanden, für die fälschlicherweise das Vorliegen des Merkmals Lernschwäche diagnostiziert wird (Falsch-Positive; 1-Spezifität), auf der x-Achse markiert wird (Abbildung 2). Im Falle einer Zufallsvorhersage und damit einer generellen Unbrauchbarkeit des Prognosemodells entspricht die ROC-Kurve der Winkelhalbierenden im Koordinatensystem.

Je besser die Vorhersagegüte des Prognosemodells ausfällt, desto stärker weicht also die Kurve von der Winkelhalbierenden ab. Als *AUC* (Area-Under-the-Curve = Fläche unter der Kurve) wird die Fläche unter der ROC-Kurve bezeichnet. Im Falle einer Zufallsvorhersage beträgt diese 0,5 ($AUC = 0.5$). Je besser die generelle Vorhersagegüte (Genauigkeit) des Screening-Verfahrens ist, desto mehr nähert sich die *AUC* dem Wert 1 an. Der beste Risiko-Cutoff ist derjenige, der den Punkt der Kurve erzeugt, der dem Wert 1 der y-Achse am nächsten liegt (s. Abbildung 2, durch Kreis markiert).

Der *AUC*-Wert kann auch als geschätzte Wahrscheinlichkeit dafür interpretiert werden, wie akkurat das Screening-Verfahren ein zufällig aus der Gruppe der späteren Risiko- und Nicht-Risiko-Kinder herausgegriffenes Individuum (positiv oder negativ) klassifiziert (Babu, 2015). *AUC*-Werte von größer als .80 können als gut und solche größer als .90 als hervorragend bezeichnet werden (Catts et al., 2015, S. 281).

Analog zum beschriebenen Prozedere wurde auch die Überprüfung der Hypothese 3 vorgenommen.

Um die Stabilität des identifizierten Prognose-Modells zu überprüfen sowie Hinweise für ein mögliches Overfitting zu bekommen (Hypothese 2), wurde die Gesamtstichprobe per Zufall in zwei etwa gleich große unabhängige Untergruppen (Gruppe A und Gruppe B) aufgeteilt (Holdout-Sample-Verfahren). In einem ersten

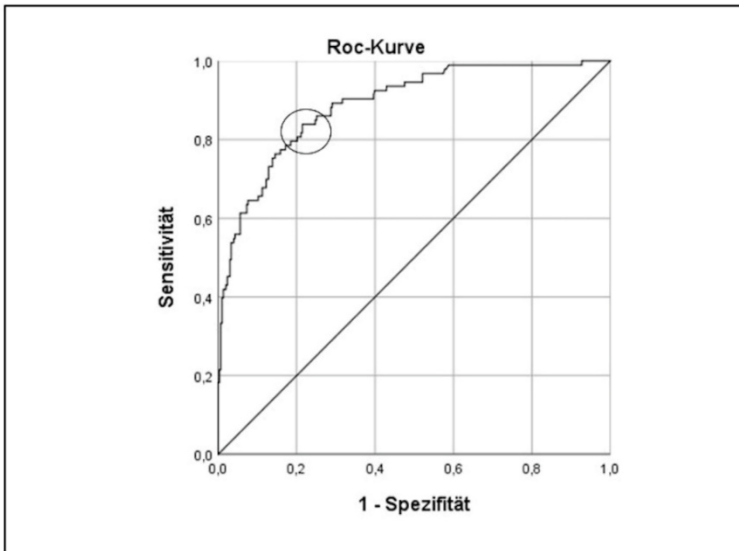


Abbildung 2. Beispiel für eine ROC-Kurve mit einem AUC-Wert = .89

Überprüfungsschritt wurde mithilfe der Gruppe A (Lerngruppe, $N = 81$) eine logistische Regressionsanalyse (Einschlussverfahren) vor dem Hintergrund des jeweiligen Kriterium-Cutoffs gerechnet und auf der Basis der ermittelten Modellparameter dieser Teilgruppe eine Risiko-Klassifikation der Probanden auch für die Gruppe B (Testgruppe, $N = 93$) vorgenommen. In einen weiteren Schritt wurde dies genau andersherum realisiert (Rollentausch). Die Überprüfung der Übereinstimmung der Klassifikationsgüte beider Gruppen wurde auf Basis einer ROC-Analyse (Vergleich der AUC-Werte) auch inferenzstatistisch umgesetzt (Hanley & Mc Neil, 1983; Hajian-Tilaki, 2018).

Während sich die Fläche unter der ROC-Kurve (AUC) über alle möglichen Trennwerte (Cutoffs des Prädiktors) berechnet und damit einen globalen Index darstellt, soll in einem weiteren Auswertungsschritt ermittelt werden, welche üblichen Güte-Aussagen über das Screening-Verfahren (Tröster, 2009) vor dem Hintergrund des z.B. am besten zwischen Risiko- und Nicht-Risiko-Kindern trennenden Risiko-Cutoffs gemacht werden können.

In diesem Zusammenhang ist zunächst die *Selektionsquote* (SQ) zu nennen. Sie gibt den prozentualen Anteil der durch das Screening-Verfahren angezeigten Risiko-Kinder relativ zur Gesamtheit aller Kinder wieder. Die *Gesamttrefferquote* (GT) zeigt den Anteil aller richtigen Screening-Entscheidungen an der Gesamtheit aller Entscheidungen an.

Einen sehr wichtigen Güte-Index stellt die *Sensitivität* (SN) eines Verfahrens dar. Die Sensitivität stellt den Anteil der Kinder mit einer späteren Schwäche im Bereich der Schriftsprache dar, der durch das Screening frühzeitig erkannt wird (Tröster, 2009, S. 88).

Damit eine Sensitivitätsrate als praktisch bedeutsam bezeichnet werden kann, sollte diese nach Catts et al. (2009, S. 170) bei mindestens 80% liegen. Eine solche kann in der Regel jedoch nur auf Kosten einer relativ niedrigen *Spezifität* (Anteil der Kinder mit einem richtig-negativen Screening-Befund) bzw. einer hohen Rate von falsch-positiven Klassifikationen erreicht werden (Poulsen et al., 2017).

