technische universität dortmund

Fakultät Rehabilitations-wissenschaften

Rechenstörungen im Grundschulalter: Entwicklungsdefizite und -verzögerungen, Risikoerkennung und Einflussfaktoren auf die diagnostische Einschätzung von Lehrkräften

vorgelegt von

Sarah Lamb

Dissertation zur Erlangung des Grades einer Doktorin der Philosophie (Dr. phil.)

in der

Fakultät Rehabilitationswissenschaften

der Technischen Universität Dortmund

Dortmund

2025

Betreuer: Prof. Dr. Jörg-Tobias Kuhn

Betreuerin: Prof. Dr. Janin Brandenburg

Die vorliegende Arbeit wurde von der Fakultät Rehabilitationswissenschaften der Technischen Universität Dortmund als Dissertation angenommen.

Gutachter: Prof. Dr. Jörg-Tobias Kuhn

Gutachterin: Prof. Dr. Janin Brandenburg

Tag der Disputation: 14.08.2025

Dortmund

Danksagung

Meine Promotionszeit an der Fakultät Rehabilitationswissenschaften waren für mich nicht nur akademisch, sondern auch persönlich eine prägende und inspirierende Erfahrung. Dies verdanke ich den vielen wunderbaren Menschen, die mich auf diesem Weg begleitet, unterstützt und an mich geglaubt haben.

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. Jörg-Tobias Kuhn. Seine konstante Unterstützung, seine fachliche Expertise und sein Vertrauen in meine Arbeit haben diese Dissertation erst möglich gemacht. Mit seinen wertvollen Anregungen und seinem konstruktiven Feedback hat er zum Entstehen dieser Arbeit maßgeblich beigetragen. Bedanken möchte ich mich auch bei Prof. Dr. Janin Brandenburg und Prof. Dr. Jan Kuhl für die Betreuung meiner Dissertation. Ann-Katrin Schulz und Dr. Florian Krieger danke ich für ihre Mitwirkung an den Publikationen!

An dieser Stelle möchte ich mich bei meinen Kolleg*innen am Fachgebiet der Methoden der empirischen Bildungsforschung und der Rehabilitationssoziologie bedanken, die mich in den wichtigen Phasen meiner Promotion entlastet, unterstützt und für eine inspirierende und motivierende Arbeitsatmosphäre gesorgt haben! Besonders danke ich: Dr. Teresa Sartor, Christian Kißler, Dr. Jana York, Dr. Sarah Schulze und Vertr.-Prof. PD Dr. Bastian Pelka.

Nicht zuletzt gilt mein tiefer Dank meiner Familie und meinen Freund*innen, die mich emotional in den Höhen und Tiefen dieser Zeit begleitet und jeden Meilenstein mit mir gefeiert haben. Ihr Verständnis und ihre aufmunternden Worte haben mir Kraft gegeben.

Diese Arbeit wäre ohne die Unterstützung dieser Menschen nicht möglich gewesen – ihnen allen gilt mein herzlichster Dank!

Inhaltsverzeichnis

Zusammenfassung

Kinder mit einer Rechenstörung sollten frühzeitig identifiziert und in ihren mathematikspezifischen Fertigkeiten gefördert werden, um einen persistierenden Entwicklungsverlauf zu verhindern und langfristige negative Folgen abzumildern. Die Hauptverantwortung für diese Aufgabe liegt in erster Linie bei den Bildungsinstitutionen, insbesondere der Schule und damit bei den Lehrkräften. Daher untersucht die vorliegende Dissertation, *wie Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren können.*

Evidenzbasiertes Wissen über die Ursachen, Schwierigkeiten und Entwicklungsverläufe von Rechenstörungen ist eine notwendige Voraussetzung für eine erfolgreiche Risikoidentifikation. Allerdings sind die Ursachen der Rechenstörung nicht abschließend geklärt und auch bezüglich des Entwicklungsverlaufs bestehen offene Fragen. Daher untersuchte *Studie I*, ob die Defizite in den basisnumerischen Fertigkeiten von Kindern mit einer Rechenstörung eher auf eine Entwicklungsverzögerung oder auf eine rechenstörungsspezifische qualitative Abweichung hindeuten und, ob diese Defizite auf eine (selektive) Beeinträchtigung in einem der zentralen Kernsysteme der Zahlenverarbeitung zurückzuführen sind. Dazu wurden die basisnumerischen Fertigkeiten von $N = 480$ Kindern ($n = 68$ Kinder mit Rechenstörung, Klassenstufe zwei bis vier) untersucht. Multilevel-Analysen legen nahe, dass die Beeinträchtigungen in der Basisnumerik auf eine Entwicklungsverzögerung hinweisen, die eher aus einem Defizit in der approximativen Mengenverarbeitung (Approximate Number System) und weniger aus einem Defizit in der Verarbeitung abstrakt dargestellter Mengen in Form von Symbolen (Access Deficit) resultieren. Es gab keinen Hinweis auf ein spezifisches Defizit in der exakten Repräsentation kleinerer Mengen (Object Tracking System).

Vor dem Hintergrund dieser Erkenntnisse wurde in *Studie II* ein theoriebasierter Screeningfragebogen für Lehrkräfte zur Identifikation von Kindern mit einem Risiko für

eine Rechenstörung vorgestellt und hinsichtlich seiner psychometrischen Eigenschaften untersucht. Dazu wurden die Daten von N = 377 Schüler*innen (n = 45 Kinder mit Anzeichen für eine Rechenstörung) der Klassenstufe zwei bis vier und N = 33 Lehrkräften erhoben. Die Ergebnisse der psychometrischen Analysen zeigten, dass der Fragebogen zur Erfassung mathematischer Fertigkeiten (FERMAT) über gute bis zufriedenstellende Screeningeigenschaften verfügt und Kinder mit und ohne testdiagnostische Anzeichen für eine Rechenstörung ökonomisch, reliabel und valide identifiziert.

Da die diagnostischen Einschätzungen der Lehrkräfte allgemein auf urteilsrelevanten und -irrelevanten Informationen beruhen, wurde in *Studie III* (auf Basis der Daten aus Studie II) untersucht, welche mathematikspezifischen und -unspezifischen Schüler*innenmerkmale die Lehrkräfteeinschätzung im FERMAT beeinflussen und ob es Variationen zwischen den Lehrkräften gibt. Multilevel-Analysen zeigten, dass das Lehrkräfteurteil über die mathematischen Schwierigkeiten ihrer Schüler*innen vorrangig von deren mathematischen Leistung bestimmt wird. Urteilsvariationen wiesen auf interindividuelle Unterschiede in der Urteilsstrenge zwischen den Lehrkräften hin. Die Gewichtung verschiedener Schüler*innenmerkmale variierte jedoch nicht.

Insgesamt bieten die Ergebnisse dieser Dissertation praxis- und forschungsrelevante Implikationen, die die Früherkennung von Grundschulkindern mit einer Rechenstörung im schulischen Kontext unterstützen. Dennoch besteht weiterhin Optimierungsbedarf, um die frühzeitige Risikoidentifikation von Kindern mit einer Rechenstörung zu verbessern.

Abstract

Children with a mathematical learning disorder should be identified at an early stage and supported in their mathematical skills in order to prevent a persistent developmental trajectory and mitigate long-term negative consequences. The primary responsibility for this lies with educational institutions, especially the school and therefore with the teachers. This doctoral dissertation examines *how teachers can effectively identify elementary school children at risk of mathematical learning disorders at an early stage.*

Evidence-based knowledge of the causes, difficulties, and developmental trajectories of mathematical learning disorders is essential for successful risk identification. However, the causes of mathematical learning disorders remain not completely understood, and questions regarding developmental trajectories still exist. *Study I* investigated whether deficits in basic numerical skills in children with mathematical learning disorders are more indicative of a developmental delay or a specific qualitative deviation and if these deficits result from a (selective) impairment of cognitive core systems involved in numerical processing. To address this, data from N = 480 children (n = 68 children with mathematical learning disorders, grades two to four) were analysed. Multilevel analyses did not indicate qualitatively different basic numerical skills but pointed to a specific developmental delay, primarily resulting from a deficit in the approximate number system rather than an access deficit. No evidence was found for impairments in the object tracking system.

Considering this, *Study II* introduced and psychometrically evaluated a screening instrument designed to help teachers in identifying children at risk of mathematical learning disorder. Data were collected from N = 377 students (n = 45 children with test-diagnostic indications of mathematical learning disorders) grades two to four and N = 33 teachers. Psychometric analyses demonstrated that the questionnaire to assess mathematical skills of children in grades two to four (german: Fragebogen zur Erfassung

mathematischer Fertigkeiten [FERMAT]) exhibits good to mostly satisfactory screening characteristics and differentiates between children with and without test-diagnostic indications of mathematical learning disorders in an effective, reliable, and valid manner.

Teacher-based assessments of students' mathematical difficulties are influenced by both relevant and irrelevant information. *Study III* (based on data from Study II) investigated if teachers' assessments of students' mathematical difficulties depend primarily on their actual mathematics-specific performance or whether mathematics-unspecific characteristics also play a role. We also examined whether the strictness of assesements varied across teachers and whether, and to what extent, the consideration of student characteristics depended on the teacher. Multilevel analyses indicated that teachers' assessments of students' mathematical difficulties were primarily based on their actual mathematical performance. Variations in teachers' assessments suggested inter-individual differences in strictness, but the weighting of different student characteristics remained consistent.

Overall, the findings of this doctoral dissertation provide valuable insights for the early identification of elementary school children at risk of mathematical learning disorders. Nevertheless, there is still scope for optimisation in order to improve the early risk identification of children with mathematical learning disorders.

Promotionsrelevante Studien

Die drei Teilstudien der vorliegenden kumulativen Dissertation mit dem Titel – *Rechenstörungen im Grundschulalter: Entwicklungsdefizite und -verzögerungen, Risikoerkennung und Einflussfaktoren auf die diagnostische Einschätzung von Lehrkräften* – wurden in drei verschiedenen Fachzeitschriften mit einem Peer-Review-Verfahren veröffentlicht. Eine zusammenfassende Übersicht der drei Studien ist in Tabelle 1 (S. 81) zu finden.

Studie I (Anhang A): **Lamb, S**., Krieger, F., & Kuhn, J.-T. (2024a). Delayed development of basic numerical skills in children with developmental dyscalculia. *Frontiers in Psychology, 14*, 1187785. https://doi.org/10.3389/fpsyg.2023.1187785

Studie II (Anhang B): **Lamb, S**., Schulz, A.-K., & Kuhn, J.-T. (2024b). Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule. *Lernen und Lernstörungen, 13*(4), 165–177. https://doi.org/10.1024/2235-0977/a000456

Studie III (Anhang C): **Lamb, S**., Schulz, A.-K., & Kuhn, J.-T. (2025). Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule. *Empirische Sonderpädagogik, 17*(1), 35–49. https://doi.org/10.25656/01:34364

Abbildungs- und Tabellenverzeichnis

1 Einleitung

Ein grundlegendes Zahlenverständnis und adäquate Rechenfertigkeiten sind essenziell für die Bewältigung der Komplexität des Lebens (Gerardi et al., 2013). Dies gilt für alltägliche Aufgaben wie dem Ablesen der Uhrzeit und dem Verstehen von Statistiken (Fischer et al., 2013), aber auch für die schulische und berufliche Entwicklung (Claessens & Engel, 2013; Duncan et al., 2007; Hakkarainen et al., 2013; Korhonen et al., 2014).

Mathematikbezogene Fertigkeiten entwickeln sich nicht erst im Schulalter. So verfügen Menschen beispielsweise von Geburt an über ein Gefühl für Zahlen (z. B. Dehaene, 1992). Aufbauend auf diesem angeborenen *Zahlensinn* entwickeln sich schon in der frühen Kindheit mathematische Vorläuferfertigkeiten (z. B. Dehaene, 1992; Krajewski, 2008; vgl. Kapitel 2.2). Im Vorschulalter differenziert sich das Fähigkeitsrepertoire weiter aus und reicht zu diesem Zeitpunkt bereits von der Mengenunterscheidung über die Zahlwortreihe bis hin zur basalen Arithmetik (z. B. Fischer et al., 2017; Fritz et al., 2018). Die Entwicklung mathematischer Kompetenzen setzt sich unter formeller Beschulung meist erfolgreich fort (von Aster & Shalev, 2007; Geary, 2013; Krajewski, 2008).

Doch etwa ein Viertel der Schüler*innen in Deutschland weist am Ende der Grundschulzeit unterdurchschnittlich ausgeprägte mathematische Kompetenzen auf (Selter et al., 2020). Ungefähr 2 bis 8 % entwickeln eine klinisch relevante Rechenstörung (Fischbach et al., 2013; Landerl & Moll, 2010; Moll et al., 2014; Wyschkon et al., 2009; vgl. Kapitel 2.1). Gemäß klinisch-diagnostischer Klassifikationssysteme wird die Rechenstörung als eine spezifische Lernstörung klassifiziert, die durch multiple Beeinträchtigungen beim Erwerb mathematischer Fertigkeiten gekennzeichnet ist (Diagnostic and Statistical Manual of Mental Disorders [DSM-5 / DSM-5-TR], American Psychiatric Association [APA], 2013, 2022; International Classification of Diseases [ICD-10/-11], World Health Organization [WHO], 2025).