Neben der allgemeinen Prognosequalität eines Screenings (die Menge der Treffer) ist

gerade auch aus Sicht der Praxis die Sicherheit eines Screenings von großer Bedeutung. Hierüber geben nach Tröster (2009, S. 89–91) u.a. die Werte der *positiven Korrektheit* (*PK*, auch positiv prädiktiver Wert) und der *negativen Korrektheit* (*NK*, auch negativ prädiktiver Wert) Auskunft. Die *PK* ist der Anteil der richtig-positiven Screening-Befunde an der Gesamtzahl der positiven Screening-Befunde, gibt also die Wahrscheinlichkeit an, mit der ein positiver Screening-Befund tatsächlich zutrifft. Dies ist für die Praxis von hoher Relevanz. Die negative Korrektheit (*NK*) bezeichnet auf der anderen Seite die Wahrscheinlichkeit, mit der ein negativer Befund tatsächlich zutrifft.

Neben der *PK* gibt das *positive Likelihood Ratio* (*LR+*) an, um wieviel wahrscheinlicher ein positiver Prognosebefund beim Vorliegen einer (späteren) Störung als beim Nicht-Vorliegen einer (späteren) Störung ist. Je größer *LR+* ist, desto sicherer kann bei einem positiven Screening-Befund auf das Vorhandensein einer Störung geschlossen werden (Tröster, 2009, S. 95). Nach Bender (2001) kann ein $LR+ > 10$ (Screening-Befund bei Vorliegen einer Störung zehnfach wahrscheinlicher) als sehr gut bezeichnet werden, eines zwischen 5 und 10 als gut, eines zwischen 2 und 5 als mäßig und ein $LR+ < 2$ als schlecht.

Das *Odds Ratio* (*OR*) als Index der Vorhersagegenauigkeit gibt die Chance an, dass bei einem positiven Screening-Befund später Lernprobleme auftauchen x -mal größer ist als nach einem negativen Befund.

Neben dem international gebräuchlichen *AUC*-Wert wird ebenfalls zur Beurteilung der generellen Prognosequalität eines Screenings im deutschsprachigen Raum üblicherweise der *RATZ-Index* (Relativer Anstieg der Treffer gegenüber der Zufallstrefferquote) zur generellen Leistungsfähigkeit eines Screenings herangezogen (Tröster, 2009, S. 143), der jedoch nur vor dem Hintergrund eines konkreten Prädiktorschwellenwertes berechnet werden kann. Der Index relativiert die Gesamttrefferquote (*GT*) im Hinblick auf die Zufallstrefferquote,

so dass die Güte von Screenings (wie auch beim *AUC*-Wert) unabhängig von der Prävalenzrate (*GR*) und der Selektionsquote (*SQ*) miteinander verglichen werden kann. Nach Jansen et al. (2002) sind *RATZ-Indizes* $> 66\%$ als sehr gut, solche zwischen 34% und 66% als gut aber unspezifisch und diejenigen unter 34% als unzureichend zu betrachten.

Ergebnisse

Wie aus Tabelle 1 (linke Spalte) im Rahmen der Überprüfung von Hypothese 1 zu ersehen ist, zeigen sich bei der schrittweisen logistischen Regression drei Schritte der Modellbildung (= Anzahl der aufgenommenen Variablen). Ganz rechts in der Tabelle (ΔX^2) ist die Signifikanz der Verbesserung der Modellgüte (gemessen als Veränderung von X^2) von Schritt zu Schritt der Variablenaufnahme angegeben.

Zur Überprüfung der Signifikanz des Gesamt-Modells wird in der Regel ein X^2 -Test herangezogen (Tabelle 1, letzte Zeile). Dieser signalisiert mit einem Prüfwert von $X^2 = 43.45$ ($df = 3$, $p = .021$), dass das Modell im Vergleich zu einem Null-Modell in seiner Gänze einen höchst signifikanten Erklärungsbeitrag leistet.

Etwas anschaulicher kann Nagelkerkes $R^2 = .358$ interpretiert werden (Tabelle 1, letzte Zeile), nämlich wie das Bestimmtheitsmaß der linearen Regression, das als Anteil der Varianz der abhängigen Variable (Leistungsstatus eines Kindes am Ende der 1. Klasse), der durch die unabhängigen Variablen (IQ, SVS, PS) erklärt wird. Im vorliegenden Fall wird quasi eine „Varianzaufklärung“ von guten 35.8% erreicht. Eine Umrechnung des R^2 in das Effektstärkemaß von Cohen ($f^2 = R^2 / [1 - R^2]$) ergibt einen Wert von 0.56 , was als sehr starker Effekt interpretiert werden kann (Cohen, 1988).

Der Hosmer-Lemeshow-Test (HL-Test) liefert einen Beleg für das Ausmaß, wie gut das Modell (Schritt 3) an die empirischen Daten angepasst ist. Der Unterschied sollte also

Tabelle 1: Ergebnisse der schrittweisen logistischen Regressionsanalyse ($N = 173$) zur Prognose eines Risikos bezüglich der Schriftsprachkompetenz ($PR \leq 16$) auf der Basis der Summenwerte der zehn Untertests (Latent-Trait)

		<i>B</i>	<i>SE(B)</i>	<i>Wald</i>	<i>df</i>	<i>Sig.</i>	<i>Exp(B)</i>	ΔX^2
Schritt 1								$p = .000$
	IQ (CFT-Mat)	-.355	.075	22.462	1	.000	.701	
	Konstante	.908	.484	3.517	1	.061	2.479	
Schritt 2								$p = .001$
	IQ (CFT-Mat)	-.345	.079	19.098	1	.000	.708	
	Satzverstehen (SVS)	-.459	.145	10.078	1	.002	.632	
	Konstante	4.070	1.135	12.856	1	.000	58.562	
Schritt 3								$p = .021$
	IQ (CFT-Mat)	-.277	.085	10.534	1	.001	.758	
	Satzverstehen (SVS)	-.400	.147	7.447	1	.006	.670	
	Phonem-Synthese (PS)	-.136	.060	5.216	1	.022	.873	
	Konstante	5.124	1.286	15.867	1	.000	168.020	
Modell: $X^2 = 43.45$ ($df = 3$, $p = .021$); $R^2 = .358$; $HL-X^2 = 9.4$ ($df = 8$, $p = .31$)								

Anmerkungen. *B* = Regressionskoeffizient; *SE(B)* = Standardfehler des Regressionskoeffizienten; *Wald* = Waldkriterium zur Bestimmung der Signifikanz des Einzelprädiktors; *df* = Freiheitsgrade; *Sig.* = Signifikanz des Einzelprädiktors; *Exp(B)* = Effekt-Koeffizient; ΔX^2 = Veränderung der Modellgüte je Schritt; $HL-X^2$ = Hosmer-Lemeshow- χ^2 -Test; R^2 = Nagelkerkes R-Quadrat

möglichst klein sein (Beibehaltung von H_0) und wird durch die Prüfgröße X^2 ermittelt. Wie in Tabelle 1 (letzte Zeile) dargestellt, liefert der HL-Test ein nicht signifikantes Ergebnis ($p = .31$), was einen guten Modell-Fit belegt.

Nach der Feststellung der Eigenschaften des Gesamt-Modells können die einzelnen Koeffizienten der Prädiktorvariablen betrachtet werden. Die jeweiligen Signifikanzen (*Sig.*, $p < .05$) bei Schritt 3 bedeuten, dass die einzelnen Prädiktoren, nämlich der IQ (CFT-Matrizen), das Satzverstehen (SVS) sowie die Phonem-Synthese (PS) einen Einfluss auf die abhängige Variable (Schriftsprachkompetenz) besitzen. Das deutet sich bereits in den Signifikanzen (*Sig.*) im *Wald*-Test (*Wald*) bei den vorherigen Schritten 1 und 2 des Modellaufbaus an.

Die negativen Vorzeichen des jeweiligen Regressionskoeffizienten *B* (2. Spalte in Tabelle 1), der einen direkten Vergleich zwi-

schen dem Gewicht der Variablen erlaubt, zeigen an, dass bei steigenden Werten im IQ (Matrizen), Satzverstehen und der Phonem-Synthese das Risiko einer Schwäche im Bereich der Schriftsprachkompetenz jeweils sinkt.

Das *Exp(B)* in der letzten Spalte der Tabelle 1 gibt den entlogarithmierten Logit-Koeffizienten (Effekt-Koeffizienten) als Odds Ratio (Chancenverhältnis) wieder. Ein Wert von genau 1 bedeutet keine Veränderung und somit kein Einfluss der jeweiligen Prädiktorvariablen. Die Logits lassen sich jedoch anschaulicher als Wahrscheinlichkeiten ausdrücken. Das *Exp(B) = .670* beim SVS zeigt zum Beispiel an, dass hier ein jeweils zusätzlich erzielter Punkt die Wahrscheinlichkeit einer Schwäche im Bereich des Schriftspracherwerbs um $100 \cdot (1 - 0.670) = 33.0\%$ senkt. Für den IQ sind dies demnach 24.2% und für die Phonem-Synthese 12.7% .

Das Prädiktoren-Set aus IQ, SVS und PS diene als Grundlage für alle weiteren Analysen. In einem weiteren Auswertungsschritt wurde die grundsätzliche Klassifikationsgenauigkeit des Prognose-Modells für unterschiedliche Kriterium-Cutoffs auf Basis jeweils einer logistischen Regressionsanalyse (Einschlussverfahren) mit anschließender ROC-Analyse ermittelt (s. Tabelle 2).

Wie hierzu aus Tabelle 2 zu ersehen ist, ergeben sich gute AUC-Werte ($\geq .80$) und angemessene Modell-Fits (HL-Test).

Um die Stabilität des IQ-SVS-PS-Modells zu überprüfen (Hypothese 2), wurde eine doppelte Kreuzvalidierung vorgenommen. Tabelle 3 gibt jeweils zeilenweise darüber Auskunft, wie sich die Modellparameter der entsprechenden Lerngruppe ($N = 81$) auf die Klassifikationsgüte für die jeweilige Testgruppe ($N = 93$) und umgekehrt auswirken.

Wie aus Tabelle 3 zu ersehen ist, ergeben sich sowohl für die entsprechenden Lern- als auch für die Testgruppen bis auf wenige Ausnahmen gute AUC-Werte ($\geq .80$), die sich jeweils bis auf eine Ausnahme (Lerngruppe A \rightarrow Testgruppe B, $PR \leq 10$) nicht statistisch signifikant voneinander unterscheiden (Hanley & Mc Neil, 1983; Hajian-Tilaki, 2018). Dies weist auf eine gute Stabilität des IQ-SVS-PS-Modells hin. Anzeichen für ein Overfitting sind nicht erkennbar.

In einem dritten Auswertungsschritt wurde Hypothese 3 überprüft, nach der sich vor dem Hintergrund eines logistischen Regressionsmodells auf Item-Ebene (formatives Modell, Tabelle 4) bessere Prognosebefunde erzielen lassen sollten als auf der Ebene der Summenwerte (Latent-Trait-Modell) der Untertests (Tabelle 2).

Tabelle 2: Ergebnisse der logistischen Regressionsanalysen (Summenwerte der einzelnen Prädiktoren = Latent-Trait-Modell; Einschlussverfahren) mit anschließender ROC-Analyse auf der Basis unterschiedlicher Cutoffs für das Kriterium Schriftsprachkompetenz ($N = 173$)

Prädiktoren	Cutoff Kriterium	X ² Omnibus (p)	Nagelkerkes R ²	HL-Test ($p =$)	AUC-Wert	95% KI für AUC
IQ, SVS, PS	$PR \leq 10$.000	.359	.519	.86	[.78, .93]
IQ, SVS, PS	$PR \leq 16$.000	.354	.114	.83	[.77, .90]
IQ, SVS, PS	$PR \leq 25$.000	.347	.166	.81	[.74, .88]

Anmerkungen. IQ = CFT-Matrizen; SVS = Satzverstehen; PS = Phonemsynthese; X² Omnibus (p) = Signifikanz des Gesamtmodells (X²-Test); HL-Test = Hosmer-Lemeshow-Test; KI = Konfidenz-Intervall; AUC = Fläche unter der ROC-Kurve