Werden Rechenstörungen nicht frühzeitig erkannt und entsprechende Interventionen eingeleitet, steigt das Risiko für einen persistierenden Entwicklungsverlauf mit erheblichen Einschränkungen für die psychische Gesundheit und gesellschaftliche Teilhabe (Esser, 1992; Gerardi et al., 2013; Kohn et al., 2013; Parsons & Bynner, 2005; Saga et al., 2022; Schulz et al., 2018; Vigna et al., 2022). Um die langfristigen Folgen möglichst abzumildern, sollten Kinder mit Anzeichen für eine Rechenstörung frühzeitig identifiziert und in ihren mathematikspezifischen Fertigkeiten gefördert werden, noch bevor sich anfängliche Schwierigkeiten in einer negativen Abwärtsspirale manifestieren (Betz & Breuninger, 1982). Der erste Schritt auf dem Weg zur geeigneten Intervention besteht in der frühzeitigen Risikoidentifikation (Kaufmann & Wessolowski, 2021; Sikora & Voß, 2018; Tröster, 2009; Voß, 2017; vgl. Kapitel 3).

Die Grundschulzeit stellt ein wichtiges Zeitfenster für die Identifikation von Kindern mit Anzeichen für eine Rechenstörung dar (Bender et al., 2024; Sousa et al., 2017). Die Feststellung von Lernschwierigkeiten fällt in den Verantwortungsbereich der Lehrkräfte (z. B. Hesse & Latzko, 2017). Obwohl die diagnostischen Einschätzungen von Lehrkräften generell präzise sind (z. B. Hoge & Coladarci, 1989; Kaufmann, 2020; Mack et al., 2023; Südkamp et al., 2012), gelingt ihnen die Beurteilung leistungsschwacher Schüler*innen allgemein weniger gut (z. B. Coladarci, 1986; Lorenz, 2011; Wagner, 2024). Dies wirft die Frage auf, *wie Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren können?*

Diese Frage bildet den Ausgangspunkt der vorliegenden Dissertation. Eine wesentliche Voraussetzung für eine gelingende Risikoidentifikation ist das evidenzbasierte Wissen über die Ursachen und Schwierigkeiten der Kinder mit Rechenstörung sowie deren Entwicklungsverlauf (z. B. Kaufmann & Wessolowski, 2021). Es besteht wissenschaftlicher Konsens darüber, dass die basisnumerischen Fertigkeiten von Kindern mit einer Rechenstörung beeinträchtigt sind (z. B. Butterworth

et al., 2011; Kißler et al., 2021; Kuhn et al., 2013; Schwenk et al., 2017). Ungeklärt ist jedoch, ob diese Beeinträchtigungen einen qualitativ unterschiedlichen oder verzögerten Entwicklungsverlauf darstellen (vgl. Kapitel 2.2). Die meisten Studien sprechen für eine entwicklungsverzögerte numerische Kognition (z. B. Schwenk et al., 2017; vgl. Kapitel 2.2). Einzelne Befunde deuten jedoch auf überproportional starke Beeinträchtigungen in einzelnen Bereichen hin, was eine abweichende numerische Verarbeitung nahelegt (z. B. Landerl, 2013; vgl. Kapitel 2.2). Eng verbunden mit der Charakterisierung der basisnumerischen Defizite ist die nicht abschließend beantwortete Frage nach deren Ursachen.

Domänenspezifische Erklärungsansätze führen Rechenstörungen auf Beeinträchtigungen in den zentralen Kernmechanismen der Zahlenverarbeitung zurück (für eine Übersicht s. Andersson & Östergren, 2012). Die Erklärungsansätze unterscheiden sich in der Annahme darüber, welcher Kernmechanismus beeinträchtigt ist. Diskutiert werden 1) ein generelles Defizit in der approximativen Mengenverarbeitung (*Approximate Number System* [ANS]), 2) ein spezifisches Defizit in der exakten Repräsentation kleinerer Mengen (*Object Tracking System*, [OTS]), und 3) eine Beeinträchtigung in der Verarbeitung abstrakt dargestellter Mengen in Form von Symbolen (*Access Deficit*, [AD]). Welcher Kernmechanismus beeinträchtigt ist, bleibt aufgrund der heterogenen Evidenzlage unklar (z. B. Andersson & Östergren, 2012; Dowker, 2024; Rousselle & Noël, 2007; vgl. Kapitel 2.2). Daher untersuchte *Studie I* (Lamb et al., 2024a), ob die Defizite in den basisnumerischen Fertigkeiten von Grundschulkindern mit einer Rechenstörung eher auf eine Entwicklungsverzögerung oder auf eine rechenstörungsspezifische qualitative Abweichung hindeuten und ob diese Defizite auf eine (selektive) Beeinträchtigung in einem der zentralen Kernsysteme der Zahlenverarbeitung zurückzuführen sind (vgl. Kapitel 2.3).

Evidenzbasiertes Wissen über Ursachen, Defizite und Entwicklungsverläufe von Rechenstörungen ist eine notwendige, aber keine hinreichende Bedingung dafür, dass Lehrkräfte erkennen, bei welchen Kindern ein Risiko für eine Rechenstörung besteht. Diagnostische Verfahren können Lehrkräfte dabei unterstützen (Sikora & Voß, 2018; vgl. Kapitel 3). Es existieren verschiedene standardisierte Testverfahren (vgl. Kapitel 3.1), die allgemein anerkannte Gütekriterien (z. B. *Reliabilität, Objektivität, Validität*; American Educational Research Association [AERA], American Psychological Association [APA] & National Council for Measurement in Education [NCME], 2014) erfüllen und eine objektiv-reliable Feststellung ermöglichen (für eine Übersicht s. Leitlinie zur Diagnostik und Intervention bei Rechenstörung [S3-Leitlinie] der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften [AWMF]; Schulte-Körne & Haberstroh, 2018). Allerdings werden sekundäre Gütekriterien – *Ökonomie* und *Akzeptanz* – die für die Beurteilung der Nützlichkeit im schulischen Kontext besonders relevant sind, teils eher nachrangig berücksichtigt (Walter, 2020). Damit ist der Einsatz dieser Instrumente im schulischen Kontext mit Einschränkungen verbunden. Denn eine umfangreiche und zeitintensive diagnostische Testung schränkt die tatsächliche Anwendung in der Unterrichtspraxis ein (Beauducel & Leue, 2014).

Indirekte Beurteilungsverfahren, beispielsweise Einschätzungen durch Lehrkräfte, die über einen Fragebogen erfasst werden, können eine zeitsparende Alternative darstellen (Kenny & Chekaluk, 1993). Denn Fragebögen sind allgemein relativ einfach zu handhaben (Reid et al., 2014) und erfordern keine Testung des Kindes (Cabell et al., 2009). Im deutschsprachigen Raum wurden zahlreiche Fragebögen und Symptom-Checklisten veröffentlicht, die darauf abzielen, Kinder mit Anzeichen einer Rechenstörung zu erkennen (z. B. Checkliste für Lehrkräfte; Jacobs & Petermann, 2012; vgl. Kapitel 3.1). Allerdings handelt es sich dabei nicht um psychometrisch validierte Instrumente. So fehlen Angaben zu den psychometrischen Eigenschaften und

Gütekriterien. Somit kann die Güte dieser Instrumente und dessen Eignung für den Einsatz in der Praxis nicht beurteilt werden (vgl. Kapitel 3.1). Daraus ergibt sich die Notwendigkeit, alternative Screeninginstrumente für Lehrkräfte zu entwickeln, die sie bei der Identifikation von Grundschulkindern mit Anzeichen einer Rechenstörung ökonomisch und reliabel unterstützen. Daher wurde in *Studie II* (Lamb et al., 2024b) ein niedrigschwellig einsetzbarer Screeningfragebogen für Lehrkräfte zur Früherkennung von Rechenstörungen in der Grundschule (Fragebogen zur Erfassung mathematischer Fertigkeiten [FERMAT]) vorgestellt und hinsichtlich seiner psychometrischen Eigenschaften und screeningrelevanten Gütekriterien untersucht (vgl. Kapitel 3.2).

Wenn Lehrkräfte einschätzen sollen, ob die mathematischen Fertigkeiten ihrer Schüler*innen beeinträchtigt sind, spielt ihre diagnostische Kompetenz eine zentrale Rolle (vgl. Kapitel 4.1). Meta-Analysen zeigen, dass die Einschätzungen der Lehrkräfte über die Fähigkeiten ihrer Schüler*innen im Allgemeinen recht präzise sind (Kaufmann, 2020; Südkamp et al., 2012). Allerdings gibt es zahlreiche Faktoren, die das diagnostische Urteil der Lehrkräfte beeinflussen können (z. B. Hoge & Coladarci, 1989; Kaufmann, 2020; Südkamp et al., 2012; Wagner, 2024; vgl. Kapitel 4.2). So tendieren Lehrkräfte beispielsweise im Sinne des *logischen Fehlschlusses* dazu, die Ausprägung eines Schüler*innenmerkmals (z. B. Leseleistung) als Indiz für ein anderes Merkmal (z. B. Mathematikleistung) heranzuziehen, ohne dass dies empirisch begründet wäre (z. B. Helmke, 2017). Die Einschätzungen der Lehrkräfte bestehen also aus einer Kombination von urteilsrelevanten und -irrelevanten Informationen, die nicht unbedingt in direktem Zusammenhang mit der Leistung der Schüler*innen stehen (z. B. Kaiser et al., 2015; Südkamp et al., 2012). Hinzu kommen Schwankungen in der Urteilsgenauigkeit und -strenge zwischen den Lehrkräften (z. B. Hesse & Latzko, 2017; Karing & Artelt, 2014). Daher untersuchte *Studie III* (Lamb et al., 2025), ob das Lehrkräfteurteil über die mathematikbezogenen Schwierigkeiten ihrer Schüler*innen im FERMAT auf deren

tatsächlichen mathematikspezifischen Leistungen basiert, oder ob auch mathematikunspezifische Merkmale der Schüler*innen (Geschlecht, Lesefertigkeit und Intelligenz) in das Urteil der Lehrkräfte einfließen. Zudem wurde erforscht, ob die Urteilsstrenge beziehungsweise -milde zwischen den Lehrkräften variiert und ob das Ausmaß, in dem die Merkmale der Schüler*innen herangezogen werden, lehrkräfteabhängig ist (vgl. Kapitel 4.3).

Zur Beantwortung der Ausgangsfrage – *Wie können Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren?* – werden die drei Teilstudien in dieser Dissertation inhaltlich in Beziehung gesetzt, in einen größeren theoretischen Rahmen eingeordnet und anschließend vor dem Hintergrund der übergeordneten Fragestellung diskutiert. Dazu wird in Kapitel 2.1 zunächst das klinisch-diagnostische Verständnis von Rechenstörung durch die Darstellung unterschiedlicher Definitionen und diagnostischer Kriterien erörtert. Anschließend werden in Kapitel 2.2 domänenspezifische Ursachen der Rechenstörung sowie Schwierigkeiten in der Basisnumerik dargestellt, um ein vertieftes Verständnis der zugrunde liegenden Ursachen und Symptomatik zu vermitteln. Das Kapitel mündet in der Skizzierung offener Forschungsfragen, die in *Studie I* (Lamb et al., 2024a) beantwortet werden (Kapitel 2.3; s. Abbildung 1). In Kapitel 2.4 werden die Rechenfertigkeiten in den Blick genommen, da diese neben den Defiziten in der Basisnumerik für die Identifikation von Grundschulkindern mit Anzeichen für eine Rechenstörung ebenfalls zentral sind.