Tabelle 3: Ergebnisse einer doppelten Kreuzvalidierung des IQ-SVS-PS-Modells durch einen Vergleich der ROC-Kurven vor dem Hintergrund unterschiedlicher Kriterium-Cutoffs

Gruppe	Cutoff	AUC-Wert	95% KI für AUC	Gruppe	AUC-Wert	95% KI für AUC	p (zweis.) Differenz
Lern A	$PR \leq 10$.93	[.85, .99]	Test B	.79	[.68, .90]	.047
Lern B	$PR \leq 10$.79	[.68, .91]	Test A	.92	[.84, .99]	.069
Lern A	$PR \leq 16$.89	[.81, .98]	Test B	.80	[.70, .90]	.161
Lern B	$PR \leq 16$.81	[.71, .90]	Test A	.84	[.74, .94]	.604
Lern A	$PR \leq 25$.87	[.79, .96]	Test B	.74	[.63, .85]	.064
Lern B	$PR \leq 25$.77	[.67, .87]	Test A	.82	[.71, .93]	.523

Anmerkungen. Cutoff = Kriterium-Cutoff; Lern = Lerngruppe; Test = Testgruppe

Tabelle 4: Ergebnisse der logistischen Regressionsanalysen (Item-Ebene der Prädiktoren = formatives Messmodell; Einschlussverfahren) mit anschließender ROC-Analyse auf der Basis unterschiedlicher Cutoffs für das Kriterium Schriftsprachkompetenz (N = 173)

Prädiktoren	Cutoff Kriterium	X ² Omnibus (p =)	Nagelkerkes R ²	HL-Test (p =)	AUC-Wert	95% KI für AUC
IQ, SVS, PS	PR ≤ 10	.000	.695	.309	.96	[.92, 1.00]
IQ, SVS, PS	PR ≤ 16	.000	.597	.512	.93	[.88, .97]
IQ, SVS, PS	PR ≤ 25	.000	.567	.885	.90	[.85, .95]

Anmerkungen. IQ = CFT-Matrizen; SVS = Satzverstehen; PS = Phonemsynthese; X² Omnibus (p) = Signifikanz des Gesamtmodells (X²-Test); HL-Test = Hosmer-Lemeshow-Test; KI = Konfidenz-Intervall; AUC = Fläche unter der ROC-Kurve

Tabelle 5: Güte-Kennwerte der klassifikatorischen Vorhersage für Kinder mit schwacher Schriftsprachkompetenz (PR ≤ 16) zu Beginn der 2. Klasse hinsichtlich des Prädiktoren-Sets aus IQ, SVS und PS auf Item-Ebene (formatives Messmodell) vor dem Hintergrund unterschiedlicher Risiko-Schwellenwerte p (y=1)

Kriterium	p (y=1) ≥	SQ	GT	SN	SP	PK	NK	LR+	OR	RATZ
Schriftsprachkompetenz	.10	43.9	73.4	94.1	68.3	42.1	97.9	3.0	34.5	89.5
	.19	30.0	86.1	91.2	84.9	59.6	97.5	6.0	58.1	87.4
PR ≤ 16	.25	25.4	87.3	82.4	88.5	63.6	95.3	7.2	35.9	76.3
N = 173	.35	20.2	87.9	70.6	92.1	68.6	92.8	8.9	27.9	63.1
GR = 19.7%	.45	16.2	88.4	61.8	95.0	75.0	91.0	12.3	30.5	68.9

Anmerkungen. GR = Grundrate (Prävalenz); SQ = Selektionsquote; GT = Gesamt-Trefferquote; SN = Sensitivität; SP = Spezifität; PK = Positive Korrektheit (positiv prädiktiver Wert); NK = Negative Korrektheit (negativ prädiktiver Wert); LR+ = Positives Likelihood Ratio; OR = Odds Ratio; RATZ = RATZ-Index

Wie schon durch Augenschein beim Vergleich der Tabellen 2 und 4 ersichtlich wird, ergeben sich für die Analysen auf Item-Ebene durchweg höhere R²-Werte (Nagelkerke) und im Schnitt um ca. 0.10 Punkte höhere AUC-Werte.

Überprüft man die Differenz der ROC-Kurven (AUC-Werte) als Resultat der beiden Messmodelle zufallskritisch (Hanley & McNeil, 1983; Hajian-Tilaki, 2018), erweist sich die AUC-Wert-Differenz von 0.10 bei einem Cutoff = PR ≤ 16 als hoch signifikant (z = -3.09; p [zweis.] = .002). Dies gilt in ähnlichem Ausmaß für alle anderen Kriterium-Cutoffs.

In einem weiteren Auswertungsschritt wurde ermittelt, welche der gängigen Güte-Aussagen (Sensitivität, Spezifität, RATZ-Index etc.) über das IQ-SVS-PS-Modell vor dem Hintergrund des durch die ROC-Analyse ermittelten optimalen Prädiktor-Cutoffs von .19 sowie benachbarte Werte gemacht werden können. Diese Güte-Kennwerte sind exemplarisch in Tabelle 5 auf Item-Ebene auf der Basis des Kriterium-Cutoffs von PR ≤ 16 dargestellt.

Betrachtet man die Gütekriterien vor dem Hintergrund des durch die ROC-Analyse ermittelten optimal trennenden Risiko-Cutoff (dritte Zeile), kann bei einer SQ = 30% mit einer GT = 86.1%, einer sehr guten SN =

91.2%, einer $SP = 84.9\%$ sowie einer guten Prognose-Sicherheit ($LR+ = 6.0$) gerechnet werden. Letztere wird sowohl durch eine hohe Vorhersagegenauigkeit ($OR = 58.1$) sowie einer PK von knapp 60% und einer NK von knapp 98% unterstrichen.

Der *RATZ-Index* liegt mit 87.4% im sehr guten spezifischen Bereich. Im Gegensatz zu den etwas höheren Prädiktor-Cutoffs (z.B. $p [y=1] \geq .25$) liegt hier jedoch im Vergleich zur Grundrate ($GR = 19.7\%$) eine gewisse Überselektion vor.

Aus Tabelle 5 wird auch der grundsätzliche Einfluss deutlich, den unterschiedliche Schwellenwerte (Risiko-Cutoffs) auf die Güte-Werte eines Screening-Verfahrens haben: Relativ niedrige Cutoffs führen zwar zu relativ hohen Sensitivitäten, „erkaufen“ sich diese jedoch mit einer niedrigeren Sicherheit (PK , $LR+$). Der über eine ROC-Analyse ermittelte Cutoff führt in der Regel zu einer gewissen Ausgeglichenheit zwischen Sensitivität und Spezifität und kann einen tragbaren Kompromiss darstellen.

Zusammenfassung und Diskussion

Im Rahmen der Schuleingangsdiagnostik (Prävention) sollten Filter-Screenings (Tröster, 2009, S. 69) zum Einsatz kommen, die nicht nur das Kriterium der prognostischen Validität als zentrales Kriterium erfüllen (Schneider & Hasselhorn, 2018), sondern sich auch zeitökonomisch zum Einsatz bringen lassen.

Dazu wurde die Entwicklung und Evaluation eines (abgesehen vom Matrizen-test) letztlich komplett neu entwickelten Filter-Screenings beschrieben. Im Rahmen der Entwicklung des Verfahrens standen die Autoren vor der Herausforderung, Aufgabenformate zu entwickeln, die anders als bei vielen gängigen Screening-Verfahren ganzen Gruppen von Kindern zu Beginn der ersten Klasse präsentiert werden können. Ziel war es zu eruieren (Hypothese 1), ob sich aus einem breiten Set von im Prinzip bewährten Prädiktorvariablen heraus

(Abbildung 1) per Datenreduktion und auf stabile Art und Weise (Hypothese 2) mithilfe einer schrittweisen logistischen Regressionsanalyse eine überschaubare und damit für die Praxis unter vertretbarem Aufwand durchführbare Screening-Prozedur entwickeln lässt, die eine statistisch signifikante und praktisch relevante Klassifikation von Risiko- und Nicht-Risiko-Kindern bezüglich ihrer Schriftsprachkompetenz über einen Prognosezeitraum von ca. einem Jahr erlaubt.

Darüber hinaus sollte überprüft werden (Hypothese 3), ob sich die prognostische Güte des Verfahrens tatsächlich dadurch steigern lässt, indem man ein formatives Messmodell anstelle des üblicherweise bevorzugten Latent-Trait-Messmodell zugrunde legt.

Als Ergebnis konnte ein Prognosemodell, bestehend aus einem Dreier-Set von Variablen, nämlich dem IQ (CFT-Matrizen), der linguistischen Kompetenz (Satzverstehen, SVS) sowie einem Aspekt der phonologischen Bewusstheit (Phonem-Synthese, PS) identifiziert werden (Tabelle 1). Die Parameter dieses Modells sowie seine Einzelkomponenten sind statistisch hoch signifikant von Null verschieden und ermöglichen vor dem Hintergrund der drei gewählten Kriterium-Cutoffs ($PR \leq 10$, $PR \leq 16$, $PR \leq 25$) jeweilige *AUC*-Werte von .86, .83 und .81 (Tabelle 2), die durchweg als gut bezeichnet werden können (Catts et al., 2015, S. 281).

Die Ergebnisse einer doppelten Kreuzvalidierung des IQ-SVS-PS-Modells durch einen Vergleich der ROC-Kurven an jeweils unabhängigen Stichproben (Tabelle 3) weisen auf eine gute Stabilität des Prädiktor-Modells sowie auf ein nicht vorhandenes Overfitting hin. Damit können die Hypothesen 1 und 2 durchaus als bestätigt betrachtet werden.

Als bemerkenswerter Befund ist festzuhalten, dass sich die Annahmen aus den Arbeiten von Seeboth und Möttus (2018), Möttus und Rozgonjuk (2021) sowie Schroeders et al. (2020) bezüglich der Vorteile der Realisierung eines formativen Messmodells auch

im Rahmen der hier durchgeführten Schulleistungsmessung bestätigen lassen (Hypothese 3, Tabelle 4). Auch wenn in diesem Zusammenhang der prognostische Beitrag einzelner Items im Einzelfall zu schwach war, um im Rahmen der logistischen Regressionsanalyse statistische Signifikanz zu erreichen (aus Platzgründen nicht dargestellt), so können jedoch die insgesamt 40 Items aus den drei Skalen (IQ, SVS, PS) bei ansonsten guten Modell-Fits im Schnitt um ca. jeweils 0.10 höhere *AUC*-Werte liefern als die jeweiligen Analysen auf der Basis von Summenwerten (Tabellen 2 und 4). Dies führt insgesamt zu einer statistisch hoch signifikanten Verbesserung der Klassifikationsgüte. Numerisch steigen die *AUC*-Werte dadurch in einen Bereich von $>.90$, was als ein hervorragendes Klassifikationsresultat zu bewerten ist (Catts et al., 2015, S. 281).

Legt man vor dem Hintergrund des formativen Messmodells auf der Basis der logistischen Regression und einen durch eine ROC-Analyse ermittelten am besten trennenden Risiko-Cutoff von $.19$ zugrunde, lässt sich ein sehr guter *RATZ-Index* von 87.4% sowie eine gute Balance zwischen Sensitivität (91.2%) und Spezifität (84.9%) erreichen (Tabelle 5). Diese Güte-Indizes liegen damit durchweg über denjenigen der Verfahren, die weiter oben im Rahmen der Einleitung genannt wurden. Möglicherweise bilden einzelne Items tatsächlich unterschiedlich prognoserelevante Kompetenz-Nuancen ab, die bei einer Summenbildung (Aggregation) nicht mehr ihre volle Wirkung entfalten können. Dieser Vermutung sollte bezüglich schulisch relevanter Basiskompetenzen in Zukunft auf jeden Fall detaillierter nachgegangen werden. Die aktuelle Datenlage lässt aufgrund des Verhältnisses der Probanden- zur Variablenanzahl die Kreuzvalidierung der Modellparameter auf der Basis eines formativen Messmodells noch nicht zu, was jedoch im Rahmen einer zweiten Untersuchungswelle geplant ist.