Kapitel 3 ist der Früherkennung von Rechenstörungen im schulischen Kontext gewidmet. In Kapitel 3.1 werden die Potenziale und Herausforderungen bestehender diagnostischer Verfahren zur Identifikation von Kindern mit Anzeichen für eine Rechenstörung im schulischen Setting dargestellt und hinsichtlich ihrer Einsatzmöglichkeiten und Gütekriterien diskutiert. Darauf aufbauend werden Anforderungen an ein Instrument formuliert, das der Identifikation von Kindern mit

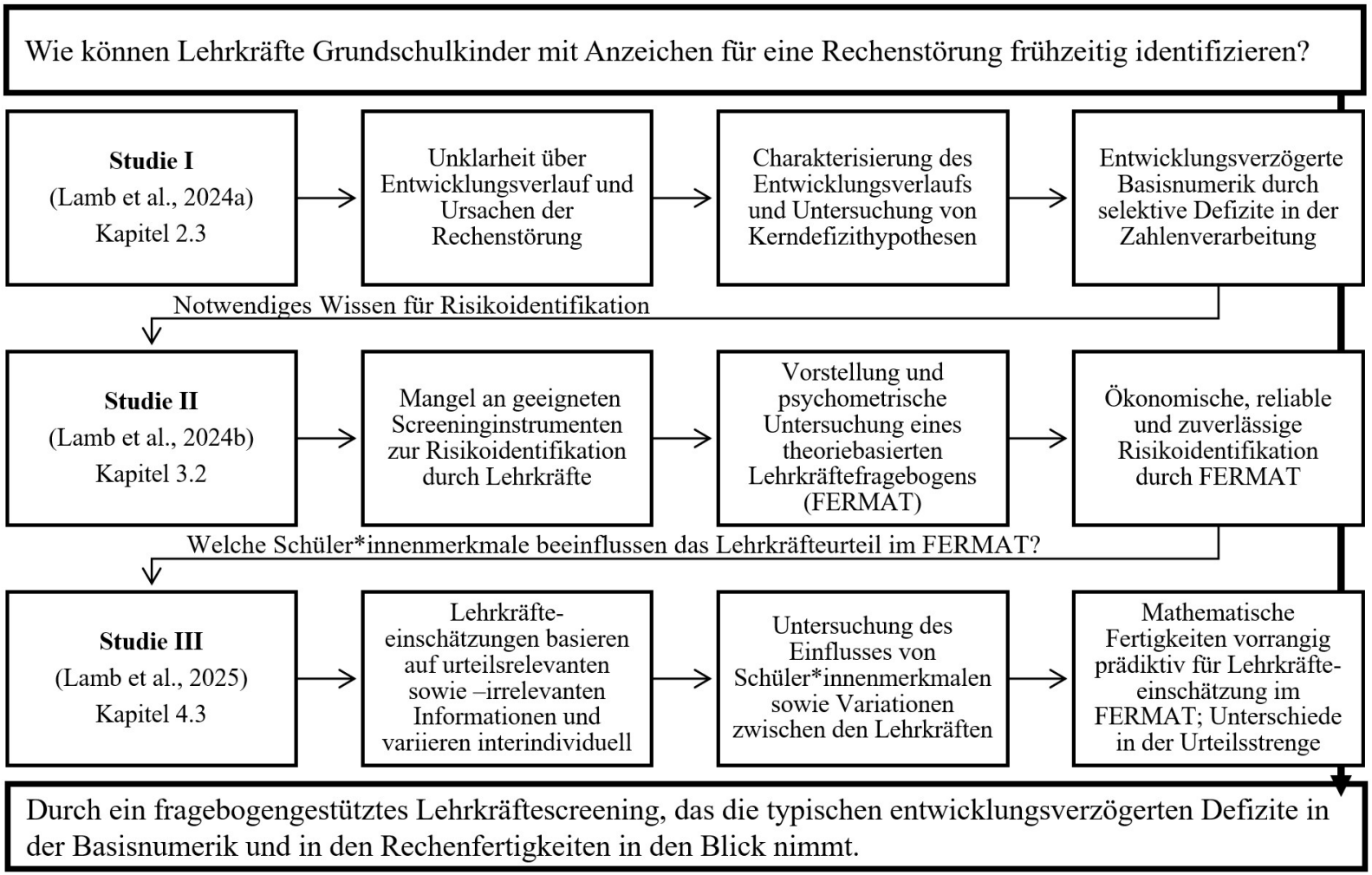
Rechenstörungen durch Lehrkräfte dienen soll. Anschließend wird in *Studie II* (Lamb et al., 2024b) ein Lehrkräftefragebogen vorgestellt, empirisch untersucht und anhand gängiger Gütekriterien bewertet (Kapitel 3.2; s. Abbildung 1).

Daran anknüpfend wird in Kapitel 4 und 4.1 das Konstrukt der Diagnosekompetenz von Lehrkräften operationalisiert. In Kapitel 4.2 werden verschiedene Einflussfaktoren, die die Einschätzungen der Lehrkräfte bei der Beurteilung mathematischer Schwierigkeiten beeinflussen können, beleuchtet und anschließend in *Studie III* (Lamb et al., 2025) untersucht (Kapitel 4.3; s. Abbildung 1).

Kapitel 5 führt die zentralen Erkenntnisse der drei promotionsrelevanten Studien zusammen und setzt sie mit Blick auf die Ausgangsfrage in einen übergeordneten Kontext, um die wesentlichen Ergebnisse und Schlüsse zu diskutieren (Kapitel 5.1 bis 5.3). Es werden Implikationen für die pädagogische Praxis und Forschung bereitgestellt, um die Identifikation und den Umgang mit Rechenstörungen im schulischen Kontext weiter zu optimieren. Die Arbeit schließt mit einem praxisorientierten Fazit und Ausblick (Kapitel 6).

Abbildung 1

Gesamtübersicht über die drei Teilstudien des Dissertationsprojektes



2 Rechenstörung

Wenn von Rechenstörungen gesprochen wird, heißt dies nicht, dass vom gleichen Konstrukt die Rede ist. Je nach Disziplin werden variierende Begriffe verwendet und unterschiedliche Perspektiven eingenommen. Der vorliegenden Dissertation liegt ein klinisch-diagnostisches Verständnis zugrunde. Dieses wird im nachfolgenden Kapitel konkretisiert.

2.1 Klinisch-diagnostisches Verständnis

Trotz variierender Begrifflichkeiten (z. B. Rechenschwäche, Rechenstörung, Dyskalkulie etc.; Kuhn, 2017) und konfligierender Diagnosekriterien (z. B. ICD-10, DSM-5, S3-Leitlinie), besteht aus klinisch-diagnostischer Sicht weitgehender Konsens darin, dass es sich bei der Rechenstörung um eine neurokognitive Störung handelt. Diese ist durch erwartungswidrige Minderleistungen in den mathematischen Fertigkeiten gekennzeichnet. Die mathematischen Schwierigkeiten dürfen nicht auf visuelle, auditive, psychische oder neurologische Störungen, eine unzureichende Schulbildung, eine nicht Beherrschung der Unterrichtssprache oder eine Intelligenzminderung (Intelligenzquotient, IQ $\leq$ 70) zurückgeführt werden (APA, 2013, 2022; WHO, 1992, 1993, 2025).

Weiter wird die *spezifische Lernstörung mit Beeinträchtigung in Mathematik* (315.1), wie sie im DSM-5/-TR bezeichnet wird, durch ein eingeschränktes Zahlenverständnis, Beeinträchtigung der Merkfähigkeit mathematischer Fakten sowie ungenaues Rechnen und mathematisch schlussfolgerndes Denken charakterisiert (APA, 2013, 2022).

Die ICD-10 der WHO (1992) definiert die *Rechenstörung* (F81.2) als *umschriebene Entwicklungsstörung schulischer Fertigkeiten* (F81), die durch eine Beeinträchtigung arithmetischer Fertigkeiten (z. B. Addition, Subtraktion, Multiplikation und Division) gekennzeichnet ist. 2022 ist in Deutschland die neue Auflage ICD-11 (WHO, 2025) in

Kraft getreten. In dieser Version wird die *Entwicklungsbedingte Lernstörung* (6A03) *mit Beeinträchtigung in Mathematik* (6A03.2) durch „bedeutsame und anhaltende Schwierigkeiten beim Erlernen akademischer Fähigkeiten im Zusammenhang mit Mathematik oder Arithmetik, wie z. B. Zahlensinn, Auswendiglernen von Zahlenfakten, genaues Rechnen, flüssiges Rechnen und genaues mathematisches Denken“ charakterisiert (WHO, 2025). Dass die ICD-11 im Vergleich zur ICD-10 nun auch den Zahlensinn und das numerische Denken berücksichtigt, ist vor dem Hintergrund empirischer Forschungsergebnisse eine relevante Ergänzung. Denn die Forschung zeigt, dass Beeinträchtigungen in den basisnumerischen Fertigkeiten charakteristisch für Rechenstörung sind (z. B. Butterworth et al., 2011; Gaupp et al., 2004; Kuhn et al., 2013; Schwenk et al., 2017; vgl. Kapitel 2.2). Dies gilt auch für erwachsene Personen mit einer Rechenstörung (z. B. Bulthé et al., 2019; Gliksman & Henik, 2019; de Visscher et al., 2018).

Operationalisiert wird die Beeinträchtigung im Rechnen im DSM-5/-TR (APA, 2013, 2022) anhand der signifikanten Abweichung der Leistung vom Altersdurchschnitt (*einfaches Diskrepanzkriterium*). Zur Quantifizierung der Abweichung wird eine Diskrepanz von 1.5 Standardabweichungen (*SD*) empfohlen (APA, 2013). Zusätzlich wird zwischen drei Schweregraden (leicht, moderat und schwer) unterschieden, die ein unterschiedliches Ausmaß an Unterstützung erfordern (APA, 2013; Falkai & Wittchen, 2015).

Im Vergleich dazu differenziert die ICD-10/-11 (WHO, 1992, 2025) nicht zwischen verschiedenen Schweregraden, sondern fordert zusätzlich zum einfachen Diskrepanzkriterium die Berücksichtigung der Diskrepanz zwischen der individuellen Intelligenz und der damit erwarteten Leistung (*doppeltes Diskrepanzkriterium*). Das bedeutet, dass die Leistung des Kindes nicht nur, wie im DSM-5 gefordert, vom Altersdurchschnitt abweichen, sondern auch unterhalb der Leistung liegen muss, die

aufgrund seiner Intelligenz zu erwarten wäre. Je nachdem, wie stark die IQ-Diskrepanz (gemessen in *SD*) ausfällt, kann in der ICD-10/-11 zwischen *Lernschwächen* (IQ-Diskrepanz < 2 *SD*) und *Lernstörungen* (IQ-Diskrepanz ≥ 2 *SD*) unterschieden werden (Thomas et al., 2015).

Diese Unterscheidung hat in Deutschland Konsequenzen für die staatliche Finanzierung außerschulischer Fördermaßnahmen. Diese sind an das doppelte Diskrepanzkriterium des ICD-10/-11 gekoppelt. Kinder, bei denen die Diagnose einer Rechenstörung (IQ-Diskrepanz ≥ 2 *SD*) nicht gestellt wird, haben im Zweifelsfall keinen Zugang zum außerschulischen Unterstützungssystem. Zudem werden die Kinder innerschulisch benachteiligt, da die Gewährung von Nachteilsausgleichen in der Regel an die ICD-10/-11-Diagnose einer Rechenstörung gekoppelt ist (Mähler, 2021).

Die diagnostische und pädagogische Sinnhaftigkeit des doppelten Diskrepanzkriteriums wird unlängst infrage gestellt (Mähler, 2021; Schulte-Körne, 2021). Grund dafür sind unter anderem Studien, die zeigen, dass Kinder mit und ohne Erfüllung der IQ-Diskrepanz gleichermaßen von mathematikspezifischen Interventionen profitieren (z. B. Chodura et al., 2015) und sich nicht substanziell in ihren Beeinträchtigungen unterscheiden (z. B. Kuhn et al., 2013). Kritisch ist zudem, dass ein niedrigerer IQ tendenziell dazu führt, dass die IQ-Diskrepanz nicht erfüllt wird, da die mathematische Leistung in diesem Fall extrem niedrig ausfallen muss, um die erforderliche Diskrepanz zu erfüllen. So zeigt beispielsweise die Studie von Fischbach et al. (2013), dass intelligentere Kinder mit mathematischen Minderleistungen eher die IQ-Diskrepanz erreichen und somit das Diagnosekriterium einer Rechenstörung erfüllen. In der Konsequenz kann dies dazu führen, dass Kinder mit höherer Intelligenz eine höhere Chance auf außerschulische Förderungen haben.

Ähnlich zum DSM-5/-TR (APA, 2013, 2022) wird in der S3-Leitlinie der AWMF (Schulte-Körne & Haberstroh, 2018) vom doppelten Diskrepanzkriterium Abstand

genommen. Gemäß der S3-Leitlinie liegt eine Rechenstörung vor, wenn die mathematische Leistung in einem standardisierten Test trotz ausreichender Schulbildung und ohne Intelligenzminderung einem Prozentrang (PR) ≤ 7 entspricht. Liegen zusätzlich qualitative (z. B. durch Eltern und Lehrkräfte) und klinische (z. B. unbeeinträchtigte sensorische Funktionen) Informationen vor, die den Verdacht auf eine Rechenstörung unterstützen, kann ein weniger strenger Cut-Off von 1 *SD* respektive ein PR von ≤ 16 angewandt werden (Schulte-Körne & Haberstroh, 2018).

Zu den in der Praxis variierenden diagnostischen Cut-Off-Kriterien zur Operationalisierung der Rechenstörung kommen weitere Kriterien aus der Forschungspraxis hinzu (Dowker, 2024; Mammarella et al., 2021). Während der Gebrauch standardisierter Testinstrumente sowohl in der Forschung als auch in der Praxis üblich ist, erfüllen Forschungsstichproben, bedingt durch die forschungslegitime Zielsetzung eine möglichst große Stichprobe zu generieren, teils weniger strenge Einschlusskriterien. So wird, um beispielsweise die erforderliche Power für statistische Verfahren zu erreichen, in einigen Studien eine geringere Diskrepanz vom Altersdurchschnitt in Kauf genommen (PR ≤ 35; Jordan et al., 2003 oder PR < 35; Geary et al., 2000). Zudem werden aus forschungsökonomischen Gründen nicht selten differenzialdiagnostische Kriterien (z. B. die Überprüfung der Lese- und Rechtschreibleistung) vernachlässigt (Hasselhorn & Schuchardt, 2006). Dies führt dazu, dass je nach Studie unterschiedliche Subtypen von Rechenstörungen betrachtet werden. In der Konsequenz schränkt dies die Vergleichbarkeit der Ergebnisse ein.