Insgesamt kann resümiert werden, dass es sich als vorteilhaft herausgestellt hat,

sich bei der Konzeption von Untertests zur Erfassung prognoserelevanter Fähigkeiten an den bewährten Konzepten der phonologischen Informationsverarbeitung, der linguistischen Kompetenz sowie der nonverbalen Denkfähigkeit (IQ) zu orientieren (Abbildung 1). Die Identifikation des IQ-SVS-PS-Modells mit den günstigen Güteparametern spricht für sich und bestätigt das Modell von Ennemoser et al. (2012). Insofern ist die hier realisierte Vorgehensweise als deduktiv zu bezeichnen. Andererseits werden sowohl in der Grundlagenforschung, bei Interventionsstudien als auch im Rahmen der Entwicklung von Screening-Verfahren unterschiedliche Operationalisierungen für dieser Konzepte vorgenommen. So beschreibt beispielsweise Lewkowicz (1980) zehn verschiedene Aufgabentypen zur Erfassung bzw. zur Förderung der phonologischen Bewusstheit. Der nonverbale IQ im CFT-1 wird z.B. in Form von fünf Untertests ermittelt (Weiß & Osterland, 1997). Die linguistische Kompetenz im Rahmen des Modells von Ennemoser et al. (2012) wird z.B. durch ein Bündel verschiedener Untertests (Wortschatz, Bildung von Ableitungsmorphemen, Korrektur semantischer Inkonsistenzen, Benennungsflexibilität, Verstehen von Sätzen) aus zwei unterschiedlichen Sprachentwicklungstests durch einem gemeinsamen Summenwert (*z*-transformiert) operationalisiert. Analog zu der geschilderten unterschiedlichen Operationalisierung von Konzepten und dem Ziel dieser Untersuchung, ein zeitökonomisch und möglichst einfach durchführbares Screening-Instrument auf Gruppenbasis zu entwickeln, kann der Umstand, dass hier zunächst eine gut begründete Auswahl von zehn unterschiedlichen Untertests als „Startmodell“ für eine schrittweise logistische Regressionsanalyse eingeführt wurde, in gewissem Sinne als induktives Vorgehen bezeichnet werden.

Abweichend vom Strukturgleichungsmodell in Abbildung 1 wurde aus Gründen der Praktikabilität die Benennungsgeschwindigkeit nicht erfasst, was sich nicht negativ auf

die Prognosequalität auswirkt. Das Arbeitsgedächtnis (AG) erweist sich zwar nicht als signifikanter Einzelprädiktor, was aber keineswegs bedeutet, dass es keine Rolle spielen würde (Alloway, 2009; Alloway & Alloway, 2010). In separaten Analysen wurde der IQ-Wert durch den AG-Wert ersetzt, was zu keiner signifikanten Reduktion der Prognosequalität führte.

Auf der Grundlage des IQ-SVS-PS-Modells kann die Screening-Prozedur als Gruppentest an bis zu 15 Kindern innerhalb von 45 Minuten mit hoher prognostischer Validität durchgeführt werden. Zu betonen ist, dass ein Filter-Screening keine Diagnose darstellt, so dass bei Kindern mit einem positiven Screening-Befund ein individualdiagnostisches Vorgehen angesagt ist. Auf jeden Fall sollten dann gezielte Fördermaßnahmen auf der Basis evidenzbasierter Konzepte/Programme initiiert werden (Ise et al., 2012).

Die geschilderten Befunde sind jedoch mit gewissen Restriktionen versehen. Auch wenn die Prognosegüte der Prädiktoren im Rahmen einer Kreuzvalidierung an unabhängigen Gruppen auf der Basis der Summenwerte der Untertests grundsätzlich nachgewiesen werden konnte, ist eine weitere Überprüfung notwendig und auch geplant. Des Weiteren macht die in das formative Messmodell aufgenommene Anzahl der Items auch bei aktuell gutem Modell-Fit eine Vergrößerung der Stichprobe notwendig, um auch hier eine Kreuzvalidierung vornehmen zu können. Dies war aufgrund des Probanden-Variablen-Verhältnisses im Rahmen der vorliegenden Untersuchung nur vor dem Hintergrund des Latent-Trait-Messmodells möglich.

Des Weiteren basiert die vorliegende Analyse auf einer vergleichsweise im Durchschnitt schwach lesenden Stichprobe (ELFE-Gesamt: $T = 41.6$, $SD = 10.5$) mit insgesamt dennoch akzeptablen Risiko-Prävalenzraten. Da Lese- und Rechtschreibkompetenz stark übungsabhängig ist, könnte es jedoch sein, dass durch die pandemiebedingten Unterrichtsausfälle diejenigen Kinder, die schwach gestartet sind, nicht

zureichend gefördert wurden. Dies mag die aktuell ermittelte Anzahl der richtig-positiven Fälle im Vergleich zu einer „normalen“ Beschulungssituation erhöht haben, weil in dieser eine größere Chance besteht, dass anfänglich schwache Kinder im Laufe des Schuljahres aus dem Bereich einer unterdurchschnittlichen Leistungsfähigkeit herausgeführt werden können. Genau diesen Mechanismus beschreiben Schabmann et al. (2009) im Kontext von Unterrichtsqualität. Das ändert jedoch nichts an der grundsätzlichen Möglichkeit des beschriebenen Screening-Verfahrens, risikobehaftete Kinder am Anfang der Grundschulzeit zuverlässig zu identifizieren.

Literatur

- Ahmed, Y., Wagner, R. K. & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology, 106*(2), 419–434. <https://doi.org/10.1037/a0035692>
- Alloway, T. P. (2009). Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment, 25*(2), 92–98.
- Alloway, T. P. & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20–29.
- Babu, S. (2015). Various performance measures in binary classification – An overview of ROC study. *International Journal of Innovative Science, Engineering & Technology, 9*(2), 596–605.
- Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2018). *Multivariate Analysemethoden* (15. Aufl.). Springer.
- Badian, N., McNulty, G., Duffy, G. & Als, H. S. (1990). Prediction of dyslexia in kindergarten boys. *Annals of Dyslexia, 40*(1), 152–167.