In dem Gebrauch variierender Cut-Off-Kriterien zur Klassifikation der Minderleistungen im Rechnen spiegeln sich außerdem verschiedene Sichtweisen auf Rechenstörungen wider, die mit unterschiedlichen Perspektiven auf die Ursachen verwoben sind. Dies lässt sich anhand der von Geary (2013) vorgeschlagenen Zwei-Gruppen Lösung darstellen.

Kinder, deren mathematische Leistung in einem standardisierten Test in mindestens zwei aufeinanderfolgenden Schuljahren einem PR von ≤ 10 entspricht, zählt er zur stark beeinträchtigten Gruppe *Mathematical learning disability*. Der Begriff *Behinderung* deutet auf eine biologisch bedingte Abweichung hin; eine mathematikspezifische Lernbehinderung impliziert darüber hinaus eine Störung, die durch spezifische kognitive Defizite gekennzeichnet ist (Mazzocco, 2007, S. 29). Aus dieser Perspektive sind die Begriffe *mathematische Behinderung* und *mathematische Lernbehinderung* Variationen desselben Konstrukts. In ähnlicher Weise werden die Begriffe *mathematische Lernbehinderung* und *Dyskalkulie* typischerweise verwendet, um sich auf eine Zielgruppe zu beziehen, deren Schwierigkeiten eine angeborene Beeinträchtigung implizieren und nicht primär durch Umweltfaktoren verursacht werden (Mazzocco, 2007, S. 29). Damit umfasst ein strengerer Cut-Off-Wert (z. B. PR ≤ 10) vorrangig Kinder, deren Schwierigkeiten im Rechnen mit kognitiven Defiziten verbunden sind (Mazzocco, 2007; Mazzocco et al., 2011). In der Forschungsliteratur werden Kinder meistens als *dyskalkulisch* bezeichnet, wenn sie in etwa die klinischen Kriterien des ICD-10/-11 oder DSM-5 erfüllen. Dabei stehen domänenspezifische und -generelle kognitive Funktionen im Fokus der Ursachenhypothesen.

Kinder, die in zwei aufeinanderfolgenden Schuljahren einen 10 < PR < 25 erreichen, werden von Geary (2013) als *Low Achievers* bezeichnet. Diese Kinder werden in der Literatur meist unter dem weiter gefassten Begriff *Rechenschwierigkeiten* subsumiert. Mazzocco et al. (2011) argumentieren jedoch, dass eine größer angelegte Range (z. B. PR ≤ 25) Kinder mit Rechenschwierigkeiten identifiziert, deren Ursachen in einer breiteren ätiologischen Basis liegen. Liberalere Cut-Off-Werte schließen nach diesem Verständnis auch Kinder ein, bei denen unter anderem sekundäre Einflussfaktoren eine größere Bedeutung für die Entstehung von Rechenschwierigkeiten haben.

Empirisch gestützt wird diese Zwei-Gruppen-Kategorisierung durch Studien, die zeigen, dass sich Kinder mit schwerwiegenden Defiziten (PR ≤ 10) von der Gruppe mit weniger schweren Minderleistung (10 < PR < 25) unterscheiden (Geary et al., 2007; Geary et al., 2012; Mazzocco et al., 2011; Murphy et al., 2007). Die vorliegende Dissertation orientiert sich an der Argumentation von Mazzocco et al. (2011). Der Begriff Rechenstörungen wird bevorzugt verwendet. Operationalisiert wird diese über die signifikante Abweichung in einem standardisierten Leistungstest vom Altersdurchschnitt (mathematische Fertigkeiten: T-Wert ≤ 38 oder PR ≤ 10; PR ≤ 7).

Dass Rechenstörungen unterschiedlich operationalisiert werden, führt dazu, dass sich Forschungsstichproben von klinischen Stichproben unterscheiden. Darüber hinaus existieren auch innerhalb der Forschung Unterschiede. Diese Variabilität spiegelt sich ebenso in variierenden Prävalenzangaben wider.

Prävalenz

Je nachdem, welchen Cut-Off-Wert Studien zur Operationalisierung von Rechenstörungen heranziehen, unterscheiden sich die Prävalenzangaben. Dies erklärt, unter anderem weshalb die Prävalenzen zwischen 2 und 8 % variieren (Fischbach et al., 2013; Haberstroh & Schulte-Körne, 2019; Hasselhorn & Schuchardt, 2006; Landerl & Moll, 2010; Moll et al., 2014; Shalev et al., 2000; Wyschkon et al., 2009). Eine Studie von Moll et al. (2014) zeigte, dass je nach Strenge der Kriterien (Abweichung vom Altersdurchschnitt gemessen in *SD*) zwischen 4.84 % (1 *SD*), 3.31 % (1.25 *SD*) und 2.39 % (1.5 *SD*) der Grundschulkinder von einer Rechenstörung betroffen sind. Eine ähnliche Schwankungsbreite wurde in einer Studie von Fischbach et al. (2013) festgestellt, wobei die Prävalenzraten bei Anwendung der doppelten Diskrepanz (gemäß ICD-10) bei 2.6 % und bei Anwendung des einfachen Diskrepanzkriteriums bei 5.0 % lagen. Wyschkon et al. (2009) berichten von einer noch größeren Varianz in Abhängigkeit der Kriterien. Diese

lag zwischen 0.1 % (Forschungskriterien der ICD-10 mit Bezug zur Gesamtintelligenz) und 8.1 % (einfache Diskrepanz).

Die unterschiedlichen Cut-Off-Kriterien sind nicht der einzige Grund für die variierenden Prävalenzangaben (z. B. Desoete et al., 2004; Hasselhorn & Schuchardt, 2006). Auch die Berücksichtigung von Ein- und Ausschlusskriterien (z. B. Fischbach et al., 2013), die Verwendung diagnostischer Verfahren mit unterschiedlichen Schwerpunkten sowie die Betrachtung verschiedener Alters- oder Klassenstufen (z. B. Wyschkon et al., 2009) verursachen Schwankungen in den Prävalenzen.

Fischbach et al. (2013) zeigten beispielsweise, dass die Prävalenz der Rechenstörung von der Kontrolle weiterer Ein- und Ausschlusskriterien (z. B. Lesefertigkeit) beeinflusst wird. Wurde neben der mathematischen Leistung auch die Lese- und Rechtschreibleistung kontrolliert, wiesen 5.0 % der Kinder eine Rechenschwäche (nach ICD-10) und 2.6 % eine Rechenstörung (nach ICD-10) auf. Ohne die Kontrolle der Lese- und Rechtschreibleistung waren es 9.2 % respektive 5.0 %. Auch von Aster et al. (2007) ermittelten eine Prävalenzrate von 6 % für Rechenstörungen, wobei nur 1.8 % der Kinder von einer isolierten Rechenstörung betroffen waren. Die übrigen 4.2 % hatten zusätzlich substanzielle Schwierigkeiten im Lesen und Rechtschreiben, was darauf hindeutet, dass ein beträchtlicher Anteil der Forschungsbefunde nicht das Profil einer isolierten Rechenstörung, sondern vielmehr eine kombinierte Form schulischer Lernschwierigkeiten/-störungen beschreibt, die durch Schwierigkeiten in den drei Kernkompetenzen Lesen, Schreiben und Rechnen gekennzeichnet ist.

Auch Wyschkon et al. (2009) untersuchten die Auswirkungen verschiedener Definitionsansätze zur Klassifikation von Lernstörungen auf die Prävalenzraten. Dazu wendeten sie die Operationalisierungen anderer Studien auf ihre eigene Stichprobe an und verglichen die Ergebnisse mit den Originalstudien. Dabei zeigte sich, dass bei der

Anwendung gleicher Cut-Off-Kriterien dennoch große Prävalenzschwankungen zu beobachten sind, die vermutlich auf die Betrachtung unterschiedlicher Alters- oder Klassenstufen und Testverfahren mit unterschiedlichen inhaltlichen Schwerpunkten zurückzuführen sind.

Ähnlich heterogen sind die Ergebnisse im Hinblick auf Geschlechterunterschiede. Die meisten Studien zeigen, dass Mädchen signifikant häufiger von Rechenschwierigkeiten/-störungen betroffen sind als Jungen (z. B. Fischbach et al., 2013; Landerl & Moll, 2010; Moll et al., 2014). Andere Studien belegen Geschlechterunterschiede zugunsten der Mädchen (z. B. Barbaresi et al., 2005) oder finden keine eindeutigen Unterschiede zwischen den Geschlechtern (z. B. Desoete et al., 2004; Reigosa-Crespo et al., 2012). Daher sollte allen Geschlechtern die gleiche Aufmerksamkeit geschenkt werden (Devine et al., 2013).

Bei der Betrachtung der Geschlechterunterschiede werden neben testdiagnostischen Kriterien auch psychosoziale Einflussfaktoren diskutiert, die die Varianz der Prävalenzen erklären könnten. Dazu zählt unter anderem das *mathematische Selbstkonzept*, welches bei Jungen höher ausgeprägt ist als bei Mädchen (Cvencek et al., 2011; Lindberg et al., 2013; Ehm et al., 2011; Schütky, 2022) und mit der (mathematischen) Leistung in Zusammenhang steht (z. B. Ehm et al., 2011). In einer Meta-Analyse zeigten Wu et al. (2021), dass der Zusammenhang zwischen der Leistung und dem Selbstkonzept reziprok ist, wobei der Effekt von der Leistung auf das Selbstkonzept stärker ausfiel. Zu einem ähnlichen Ergebnis kamen Ehm et al. (2021). Darüber hinaus gibt es Hinweise darauf, dass das schulische Selbstkonzept stärker von *sozialen Erwartungen* als von der tatsächlichen Leistung beeinflusst wird. Krajewski (2008) zeigte, dass Jungen im Vorschulalter und in der ersten Klasse objektiv bessere mathematische Leistungen zeigen als Mädchen, dieser Unterschied verschwindet aber in der zweiten Klasse, während die Unterschiede im mathematischen Selbstkonzept

zugunsten der Jungen stabil bleiben. Krajewski (2008) sieht in diesem Ergebnis einen Beleg dafür, dass die Genese des schulischen Selbstkonzepts stärker von sozialen Erwartungen als von tatsächlichen Leistungen beeinflusst wird (S. 201f.). Auch Steinmayr et al. (2019) stellen fest, dass Geschlechterunterschiede im mathematischen Selbstkonzept zuungunsten der Mädchen nicht durch tatsächliche Leistungen, sondern teilweise durch Schulnoten sowie durch Einschätzungen der Eltern und Lehrkräfte erklärt werden.

Denkbar wäre also, dass *Geschlechterstereotype* das mathematische Selbstkonzept beeinflussen. Stereotype beziehen sich allgemein auf generalisierte Überzeugungen über Menschen verschiedener sozialer Gruppen (Eagly & Chaiken, 1993). Geschlechterstereotype beinhalten Attribute, die Individuen aufgrund ihres Geschlechts zugeschrieben werden (Eagly, 1987). Obwohl die klinische Rechenstörung nicht auf Geschlechterstereotype zurückzuführen ist, zeigen Studien, dass Stereotype die mathematische Leistung beeinflussen können. Es gibt Hinweise darauf, dass Grundschullehrkräfte bei objektiv gleichen mathematischen Leistungen Schülerinnen schlechter bewerten als Schüler (Cimpian et al., 2016). Somit könnten Mädchen bewusst oder unbewusst durch Geschlechterstereotype beeinflusst werden und schlechtere mathematische Leistungen zeigen, die unter ihrem tatsächlichen Fähigkeitsniveau liegen (z. B. Huguet & Régner, 2007). Unterstützt wird diese Annahme durch eine aktuelle Studie von Olczyk et al. (2023). Die Autor*innen zeigten nicht nur eine systematische Verzerrung der Einschätzungen von Lehrkräften zuungunsten der Mädchen im Unterrichtsfach Mathematik, sondern auch, dass Leistungsunterschiede zwischen Jungen und Mädchen während der Grundschulzeit weiter zunehmen. Dieses Ergebnis sehen die Autor*innen als Beleg für eine sich selbst erfüllende Prophezeiung. Ähnliche Befunde liefern Robinson-Cimpian et al. (2014).

Auch wenn umweltbezogene Einflussfaktoren keine direkte Ursache für eine klinisch relevante Rechenstörung darstellen, verdeutlicht dieses Beispiel, dass Rechenstörungen in Wechselwirkung mit internalen und externalen Faktoren stehen. Dies ist insbesondere für die Risikoidentifikation im Grundschulalter relevant und wird daher in Kapitel 4 erneut aufgegriffen und in *Studie III* (Lamb et al., 2025) untersucht.