- Barth, K. & Gomm, B. (2014). *Gruppentest zur Früherkennung von Lese- und Rechtschreibschwierigkeiten (PB-LRS)*. Reinhardt.
- Bender, R. (2001). Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung. *Medizinische Klinik, 96*(2), 116–121.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte Auflage). Pearson.
- Catts, H. W., Fey, M. E., Zhang, X. & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32*(1), 38–50.
- Catts, H. W. & Hogan, T. (2003). Language basis of reading disabilities and implications for early identification and remediation. *Reading Psychology, 24*(3–4), 223–246. <https://doi.org/10.1080/02702710390227314>
- Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S. & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3), 281–297. <https://doi.org/10.1177/0022219413498115>
- Catts, H. W., Nielsen, D. C., Bridges, M. S. & Liu, Y. S. (2016). Early identification of reading comprehension difficulties. *Journal of Learning Disabilities, 49*(5), 451–465. <https://doi.org/10.1177/0022219414556121>
- Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S. & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163–176. <https://doi.org/10.1177/0022219408326219>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum Associates.
- Daseking, M. & Petermann, F. (2009). *BA-SIC-Preschool. Screening für kognitive Basiskompetenzen im Vorschulalter*. Huber.
- Ehri, L., Nunes, S. R., Willows, D. M., Schuster, B. V., Yakhoub-Zadeh, Z. & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36*(3), 250–287.
- Endlich, D., Berger, N., Küspert, W., Lenhard, W., Marx, P., Weber, J. et al. (2016). *Würzburger Vorschultest (WVT). Erfassung schriftsprachlicher und mathematischer (Vorläufer-)Fertigkeiten und sprachlicher Kompetenzen im letzten Kindergartenjahr*. Hogrefe.
- Endlich, D., Küspert, P., Lenhard, W., Marx, P. & Schneider, W. (2019). *LRS-Screening. Laute, Reime, Sprache – Würzburger Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten*. Hogrefe.
- Ennemoser, M., Marx, P., Weber, J. & Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*(2), 53–67. <https://doi.org/10.1026/0049-8637/a000057>
- Fischbach, A., Schuchardt, K., Brandenburg, J., Kleczewski, J., Balke-Melcher, Chr., Schmidt, C. et al. (2013). Prävalenz von Lernschwächen und Lernstörungen: Zur Bedeutung der Diagnosekriterien. *Lernen und Lernstörungen, 2*(2), 65–76. <https://doi.org/10.1024/2235-0977/a000035>
- Fischer, M. Y. & Pfost, M. (2015). Wie effektiv sind Maßnahmen zur Förderung der phonologischen Bewusstheit? Eine meta-analytische Untersuchung der Auswirkungen deutschsprachiger Trainingsprogramme auf den Schriftspracherwerb. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(1), 35–51. <https://doi.org/10.1026/0049-8637/a000121>


- Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K. et al. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86*(1), 6–23. <https://doi.org/10.1037/0022-0663.86.1.6>
- Fritz, A., Ricken, G. & Gerlach, M. (2013). *Kalkulie. Diagnose- und Trainingsprogramm für rechenschwache Kinder* (4. Aufl.). Cornelsen.
- Gallagher, A., Frith, U. & Snowling, M. J. (2000). Precursors of literacy delay among children at genetic risk of dyslexia. *Journal of Child Psychology and Psychiatry, 41*(2), 203–213.
- Glück, C.W. (2007). *Wortschatz- und Wortfindungstest für 6- bis 10-Jährige (WWT 6-10)*. Urban & Fischer.
- Goldammer, A. von, Mähler, C., Bockmann, A-K. & Hasselhorn, M. (2010). Vorhersage früher Schriftsprachleistungen aus vorschulischen Kompetenzen der Sprache und der phonologischen Informationsverarbeitung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*(1), 48–56. <https://doi.org/10.1026/0049-8637/a000005>
- Gorecki, B. & Landerl, K. (2015). Ist die phonologische Bewusstheit ein Prädiktor für die Leseleistung? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(3), 139–146. <https://doi.org/10.1026/0049-8637/a000135>
- Gough, P. B. & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6–10.
- Grimm, H. (2001). *SETK 3-5. Sprachentwicklungstest für drei- bis fünfjährige Kinder*. Hogrefe.
- Grube, D. & Hasselhorn, M. (2006). Längsschnittliche Analysen zur Lese-, Rechtschreib- und Mathematikleistung im Grundschulalter: zur Rolle von Vorwissen, Intelligenz, phonologischem Arbeitsgedächtnis und phonologischer Bewusstheit. In I. Hosenfeld & F.-W. Schrader (Hrsg.), *Schulische Leistung. Grundlagen, Bedingungen, Perspektiven* (S. 87–105). Waxmann.
- Hajian-Tilaki, K. (2018). Receiver operator characteristic analysis of biomarkers evaluation in diagnostic research. *Journal of Clinical and Diagnostic Research, 12*(6), LE01-LE08. <https://doi.org/10.7860/JCDR/2018/32856.11609>
- Hanley, J. A. & Mc Neil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*(3), 839–843.
- Hayiou-Thomas, M. E., Harlaar, N., Dale, P. S. & Plomin, R. (2006). Genetic and environmental mediation of the prediction from preschool language and nonverbal ability to 7-year reading. *Journal of Research in Reading, 29*(1), 50–74.
- Hoover, W. A. & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*(2), 127–160.
- Ise, E., Engel, R.R. & Schulte-Körne, G. (2012). Was hilft bei der Lese-Rechtschreibstörung? Ergebnisse einer Metaanalyse zur Wirksamkeit deutschsprachiger Förderansätze. *Kindheit und Entwicklung, 21*(2), 122–136. <https://doi.org/10.1026/0942-5403/a000077>
- Jansen, H., Mannhaupt, G., Marx, H. & Skowronek, H. (2002). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC)* (2., überarbeitete Auflage). Hogrefe.
- Klicpera, C. & Schabmann, A. (1993). Do German-speaking children have a chance to overcome reading and spelling difficulties? A longitudinal survey from the second until the eighth grade. *European Journal of Psychology of Education, 8*(3), 307–323.