Zusammenfassend kann festgehalten werden, dass das klinisch-diagnostische Verständnis von Rechenstörungen kein einheitlich operationalisiertes Konstrukt darstellt. Klinische Stichproben unterscheiden sich von Forschungsstichproben und auch innerhalb der Forschung gibt es Unterschiede zwischen den Gruppen, die unter dem Begriff Rechenstörung subsumiert werden. Diese Unterschiede sind im Wesentlichen auf variierende diagnostische Ein- und Ausschlusskriterien, die Verwendung diagnostischer Verfahren mit verschiedenen inhaltlichen Schwerpunkten sowie die Betrachtung unterschiedlicher Alters- und Klassenstufen zurückzuführen. Dies spiegelt sich einerseits in variierenden Prävalenzangaben, aber auch in unterschiedlichen Ergebnissen über die Ursachen und Schwierigkeiten der Rechenstörung wider, welche im nachfolgenden Kapitel thematisiert werden.

2.2 Ursachen und Symptome

Eng verbunden mit der Charakterisierung der beobachtbaren Schwierigkeiten bei Rechenstörungen ist die Frage nach deren Ursachen. Doch „[d]ie potenziellen Verursachungsfaktoren von Rechenstörungen sind vielschichtig“ (Kaufmann & von Aster, 2012, S. 770) und nach wie vor nicht abschließend geklärt. Weitgehender Konsens besteht darin, dass Rechenstörungen durch mehrere innerhalb (z. B. biologische Faktoren, Kognition) und außerhalb des Individuums liegende Faktoren (z. B. familiäres und schulisches Umfeld) multikausal bedingt sind (z. B. Jacobs & Petermann, 2003). Sowohl das *Multiple-Defizit-Modell* (Butterworth & Kovas, 2013) als auch das *Kausalmodell* von

Butterworth et al. (2011) betrachten die Entstehung von Lernstörungen vor dem Hintergrund eines komplexen Zusammenspiels genetischer, neuronaler, kognitiver und behavioraler Faktoren, die wiederum in Wechselwirkung mit der Umwelt des Individuums stehen (Butterworth & Kovas, 2013; Buttwerwoth et al., 2011). Die postulierten multikausalen Zusammenhänge zwischen den einzelnen Faktoren sind jedoch aufgrund ihrer Komplexität und gegenseitigen Beeinflussung empirisch schwer zu belegen. Dennoch bieten diese Modelle einen konzeptionellen Rahmen, um mögliche Verursachungs- und Risikofaktoren auf verschiedenen Ebenen systematisch zu betrachten.

Die vorliegende Dissertation untersucht Grundschulkinder, deren Minderleistungen im Rechnen nach klinischem Verständnis vorrangig mit kognitiven Defiziten verbunden sind (vgl. Kapitel 2.1). Da mathematikspezifische Beeinträchtigungen die Lernstörung im Rechnen charakterisieren (vgl. Kapitel 2.1), werden in dieser Arbeit vorrangig Defizite in den Kernmechanismen der Zahlenverarbeitung als ursächliche Faktoren diskutiert. Dazu wird im nachfolgenden Kapitel der Forschungsstand über die rechenstörungsspezifischen Schwierigkeiten in der Basisnumerik vor dem Hintergrund verschiedener domänenspezifischer Kerndefizithypothesen entlang unterschiedlicher Aufgabenparadigmen skizziert.

Mit Blick auf die Ausgangsfrage – *Wie können Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren?* – ist es zudem sinnvoll, sich auf Schwierigkeiten zu beschränken, die auf behavioraler Ebene sichtbar sind. Neurologische Aspekte (z. B. Hirnanomalien) werden daher nicht berücksichtigt. Domänengenerelle Defizite treten im Zusammenhang mit Rechenstörungen ebenfalls auf, zählen aber nicht zu den zentralen Leitsymptomen der Lernstörung (vgl. Kapitel 2.1). Daher werden diese nachfolgend komprimiert dargestellt und stehen nicht im Mittelpunkt der vorliegenden Arbeit.

Domänengenerelle Ursachen und Defizite

Domänengenerelle Funktionen beziehen sich auf Prozesse, die nicht immanenter Bestandteil der mathematischen Fertigkeiten, sondern im Allgemeinen für das Lernen und die Informationsverarbeitung wichtig sind (z. B. Kuhn et al., 2019). Einige Autor*innen gehen davon aus, dass Rechenstörungen mit Defiziten in den domänengenerellen Funktionen einhergehen. In diesem Zusammenhang wurden vor allem Beeinträchtigungen im Arbeitsgedächtnis und der zentralen Exekutive diskutiert.

In der europäischen Forschungslandschaft wird das *Arbeitsgedächtnis* häufig auf Grundlage des Mehrkomponenten-Modells von Baddeley (1986) operationalisiert. Baddeley (1986) geht davon aus, dass sich das Arbeitsgedächtnis aus drei Komponenten zusammensetzt: 1) der *zentralen Exekutive*, die für die Überwachung und Regulierung komplexer kognitiver Prozesse zuständig ist, 2) dem *visuell-räumlichen Notizblock* und 3) der *phonologischen Schleife* über die visuell-räumliche beziehungsweise sprachlich-auditive Informationen kurzfristig gespeichert und verarbeitet werden.

Defizite im Arbeitsgedächtnis werden häufig bei Kindern mit Lernstörungen beobachtet, wobei das Ausmaß und die betroffenen Komponenten je nach Lernstörung variieren (z. B. Mähler & Schuchardt, 2016; Peng & Fuchs, 2016; Schuchardt et al., 2008). Rechenstörungen gehen vor allem mit Defiziten im visuell-räumlichen Arbeitsgedächtnis einher (z. B. Busch et al., 2018; Menon, 2016; Mähler & Schuchardt, 2016; Schuchardt et al., 2008).

Die *zentrale Exekutive* ist im Modell von Baddeley nicht abschließend spezifiziert und umfasst verschiedene Kontroll- und Steuerungsfunktionen. Studien, die die zentrale Exekutive untersuchen, rekurrieren häufig auf drei Komponenten, für die gezeigt wurde, dass sich diese faktorenanalytisch voneinander trennen lassen: Shifting, Inhibition und Updating (Miyake et al., 2000). *Shifting* beschreibt die Fähigkeit, flexibel zwischen verschiedenen Anforderungen, Aufgaben oder mentalen Zuständen zu wechseln.

Updating bezeichnet die Fähigkeit, Informationen im Arbeitsgedächtnis zu speichern und diese weiterzuverarbeiten, beispielsweise irrelevante Informationen durch neue zu ersetzen. *Inhibition* meint, die Fähigkeit, impulsive oder dominante Reaktionen zu hemmen (Miyake et al., 2000).

Peng et al. (2018) berichten in einer Meta-Analyse, dass Personen mit einer Rechenstörung Defizite in allen drei Komponenten zeigen. Dies wird auch durch die Meta-Analyse von Haberstroh und Schulte-Körne (2022) unterstützt, wobei Rechenstörungen vor allem mit Beeinträchtigungen in der *Inhibition* assoziiert sind. Allerdings schlossen nur drei der sieben Studien, die Haberstroh und Schulte-Körne (2022) berücksichtigten, komorbide Aufmerksamkeitsdefizit-/Hyperaktivitätsstörungen (AHDS) aus. Da Beeinträchtigungen in der Inhibition zum Leitsymptom der ADHS zählen (z. B. van Hulst et al., 2018) und Rechenstörungen häufig komorbid mit ADHS-Symptomen auftreten (Visser et al., 2020), könnte dieses Ergebnis konfundiert sein. Denn Kuhn et al. (2016) zeigten, dass das Profil kombiniert beeinträchtigter Kinder (Rechenstörung + ADHS) additiv auftritt. Daher wäre es denkbar, dass dies auch für das Inhibitionsdefizit gilt.

Unstrittig ist, dass domänengenerelle kognitive Funktionen in mathematikbezogene Prozesse involviert sind (z. B. Vogel & De Smedt, 2021; Wilkey et al., 2018). Doch *domänenspezifische Prozesse*, die sich auf die numerische Kognition beziehen, scheinen in einem substanziellen Zusammenhang mit der mathematischen Leistung zu stehen, der über die domänengenerellen Fähigkeiten (z. B. Intelligenz) hinaus geht (z. B. Chen & Li, 2014; Dornheim, 2008; Gallit et al., 2017; Schneider et al., 2017; Stern, 2003; Weißhaupt et al., 2006). Studien zeigen, dass gut entwickelte basisnumerische Vorläuferfertigkeiten mit besseren mathematischen Fertigkeiten einhergehen (z. B. Habermann et al., 2020; Hornung et al., 2014; Krajewski & Schneider, 2006; Lyons et al., 2014). Domänengenerelle Faktoren scheinen hingegen eher einen indirekten Einfluss auf den

Rechenerwerb beziehungsweise die mathematischen Fertigkeiten zu haben (Cirino, 2011; Dornheim, 2008; Krajewski & Schneider, 2006; Weißhaupt et al., 2006). Daher ist es sinnvoll, die domänengenerellen Funktionen als unterstützenden Rahmen für die Entwicklung mathematischer Kompetenzen und nicht als vorrangige Kernursache der Rechenstörung zu betrachten, ähnlich zu dem *Vier-Stufen-Modell der Entwicklung zahlenverarbeitender Hirnfunktionen* nach von Aster & Shalev (2007), auf das im folgenden Abschnitt eingegangen wird.

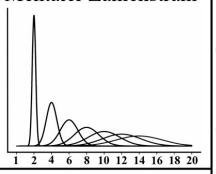
Domänenspezifische Ursachen und Defizite

Zu den elementaren domänenspezifischen kognitiven Kernsystemen der Zahlenverarbeitung zählen das *Object Tracking System* (OTS), das der genauen Unterscheidung respektive dem simultanen Erkennen (englisch: Subitizing) von drei bis vier Objekten dient (Feigenson et al., 2004) und das *Approximate Number System* (ANS), welches für das ungefähre Erkennen und Schätzen von non-symbolischen Mengen (z. B. Punkte) zuständig ist. Das *neuropsychologische Vier-Stufen-Modell der Entwicklung zahlenverarbeitender Hirnfunktionen* (von Aster und Shalev, 2007), das auf dem *Triple-Code-Modell* ([TCM]; Dehaene, 1992; Dehaene & Cohen, 1995) basiert, geht davon aus, dass sich aufbauend auf diesen beiden angeborenen Kernsystemen (s. Abbildung 2, Stufe 1) weitere numerische Repräsentationsformen entwickeln, die über verschiedene Codes repräsentiert und jeweils durch ein auf den Code spezialisiertes Modul verarbeitet werden. Diese kognitiven Repräsentationsformen entwickeln sich in einem aufeinander aufbauenden Prozess. Umrahmt wird diese Entwicklung von domänengenerellen Funktionen (s. Abbildung 2). Beginnend mit der Sprachentwicklung entwickeln sich verbal-phonologische Zahlensymbole (Zahlwörter, z. B. /drei/), die über das *verbal-phonologische Modul* (s. Abbildung 2, Stufe 2) verarbeitet werden. Mit dem Beginn der formellen Beschulung werden visuell-arabische Zahlensymbole (z. B. 3) erworben (von

Aster, 2013; von Aster & Shalev, 2007). Diese werden über das *visuell-arabische Modul* (s. Abbildung 2, Stufe 3) verarbeitet. Die Module werden als autonome Funktionseinheiten betrachtet, die in verschiedenen Regionen des Gehirns lokalisiert und über Transkodierungsprozesse miteinander verbunden sind (Dehaene & Cohen, 1995; s. Abbildung 2). Auf diese Weise können die verschiedenen numerischen Informationen ineinander übersetzt werden. Die Entwicklung setzt sich im weiteren Schulalter fort und mündet in einer abstrakten Zahlenraumvorstellung. Der *mentale Zahlenstrahl* (s. Abbildung 2, Stufe 4) entwickelt sich sukzessive, aufbauend auf der approximativen Mengenvorstellung und der Integration ordinaler und kardinaler Zahlaspekte (von Aster, 2013). In diesem Entwicklungsschritt sind die verbal-phonologischen und visuell-arabischen Zahlenrepräsentationen vollständig auf dem ANS repräsentiert, sodass auch von einem *symbolisch-mentalen Zahlenstrahl* gesprochen werden kann (von Aster, 2013; von Aster & Shalev, 2007). Es wird angenommen, dass die mentale Repräsentation von Zahlen räumlich von links nach rechts auf einer Linie angeordnet sind (Dehaene, 1992; Dolores de Hevia et al., 2006; s. Abbildung 2, Stufe 4). Zudem geht die Forschung davon aus, dass jede Zahl auf dem mentalen Zahlenstrahl als Gauß-Verteilung dargestellt wird, wobei diese Verteilung mit zunehmender Größe der Menge breiter wird (z. B. Vogel & De Smedt, 2021; s. Abbildung 2, Stufe 4).

Abbildung 2

Adaptierte Darstellung des Vier-Stufen-Modells der Entwicklung zahlenverarbeitender Hirnfunktionen nach von Aster und Shalev (2007)

<table>
<tr><td colspan="5">Arbeitsgedächtniskapazität</td><td rowspan="7">Umweltbezogene Einflussfaktoren</td></tr>
<tr><td rowspan="2">Phase</td><td>1. Stufe</td><td>2. Stufe</td><td>3. Stufe</td><td>4. Stufe</td></tr>
<tr><td>Kleinkindalter</td><td>Vorschulalter</td><td colspan="2">Schulalter</td></tr>
<tr><td>Behavioral (Funktionen)</td><td>Simultanerfassung, Approximation</td><td>(Verbales) Zählen, Zählstrategien, Faktenabruf</td><td>Schriftliches Rechnen, gerade/ungerade</td><td>Schätzen, Überschlagen, Vergleichen</td></tr>
<tr><td>Kognitive Repräsentation</td><td>Non-symbolische Mengen-repräsentationen  </td><td>Verbal-phonologische Zahlensymbole
/eins/ /zwei/...</td><td>Visuell-arabische Zahlensymbole
..., 13, 14, 15, ...</td><td>Mentaler Zahlenstrahl  </td></tr>
<tr><td>Kernsysteme der Zahlen-verarbeitung</td><td>Object Tracking System (OTS) und Approximate Number System (ANS)</td><td>Verbal-phonologisches Modul</td><td>Visuell-arabisches Modul</td><td>Verbale und visuelle Zahlenrepräsentation vollständig auf ANS repräsentiert</td></tr>
<tr><td>Lokalisation im Gehirn</td><td>Biparietal</td><td>Links präfrontal</td><td>Biokzipital</td><td>Biparietal</td></tr>
<tr><td colspan="5">Genetische Prädisposition</td></tr>
</table>

Anmerkungen. Eigene adaptierte Darstellung des Vier-Stufen-Modells der Entwicklung zahlenverarbeitender Hirnfunktionen nach von Aster und Shalev (2007, S. 870). Das Modell bringt die Module des Triple Code Modells (Dehaene, 1992) in eine zeitliche Abfolge und ergänzt den mentalen Zahlenstrahl. Umrahmt und beeinflusst wird der Entwicklungsprozess durch die domänengenerelle Arbeitsgedächtniskapazität und weitere umweltbezogene Faktoren (Butterworth et al., 2011). Der mentale Zahlenstrahl in Stufe 4 ist eine eigens modifizierte Darstellung nach Vogel und De Smedt (2021, S. 3).

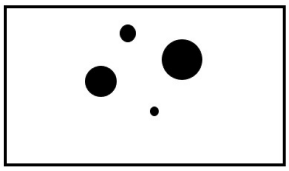
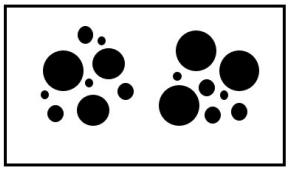
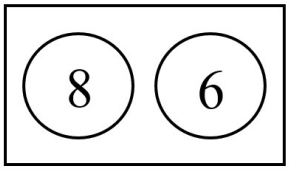
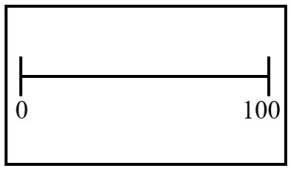
Dem Modell folgend bilden also frühe, basale, teils angeborene basisnumerische Fertigkeiten den Ausgangspunkt der mathematischen Kompetenzentwicklung (s. Abbildung 2, Stufe 1). Folglich lässt sich die These ableiten, dass sich Defizite in den frühen Entwicklungsstadien nachteilig auf die weitere mathematische Kompetenzentwicklung auswirken (z. B. Braeuning et al., 2021). Daher postuliert die Forschung, dass Rechenstörungen auf Defizite in den zentralen Kernmechanismen der Zahlenverarbeitung zurückzuführen sind. Empirisch zeigt sich dies darin, dass Kinder, Jugendliche sowie Erwachsene mit Rechenstörung Beeinträchtigungen in der

basisnumerischen Zahlenverarbeitung aufweisen (z. B. Bulthé et al., 2019; Decarli et al., 2020; Gliksman & Henik, 2019; de Visscher et al., 2018).

In der Forschung haben sich unterschiedliche *Aufgabenparadigmen* (z. B. non-symbolischer Mengenvergleich) etabliert, von denen angenommen wird, dass diese die verschiedenen Kernsysteme der Zahlenverarbeitung (z. B. ANS oder OTS) adressieren, wodurch die beobachtbare Testleistung wiederum Rückschlüsse auf Beeinträchtigungen in diesen Kernmechanismen erlaubt (s. Abbildung 3). Haberstroh und Schulte-Körne (2022) stellten in einer Meta-Analyse heraus, dass fünf dieser Aufgabenparadigmen besonders geeignet sind, um zwischen Personen mit und ohne Rechenstörung zu differenzieren. Dazu zählen die Simultanerfassung (ab vier Objekten), non-symbolische und symbolische Mengenvergleiche sowie Zahlenstrahl- und Transkodieraufgaben. Studien, die diese Paradigmen bei Personen mit und ohne Rechenstörungen untersucht haben, kommen jedoch zu unterschiedlichen Ergebnissen. Je nachdem, welche Defizite beobachtet wurden, wurden unterschiedliche Ursachenhypothesen formuliert. Diese unterscheiden sich in der Annahme, welcher Kernmechanismus der Zahlenverarbeitung beeinträchtigt ist. Im Folgenden werden drei häufig diskutierte und untersuchte domänenspezifische Kerndefizithypothesen entlang verschiedener Aufgabenparadigmen thematisiert (s. Abbildung 3).

Abbildung 3

Aufgabenparadigmen und Kerndefizithypothesen

			/drei/ 3	
Aufgabenparadigmen		**OTS**	**ANS**	**AD**
Punkte zählen (1–3 / 1–4 Punkte)		x		
Non-symbolischer Mengenvergleich			x	
Symbolischer Mengenvergleich			x	x
Zahlenstrahl			x	x

Anmerkung. Von den exemplarisch dargestellten Aufgabenparadigmen wird angenommen, dass sie die verschiedenen Kernsysteme der Zahlenverarbeitung ansprechen, so dass die beobachtete Testleistung wiederum Rückschlüsse auf Beeinträchtigungen dieser Kernmechanismen zulässt. OTS-Defizit = spezifisches Defizit im Subitizing; ANS-Defizit = allgemeines Defizit in der approximativen Mengenverarbeitung; Access Deficit = Defizite beim Zugriff symbolisch dargestellter Mengen.

OTS-Defizithypothese

Die *OTS-Defizithypothese* (Wilson & Dehaene, 2007) verortet die ursächliche Beeinträchtigung der Rechenstörung im OTS (s. Abbildung 2, Stufe 1). Diese Hypothese basiert auf einer Reihe empirischer Befunde, die feststellten, dass der *Subitizing-Prozess* bei Kindern mit Rechenstörung beeinträchtigt ist (z. B. Andersson & Östergren, 2012; Kuhn et al., 2013; Olsson et al., 2016; Raddatz et al., 2017; Schleifer & Landerl, 2011). Untersucht wird dieser Prozess über ein Aufgabenparadigma, das die Fähigkeit testet,

eine begrenzte Anzahl von Objekten (in der Regel Punkte) so schnell wie möglich zu erfassen (z. B. Landerl, 2019; s. Abbildung 3, Punkte zählen). Je nach Anzahl der Objekte werden zwei unterschiedliche (Zähl-)Prozesse aktiviert. Drei bis vier Objekte können auf einen Blick simultan über das OTS erfasst werden, während mehr als drei oder vier Objekte einen seriellen Zählprozess erfordern (Trick & Pylyshyn, 1993).

Bei Kindern mit Rechenstörung steigen die Reaktionszeiten beim simultanen Erfassen von ein bis drei Objekten steiler an als bei Kindern ohne Rechenstörung (z. B. Schleifer & Landerl, 2011). Möglicherweise müssen Kinder mit Rechenstörung bereits ab drei Objekten auf einen seriellen Zählprozess zurückgreifen (Kuhn 2017). Ebenso zeigen Kinder mit einer Rechenstörung Defizite im seriellen Zählprozess (z. B. Kuhn et al., 2013). Beim Abzählen sind die Kinder nicht zwangsläufig weniger genau, aber deutlich langsamer (z. B. Decarli et al., 2020).

Landerl (2013) stellte in einer Längsschnittuntersuchung fest, dass Kinder mit Rechenstörung über die Grundschulzeit hinweg steilere Reaktionszeitkurven für den Subitizing-Bereich aufweisen als Kinder ohne Rechenstörung. Dies weist auf eine überproportional starke Beeinträchtigung und damit auf eine qualitativ abweichende Entwicklung hin. Obwohl sich auch der Zählprozess von Grundschulkindern mit Rechenstörung von jenen ohne Schwierigkeiten signifikant unterscheidet, scheint die Entwicklung nicht qualitativ unterschiedlich zu verlaufen, was eher auf eine verzögerte Entwicklung des Zählprozesses hindeutet (Landerl, 2013). Skagerlund und Träff (2014) untersuchten Viertklässler*innen mit Rechenstörung und Viert- sowie Zweitklässler*innen ohne Rechenstörung. Die Autor*innen fanden weder signifikante Gruppenunterschiede in der Simultanerfassung (1–3 Punkte) noch beim Abzählen von Punktmengen (5–8 Punkte). Deskriptiv zeigte die Studie allerdings, dass Viertklässler*innen mit Rechenstörungen im Mittel längere Reaktionszeiten für das simultane Erfassen und Abzählen von Punkten benötigen als unbeeinträchtigte Kinder in

der vierten Klasse, im Durchschnitt aber schneller waren als Zweitklässler*innen ohne Rechenstörung. Dies deutet zumindest deskriptiv, eher auf eine verzögerte Entwicklung hin.

Ob Rechenstörungen auf eine selektive Beeinträchtigung im OTS zurückzuführen sind oder nicht, ist aufgrund der heterogenen Evidenz nach wie vor nicht endgültig geklärt. Denn neben den zahlreichen Studien, die eine Beeinträchtigung nachweisen, gibt es auch Studien, die keine Hinweise auf ein beeinträchtigtes OTS fanden (z. B. Decarli et al., 2020; Skagerlund & Träff, 2014). Das gleiche gilt für das Abzählen von Punktmengen (z. B. Andersson & Östergren, 2012; Olsson et al., 2016; Raddatz et al., 2017). Hinzu kommt, dass nicht abschließend beantwortet ist, ob die Defizite, wenn sie denn beobachtet werden, auf einen entwicklungsverzögerten oder qualitativ anderen Subitizing- beziehungsweise Zählprozess hindeuten.

ANS-Kerndefizithypothese

Die ANS-Kerndefizithypothese (Wilson & Dehaene, 2007) geht davon aus, dass Rechenstörungen aus einer generellen Beeinträchtigung in der approximativen Mengenrepräsentation resultieren (s. Abbildung 2, Stufe 1). Ein klassisches Aufgabeparadigma, welches das ANS adressiert, ist die *non-symbolische Mengenvergleichsaufgabe* (z. B. Decarli et al., 2020; Halberda et al., 2008; Landerl et al., 2009). Auch wenn die konkrete Ausgestaltung dieser Aufgabe je nach Studie variiert, besteht die Aufgabe immer darin, so schnell und genau wie möglich zu entscheiden, welche von zwei nebeneinander visuell dargestellten Mengen (in der Regel Punktwolken) quantitativ größer ist (s. Abbildung 3).

Einige Studien stellten fest, dass Grundschulkinder mit einer Rechenstörung (Klassenstufe zwei bis vier) für non-symbolische Mengenvergleichsaufgaben im Mittel signifikant längere Reaktionszeiten benötigen als Kinder ohne Rechenstörung (z. B.

Landerl et al., 2009; Landerl & Kölle, 2009). Decarli et al. (2020) hingegen zeigten, dass Kinder mit Rechenstörung zwar weniger genau beim Bearbeiten dieser Aufgabe sind, aber nicht substanziell mehr Zeit für ihre Entscheidung benötigen. Die Autor*innen schlussfolgern daraus, dass sich Kinder mit einer Rechenstörung auf eine weniger präzise interne Mengenrepräsentation verlassen.

Diese Annahme wird auch durch Ergebnisse unterstützt, die Weberbrüche als Testleistung heranzogen (z. B. Piazza et al., 2010). Der Weberbruch drückt das kleinste Verhältnis zwischen zwei Punktmengen aus, das gerade noch korrekt differenziert werden kann (Landerl et al., 2022). Piazza et al. (2010) zeigten, dass zehnjährige Kinder mit Rechenstörung in etwa eine ähnliche Präzision aufweisen wie fünfjährige Kinder ohne Rechenstörung. Auch Skagerlund und Träff (2014) stellten fest, dass Viertklässler*innen mit Rechenstörung eine geringere Präzision (Testmaß: Weberbruch) aufweisen als unbeeinträchtigte Schüler*innen der zweiten Klassenstufe. Wurden hingegen die richtigen Antworten als Testmaß verwendet, gab es keinen Unterschied zwischen den Gruppen. Mit Blick auf die Ergebnisse von Piazza et al. (2010) sowie Skagerlund und Träff (2014) wäre also denkbar, dass die non-symbolische Mengenrepräsentation von Kindern mit Rechenstörung weniger präzise ist, sich aber mit einer zeitlichen Verzögerung (z. B. von fünf Jahren; Piazza et al., 2010) zunehmend ausdifferenziert und das Niveau unbeeinträchtigter Kinder erreicht.