- Klicpera, C., Schabmann, A. & Gasteiger-Klicpera, B. (2006). Die mittelfristige Entwicklung von Schülern mit Teilleistungsschwierigkeiten im Bereich der Lese- und Rechtschreibschwierigkeiten. *Kindheit und Entwicklung, 15*(4), 216–227. <https://doi.org/10.1026/0942-5403.15.4.216>
- Kohn, J., Wyschkon, A., Ballaschk, K., Ihle, W. & Esser, G. (2013). Verlauf von Umschriebenen Entwicklungsstörungen: Eine 30-Monats-Follow-up-Studie. *Lernen und Lernstörungen, 2*(2), 77–89. <https://doi.org/10.1024/2235-0977/a000032>
- Kretschmann, R. (2007). Prävention. Schulalter. In J. Walter & F. B. Wember (Hrsg.), *Sonderpädagogik des Lernens* (Handbuch der Sonderpädagogik, Band 2, S. 245–266). Hogrefe.
- Landerl, K. & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150–161. <https://doi.org/10.1037/0022-0663.100.1.150>
- Lenhard, W., Lenhard, A. & Schneider, W. (2017). *Ein Leseverständnistest für Erst- bis Siebtklässler (ELFE II)*. Hogrefe.
- Lewkowicz, N.K. (1980). Phonemic awareness training. What to teach and how to teach it. *Journal of Educational Psychology, 72*(5), 686–700. <https://doi.org/10.1037/0022-0663.72.5.686>
- Martschinke, S., Kirschhock, E.-M. & Frank, A. (2001). *Diagnose und Förderung im Schriftspracherwerb. Der Rundgang durch Hörhäuser. Band 1: Erhebungsverfahren zur phonologischen Bewusstheit*. Auer.
- Marx, P. & Weber, J. (2006). Vorschulische Vorhersage von Lese- und Rechtschreibschwierigkeiten. Neue Befunde zur prognostischen Validität des Bielefelder Screenings (BISC). *Zeitschrift für Pädagogische Psychologie, 20*(4), 251–259. <https://doi.org/10.1024/1010-0652.20.4.251>
- Mayer, A. (2018). Benennungsgeschwindigkeit und Lesen. *Forschung Sprache, 6*, 20–42. https://www.forschung-sprache.eu/fileadmin/user_upload/Dateien/Heftausgaben/2018-1/5-70-2018-01-02.pdf
- Mayer, A. & Motsch, H.-J. (2014). Früherkennung von Schriftspracherwerbsstörungen – Zur prognostischen Validität des TE-PHOBE. *Praxis Sprache, 59*(2), 218–227.
- Mazzocco, M. M. M. & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*(3), 142–155. <https://doi.org/10.1111/j.1540-5826.2005.00129.x>
- Möttus, R. & Rozgonjuk, D. (2021). Development is in the details: Age differences in the Big Five domains, facets and nuances. *Journal of Personality and Social Psychology, 120*(4), 1035–1048. <https://doi.org/10.1037/pspp0000276>
- Müller, R. (2004). *Diagnostischer Rechen-test für 1. Klassen (DRT 1)* (2., aktualisierte Auflage). Hogrefe.
- Peng, P., Wang, T., Wang, C. & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin, 145*(2), 189–236. <https://doi.org/10.1037/bul0000182>
- Peng, P. & Fuchs, D. (2016). A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? *Journal of Learning Disabilities, 49*(1), 3–20. <https://doi.org/10.1177/0022219414521667>
- Poulsen, M., Nielsen, A.-M. V., Juul, H. & Elbro, C. (2017). Early identification of reading difficulties: a screening strategy that adjusts the sensitivity to the level of prediction accuracy. *Dyslexia, 23*(3), 251–267. <https://doi.org/10.1002/dys.1560>

- Roos, J., Schöler, H. & Treutlein, A. (2007). *Zur prognostischen Validität des Heidelberger Auditiven Screenings in der Einschulungsdiagnostik HASE*. Abschlussbericht des Projektes EVER. Heidelberg: Pädagogische Hochschule.
- Rudolf, M. & Müller, J. (2012). *Multivariate Verfahren* (2., überarbeitete Auflage). Hogrefe.
- Sassu, R. & Roebers, C. M. (2016). A multidimensional view of children's school readiness. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 48(3), 144–157. <https://doi.org/10.1026/0049-8637/a000154>
- Schabmann, A., Schmidt, B. M., Klicpera, Chr., Gasteiger-Klicpera, B. & Klingebiel, K. (2009). Does systematic reading instruction impede prediction of reading in a shallow orthography? *Psychology Science Quarterly*, 51(3), 315–338.
- Schatschneider, Ch., Fletcher, J. M., Francis, D. J., Carlson, C. D. & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, 96(2), 265–282. <https://doi.org/10.1037/0022-0663.96.2.265>
- Schneider, W. & Hasselhorn, M. (Hrsg.). (2018). *Schuleingangsdiagnostik*. Hogrefe.
- Schöler, H. & Brunner, M. (2008). *HASE. Heidelberger Auditives Screening in der Einschulungsuntersuchung* (2. Aufl.). Westra.
- Schroeders, U., Watrin, L. & Wilhelm, O. (2020). Getting older does not make you smarter - A more nuanced view on crystallized intelligence. <https://www.researchgate.net/publication/344271790>
- Seeboth, A. & Möttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, 32(3), 186–201. <https://doi.org/10.1002/per.2147>
- Tiu R. D., Thompson L. A. & Lewis, B. A. (2003). The role of IQ in a component model of reading. *Journal of Learning Disabilities*, 36(5), 424–436. <https://doi.org/10.1177/00222194030360050401>
- Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter. Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Hogrefe.
- Wagner, R. K. & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101(2), 192–212. <https://doi.org/10.1037/0033-2909.101.2.192>
- Weiß, R. & Osterland, J. (1997). *Grundintelligenztest Skala 1 – CFT 1* (5., revidierte Auflage). Hogrefe.
- Youngstrom, E. A. (2014). A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, 39(2), 204–221. <https://doi.org/10.1093/jpepsy/jst062>

AutorInnenhinweise

 Jürgen Walter
<https://orcid.org/0000-0003-0753-1499>

 Kristina Clausen-Suhr
<https://orcid.org/0000-0002-3316-9748>

Korrespondenzadresse:

Prof. Dr. Jürgen Walter und
Dr. Kristina Clausen-Suhr
Institut für Sonderpädagogik der Europa-Universität Flensburg
Abteilung Sonderpädagogik des Lernens

E-Mail:

walter@uni-flensburg.de

kristina.clausen-suhr@uni-flensburg.de