Ob die ANS-Defizit-Hypothese zutrifft oder nicht, ist nicht abschließend beantwortet. Denn obwohl eine Reihe von Studien die Hypothese unterstützt (z. B. Decarli et al., 2020; Feigenson et al., 2004; Mazzocco et al., 2011; Piazza et al., 2010), konnten andere Studien keine Evidenz für ein Defizit in der Verarbeitung non-symbolischer Mengen liefern (z. B. Busch et al., 2018; Landerl & Kölle, 2009; Rousselle & Noël, 2007).

Access-Defizit

Werden in dem Mengenvergleichsparadigma symbolische Größen (arabische Ziffern) anstelle von non-symbolischen Mengen verwendet (s. Abbildung 3, symbolischer Mengenvergleich), muss der visuell-arabische Input für den Vergleich zweier Zahlen zunächst in eine analoge Größe umgewandelt werden (Henik & Tzelgov, 1982). Diese Interaktion zwischen dem ANS und dem visuell-arabischen Modul ist nach der Annahme des *Access-Defizits* (AD) beeinträchtigt.

Die Hypothese wird vorrangig mit den Befunden von Rousselle und Noël (2007) in Verbindung gebracht. Die Autor*innen fanden keine Evidenz dafür, dass Kinder mit Rechenstörung beim Vergleich non-symbolischer Mengen beeinträchtigt sind. Allerdings hatten die Kinder Schwierigkeiten, wenn die beiden zu vergleichenden Größen symbolisch (als arabische Zahlen) dargestellt wurden. Basierend auf diesem Ergebnis wurde eine dysfunktionale Verbindung zwischen dem visuell-arabischen und dem analogen Modul postuliert, wodurch die Verarbeitung symbolisch dargestellter Mengen beeinträchtigt ist, die Verarbeitung rein non-symbolischer Mengen jedoch nicht (Noël & Rousselle, 2011; Rousselle & Noël, 2007) oder in deutlich geringerem Ausmaß (z. B. Olsson et al., 2016; Schwenk et al., 2017). Schwenk et al. (2017) stellten in einer Meta-Analyse (N = 19 Studien, N = 1630 Proband*innen im Alter von 6 bis 14 Jahren) fest, dass Kinder mit Rechenstörung im Durchschnitt signifikant längere Reaktionszeiten für den Vergleich symbolischer und non-symbolischer Mengen benötigten als Kinder ohne Beeinträchtigung. Allerdings waren die Defizite bei symbolisch dargestellten Mengenvergleichsaufgaben stärker ausgeprägt (Hedges g = 0.75; 95% CI [0.51; 0.99]) als bei non-symbolischen Mengendarstellungen (Hedges g = 0.24; 95% CI [0.13; 0.36]), was eher für die AD-Hypothese spricht.

Die Studie von Skagerlund und Träff (2014) zeigt deskriptiv, dass Viertklässler*innen mit Rechenstörung im Mittel längere Reaktionszeiten benötigten als

Viertklässler*innen ohne Rechenstörungen aber durchschnittlich schneller waren als Zweitklässler*innen ohne Beeinträchtigungen, was im Hinblick auf die symbolische Verarbeitung eher für eine verzögerte Entwicklung spricht (Skagerlund & Träff, 2014).

Doch das Vorliegen einer beeinträchtigten non-symbolischen und symbolischen Mengenverarbeitung allein reicht nicht aus, um eine der beiden Ursachenhypothesen (ANS versus AD) zu bestätigen. Da das ANS die Basis der numerischen Entwicklung bildet (s. Abbildung 2, Stufe 1) und symbolische Mengenrepräsentationen auf dem ANS abgebildet werden (Mundy & Gilmore, 2009), postuliert auch die ANS-Kerndefizithypothese Beeinträchtigungen in der symbolischen Mengenverarbeitung (z. B. Skagerlund & Träff, 2016). Diese werden auf Defizite im ANS zurückgeführt (z. B. Piazza et al., 2010).

Gleichzeitig schließt die AD-Hypothese nicht aus, dass Rechenstörungen mit Schwierigkeiten in der non-symbolischen Verarbeitung einhergehen können. Noël und Rousselle (2011) interpretieren die Defizite in der non-symbolischen Verarbeitung jedoch als Konsequenz einer nicht intakten symbolischen Mengenrepräsentation und nicht als Ursache. Diese Annahme basiert vorrangig auf einer komparativen Gegenüberstellung zehn verschiedener Studien, die das symbolische und non-symbolische Mengenvergleichsparadigma bei Kindern mit und ohne Rechenstörungen verschiedenen Alters untersuchten (Noël & Rousselle, 2011). Der Vergleich zeigte, dass Kinder mit Rechenstörungen im Alter von sechs bis elf Jahren bei symbolisch dargestellten Mengenvergleichen signifikant schlechter abschnitten als Kinder ohne Minderleistungen im Rechnen. Beim Vergleich non-symbolischer Mengen waren hingegen eher ältere Kinder ab acht Jahren beeinträchtigt (Noël & Rousselle, 2011). Auch die Übersichtsarbeit von De Smedt et al. (2013) bestätigt, dass eher ältere Kinder (ab neun Jahren) von Defiziten in der non-symbolischen Mengenverarbeitung betroffen sind. Die Ergebnisse von Skagerlund und Träff (2014) hingegen zeigten, dass Viertklässler*innen mit

Rechenstörung eine geringere ANS-Präzision aufwiesen als unbeeinträchtigte Schüler*innen in der zweiten Klassenstufe. Die symbolische Mengenverarbeitung war jedoch nicht beeinträchtigt, was insgesamt eher für das ANS-Defizit spricht. Die Meta-Analyse von Schwenk et al. (2017) fand im Gegensatz dazu keinen Hinweis auf einen Alterseffekt.

Insgesamt spricht die Forschungslage dafür, dass Schwierigkeiten in der symbolischen Mengenverarbeitung ein robuster Indikator für Rechenstörungen sind (z. B. Haberstroh & Schulte-Körne, 2022; Landerl et al., 2004; 2009; Landerl & Kölle, 2009; Schwenk et al., 2017). Doch ob diese Beeinträchtigung die Folge eines ANS-Defizits oder das eigentliche Kerndefizit darstellt, im Sinne der AD-Hypothese, ist nicht abschließend beantwortet. Zur Beantwortung dieser Fragen sollte das Alter oder die Klassenstufe der Kinder berücksichtigt werden.

Distanz- und Größeneffekt

Zwei weitere Effekte, die im Zusammenhang mit dem (non-)symbolischen Mengenvergleichsparadigma häufig untersucht und nachgewiesen wurden, sind der numerische *Distanz- und Größeneffekt*. Diese beiden Effekte beschreiben folgendes Phänomen: Je kleiner der numerische Abstand zwischen zwei zu vergleichenden Mengen ist und/oder je größer die zu vergleichenden Mengen sind, desto geringer ist die Genauigkeit und umso länger sind die Antwortzeiten (Landerl et al., 2022). Diese Effekte, die sich sowohl bei Kindern mit als auch ohne Rechenstörung zeigen, können durch den Aufbau des mentalen Zahlenstrahls erklärt werden und rekurrieren im Wesentlichen auf die ANS-Präzision (s. Abbildung 2, Stufe 4). Je stärker sich die numerischen Repräsentationen überlappen beziehungsweise je geringer der Abstand zwischen den beiden zu vergleichenden Mengen oder Zahlen ist, desto näher liegen sie in der räumlichen Dimension des mentalen Zahlenstrahls beieinander (s. Abbildung 2, Stufe 4).

Bedingt durch diese räumliche Repräsentation ist es umso schwieriger, Mengenverhältnisse zu diskriminieren, je weniger sie sich unterscheiden. Numerisch nah beieinanderliegende Zahlen (z. B. 6 versus 8) sind somit schwieriger zu diskriminieren als numerisch weiter entfernte Zahlen (z. B. 2 versus 8). Dieses beobachtbare Phänomen beschreibt den *Distanzeffekt* (Landerl et al., 2022). Die Varianz nimmt also zu, dies spiegelt die zunehmende Unsicherheit bei größeren Mengen wider (z. B. Vogel & De Smedt, 2021). Entsprechend werden Mengenrepräsentationen mit zunehmender Größe *unschärfer*, wodurch der beobachtbare *Größeneffekt* erklärt werden kann.

Gemäß der ANS-Defizithypothese zeigen sich also nicht nur Beeinträchtigungen in der Verarbeitung (non-)symbolischer Mengenrepräsentationen, sondern auch in Distanz- und Größeneffekten, da diese die Präzision des ANS widerspiegeln (Wilson & Dehaene, 2007). Einige Studien weisen einen größeren Distanzeffekt für Kinder mit Rechenstörung nach als für Kinder ohne Minderleistungen im Rechnen (Decarli et al., 2020; Price et al., 2007; Mussolin et al., 2010; Raddatz et al., 2017). Dies deutet auf eine geringere Präzision beziehungsweise einen ‚unschärferen' mentalen Zahlenstrahl hin, der zum rechten Ende hin komprimiert ist, was den Zahlenvergleich erschwert (Andersson & Östergren, 2012). Andere Studien konnten dies jedoch nicht bestätigen (z. B. Landerl et al., 2009; Landerl & Kölle, 2009; Schwenk et al., 2017; Skagerlund & Träff, 2014, 2016).

Zahlenstrahl

Auch Zahlenstrahlaufgaben erlauben Rückschlüsse über die Präzision des mentalen Zahlenstrahls. Bei diesem Paradigma besteht die Aufgabe darin, eine numerische Zahl (z. B. 76) auf einer visuell dargestellten Linie, deren Anfangs- und Endpunkte numerisch gekennzeichnet sind (z. B. 0 und 100 oder 0 und 1000), zu platzieren (z. B. Decarli et al., 2023; s. Abbildung 3). Im Hinblick auf die Defizithypothesen deuten Beeinträchtigungen beim Lösen dieser Aufgabe auf eine geringere ANS-Präzision hin. Da aber symbolisch

dargestellte Mengenrepräsentationen zunächst über das visuell-arabische Zahlensystem verarbeitet werden, findet eine Interaktion zwischen den beiden Modulen statt (Feigenson et al., 2004; Siegler & Opfer, 2003). Daher postuliert auch die AD-Hypothese Schwierigkeiten beim Lösen dieser Aufgabe.

Booth und Siegler (2006) gehen davon aus, dass die kognitive Repräsentation von Zahlen zunächst logarithmisch ist und sich durch Übung und Erfahrung zunehmend linear darstellt. Geary et al. (2008) zeigten, dass Grundschulkinder mit Rechenstörung in der ersten und zweiten Klasse im Vergleich zu einer unbeeinträchtigten Kontrollgruppe eher eine logarithmische als linear Zahlenraumvorstellung aufweisen (Zahlenraum bis 100). Die Längsschnittstudie von Landerl (2013) wies dies für Kinder mit Rechenstörung in der zweiten Klassenstufe nach, nicht jedoch für Dritt- und Viertklässler*innen. Landerl et al. (2009), die Grundschulkinder mit Rechenstörung in der zweiten bis vierten Klassenstufe untersuchten, konnten dies für den Zahlenraum bis 100 ebenfalls nicht bestätigen, wohl aber für den Zahlenraum bis 1000. Auch wenn die mittlere Abweichung der Lokalisation der Zahl von der tatsächlichen Zielposition als Testmaß herangezogen wird, zeigt sich, dass Kinder mit Rechenstörung die Zahlen auf dem Zahlenstrahl weniger präzise platzieren (Decarli et al., 2023; Geary et al., 2008; Landerl et al., 2009).

Obwohl Kinder mit Rechenstörungen im Laufe der Grundschulzeit immer präziser werden, bleibt die Ungenauigkeit im Vergleich zu Kindern ohne Rechenstörung bestehen (Landerl, 2013). Ob sich der mentale Zahlenstrahl von Kindern mit Rechenstörung verzögert entwickelt oder ob die Zahlenrepräsentation unpräzise bleibt, ist nicht abschließend beantwortet. Insgesamt sprechen die Ergebnisse eher für eine verzögerte Entwicklung.

Zwischenfazit

Zusammenfassend lässt sich festhalten, dass verschiedene domänenspezifische Kerndefizithypothesen als Ursache der Rechenstörung diskutiert werden. Diese postulieren teils ähnliche Defizite (z. B. Schwierigkeiten beim Verarbeiten symbolisch dargestellter Mengenrepräsentationen), führen diese aber auf unterschiedliche Beeinträchtigungen in den Kernmechanismen (ANS versus AD) zurück. Welcher Kernmechanismus der Zahlenverarbeitung bei Kindern mit Rechenstörung beeinträchtigt ist und die Schwierigkeiten in der Basisnumerik verursacht, ist nicht eindeutig geklärt. Auch die Frage danach, ob die Defizite in der Basisnumerik auf eine entwicklungsverzögerte oder qualitativ abweichende Beeinträchtigung hindeuten, ist nicht abschließend beantwortet. Daher wurden diese Fragen in *Studie I* (Lamb et al., 2024a) untersucht.

2.3 Studie I: Entwicklungsverzögerte Basisnumerik bei Kindern mit Rechenstörung

Studie I (Lamb et al., 2024a) untersuchte (a), ob die Defizite in der Basisnumerik von Grundschulkindern mit Rechenstörung eher auf eine Entwicklungsverzögerung oder auf eine spezifische qualitative Abweichung hindeuten (b) und ob diese Defizite auf eine (selektive) Beeinträchtigung in einem der zentralen Kernsysteme der Zahlenverarbeitung zurückzuführen sind. Dazu wurden verschiedene Aufgabenparadigmen (Punkte zählen, (non-)symbolischer Mengenvergleich, Zahlensteine und Zahlenstrahl) bei N = 480 Grundschulkindern (n = 68 Kinder mit Rechenstörung) der Klassenstufe zwei bis vier untersucht.

Orientiert an dem *Ability-Level-Matching-Ansatz* (z. B. Bradley & Bryant, 1978; Brankaer et al., 2011, 2013; Skagerlund & Träff, 2014) wurden die basisnumerischen Profile von Kindern mit und ohne Rechenstörung untersucht. Bei diesem Ansatz wird

eine Gruppe (z. B. Kinder mit Rechenstörung) mit einer altersgleichen unbeeinträchtigten Vergleichsgruppe und einer jüngeren fähigkeitsgleichen Kontrollgruppe verglichen. Unterscheiden sich die Kinder mit Rechenstörung von der altersgleichen Gruppe ohne Rechenstörung, nicht aber von der jüngeren fähigkeitsgleichen Gruppe, spricht dieses Ergebnis dafür, dass Kinder mit einer Rechenstörung qualitativ die gleiche mathematische Entwicklung durchlaufen, allerdings mit einer zeitlichen Verzögerung. Orientiert an diesem Ansatz wurden die basisnumerischen Profile von Viertklässler*innen mit einer Rechenstörung mit dem Profil der Zweitklässler*innen ohne Rechenstörung verglichen. Außerdem wurden für alle Aufgabenparadigmen Multilevel-Analysen durchgeführt, um zu prüfen, ob die basisnumerischen Defizite von Kindern mit Rechenstörung für eine verzögerte oder qualitativ abweichende Entwicklung sprechen, die nicht durch die Klassenstufe moderiert wird. Die Berücksichtigung der Klassenstufe ist außerdem wichtig, da zwei der Kerndefizithypothesen (ANS und AD) davon ausgehen, dass Kinder mit Rechenstörungen in der Verarbeitung symbolisch dargestellter Mengen beeinträchtigt sind, dieses Defizit aber je nach Zeitpunkt des Auftretens entweder als Kernursache (AD) oder als Folge (ANS) betrachten.

Die Ergebnisse der Multilevel-Analysen zeigten (a), dass die Beeinträchtigungen der Kinder mit Rechenstörung nicht auf qualitative (rechenstörungsspezifische) Unterschiede in den basisnumerischen Fertigkeiten hindeuten, sondern eher auf eine Entwicklungsverzögerung mit Ausnahme des Abzählens von Punkten (4–9/5–9 Punkte). Dieses Ergebnis wurde durch den Vergleich der basisnumerischen Profile von Viertklässler*innen mit Rechenstörung und Zweitklässler*innen ohne Rechenstörung unterstützt. (b) Die Defizite in der Basisnumerik unterstützen eher die Annahme der ANS-Defizit- als der AD-Hypothese und widersprechen der OTS-Defizit-Hypothese. Der vollständige Artikel ist dem Anhang (A) zu entnehmen.

2.4 Mathematische Kompetenzentwicklung

Studie I (Lamb et al., 2024a) zeigte, dass Kinder mit Rechenstörungen Defizite in der Basisnumerik aufweisen, die aus einem Defizit im ANS resultieren könnten. Da die basisnumerischen Fertigkeiten prädiktiv für die mathematische Entwicklung sind (z. B. Braeunig et al., 2021; Habermann et al., 2020; De Smedt et al., 2013; Schneider et al., 2017) und die weitere mathematische Kompetenzentwicklung auf dem ANS aufbaut (z. B. von Aster, 2013), ist es mit Blick auf die Ausgangsfrage der Dissertation – *Wie können Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren?*– wichtig, neben den Defiziten in der basaleren Basisnumerik (vgl. Kapitel 2.2) auch Schwierigkeiten in den komplexeren mathematischen Fertigkeiten berücksichtigen.

Dass die Defizite in der Basisnumerik auf eine Entwicklungsverzögerung hindeuten (Lamb et al., 2024a), spricht dafür, dass Kinder mit Rechenstörungen ähnliche Entwicklungsschritte durchlaufen wie unbeeinträchtigte Kinder, jedoch zeitlich verzögert. Somit eignet sich als theoretisches Rahmenmodell das Kompetenzmodell von Fischer et al. (2017): Die acht *mathematischen Kernkompetenzen* des Modells (s. Abbildung 4) stellen eine modellübergreifende Synopse drei etablierter Kompetenzentwicklungsmodelle der Zahlenverarbeitung aus dem deutschsprachigen Raum dar. Dabei wurde auch das Modell nach von Aster und Shalev (2007), das in Kapitel 2.2 thematisiert wurde, berücksichtigt. Entsprechend gibt es inhaltliche Schnittmengen zwischen den beiden Modellen. In Kapitel 2.2 standen jedoch Defizite in den *basaleren basisnumerischen Fertigkeiten* im Vordergrund, um Rückschlüsse auf ursächliche Beeinträchtigungen in den Kernsystemen der Zahlenverarbeitung ziehen zu können. Im Folgenden werden unter Einbezug weiterer Studienergebnisse Schwierigkeiten in komplexeren mathematischen Fertigkeiten entlang der acht

Kernkompetenzen von Fischer et al. (2017) skizziert, die für die Identifikation von Grundschulkindern mit Rechenstörungen im schulischen Kontext relevant sind.

Abbildung 4

Mathematische Kernkompetenzen nach Fischer et al. (2017)

Anmerkungen. Eigene Darstellung des Kompetenzentwicklungsmodells nach Fischer et al. (2017, S. 27).

Der erste Kompetenzbereich beschreibt den Ausgangspunkt der numerischen Entwicklung, nämlich die Fähigkeit 1) *Mengen zu schätzen und zu vergleichen* (entspricht dem ANS). In Kapitel 2.2 wurde ausführlich beschrieben, dass Kinder mit Rechenstörungen bereits auf dieser ersten Entwicklungsstufe Defizite aufweisen.

Darauf aufbauend entwickelt sich der zweite Kompetenzbereich 2) *Zahlwortreihe und Zählen*, indem ein ordinales Zahlverständnis erworben wird. Die Kinder erkennen, dass die Zahlwörter eine feste Reihenfolge haben, wobei die Abstände zwischen den Zahlen zu diesem Zeitpunkt noch keine Bedeutung für die Kinder haben. Die Kinder erkennen jedoch, dass Mengen durch Wegnehmen oder Hinzufügen verändert werden können. Ein Verständnis für die Mächtigkeit der gezählten Menge (Kardinalität) ist damit jedoch nicht verbunden. Das ordinale Zahlenverständnis ist eine wichtige Voraussetzung für das Lösen verschiedener Aufgaben (z. B. Fischer et al., 2017). Kinder, die diese zentrale Kompetenz nicht erworben haben, haben Schwierigkeiten bei verschiedenen basisnumerischen Aufgaben (z. B. beim Vergleich symbolischer Mengen oder beim Zählen von Punkten, vgl. Kapitel 2.1). Die Schwierigkeiten betreffen aber auch komplexere Zählvorgänge. Insbesondere beim Vorwärts- und Rückwärtszählen (30, 31,

32...) und/oder beim Zählen in Schritten (10, 12, 14...) haben Kinder mit Anzeichen einer Rechenstörung typischerweise Schwierigkeiten (z. B. Gaupp et al., 2004).

Der dritte Kompetenzbereich 3) *Zahlensymbole lesen und schreiben* beschreibt die Fähigkeit, arabische Zahlen zu schreiben und zu lesen. Schwierigkeiten in dieser Kernkompetenz wirken sich besonders negativ aus, da das Lesen und Schreiben von Zahlensymbolen relevant für die Bewältigung aller visuell dargestellten Rechenaufgaben ist (Fischer et al., 2017). Dies gilt mit Ausnahme des non-symbolischen Mengenvergleichs auch für alle in Kapitel 2.1 thematisierten Aufgabenparadigmen.

Studien zeigen zudem, dass Fehler beim Transkodieren typisch für Rechenstörungen sind (Haberstroh & Schulte-Körne, 2022). Dabei werden verbal dargebotene Zahlen fehlerhaft aufgeschrieben (Kuhn et al., 2013). Zuber et al. (2009) klassifizierten verschiedene Fehlerarten, die Kuhn et al. (2013) auch bei Grundschulkindern mit Rechenstörungen nachweisen konnten. Dabei handelt es sich um Inversionsfehler („vierunddreißig“ statt 43), Kompositionsfehler (50015 statt 515) oder lexikalische Fehler (eine oder mehrere Zahlen werden fälschlicherweise ersetzt, z. B. 53 → 54).

In der vierten Kernkompetenz 4) *Zahlengröße verstehen* (kardinales Zahlverständnis) wird das Verständnis dafür erworben, dass jede Zahl auch für eine bestimmte Menge steht. Die Kinder verstehen also, dass beispielsweise beim Zählen von Objekten die Vier nicht nur für das vierte Element, sondern für vier gezählte Objekte steht. Das sich entwickelnde Verständnis für die Kardinalität von Zahlen ermöglicht es den Kindern, einfache Rechenaufgaben zu lösen (Fischer et al., 2017). Kinder ohne kardinales Zahlenverständnis interpretieren Zahlen primär als Zählzahlen und erkennen nicht, dass diese auch Mengen darstellen können (••••• = »fünf«). Ein mangelndes kardinales Zahlenverständnis führt nicht nur zu Problemen beim Lösen basalerer basisnumerischer Aufgaben (z. B. Punkte zählen, vgl. Kapitel 2.1), sondern auch zu Schwierigkeiten bei der Anwendung von Rechenstrategien (Kaufmann & Wessolowski, 2021).

Darauf aufbauend entwickelt sich der Kompetenzbereich 5) *Zahlbeziehungen zu verstehen*. Die Kinder erkennen, dass Zahlen zueinander in Beziehung gesetzt werden können. Dabei verstehen sie, dass Differenzen zwischen Zahlen als Relationen dargestellt werden können (Fischer et al., 2017). Zudem verfügen sie über ein Teil-Ganzes-Verständnis von Zahlen und können nachvollziehen, dass beispielsweise 8 aus 5 und 3 besteht (relationales Zahlverständnis). Defizite im relationalen Zahlverständnis zeigen sich unter anderem beim Lösen von Platzhalteraufgaben (z. B. ? – 2 = 5 statt der 7 wird eine 3 eingetragen) oder bei Aufgaben, bei denen eine fehlende Zahl als Abfolge ergänzt werden muss (z. B. 126 130 __ 138). Kinder und auch Jugendliche mit Rechenstörungen haben dabei Schwierigkeiten (z. B. Meier et al., 2021). Auch für das inhaltliche Verständnis einer Textaufgabe ist es relevant, Zahlen in Beziehung zueinander setzen zu können und ein Verständnis für mathematische Operationen zu haben. Andernfalls werden die Zahlen willkürlich miteinander verknüpft (Kaufmann & Wessolowski, 2021). Auch Schwierigkeiten beim Lösen von Textaufgaben sind typisch für Rechenstörungen (z. B. Haberstroh & Schulte-Körne, 2022; Meier et al., 2021).

Die sechste Kompetenz 6*) Zahlenraumvorstellung* bezieht sich auf die zunehmende Ausdifferenzierung des mentalen Zahlenstrahls, auf dem Zahlen entsprechend ihrer numerischen Größe verortet werden. Dass Kinder mit Rechenstörung hierin beeinträchtigt sein können, wurde in Kapitel 2.2 ausführlich beschrieben.

Die Kompetenzen 7) *Rechnen* (Addition und Subtraktion) sowie das 8) *Stellenwertverständnis* entwickeln sich parallel zu den übrigen Kompetenzbereichen und beschreiben keine separaten Entwicklungsschritte. Das Stellenwertverständnis bezieht sich auf das Verständnis dafür, dass sich größere Zahlen aus der Zusammensetzung von Einern, Zehnern etc. ergeben (Fischer et al., 2017). Dieses Verständnis ist für unterschiedliche mathematische Operationen (z. B. Addition, Subtraktion, Multiplikation) relevant.

Hinsichtlich des Rechnens (Kompetenzbereich 7) zeigen mehrere Studien, dass Kinder mit Rechenstörung wenig oder gar keine Variabilität in ihren Rechenstrategien aufweisen (z. B. Meier et al., 2021) und auf weniger ausgereifte Zählstrategien, wie das verbale Zählen oder das Zählen mit den Fingern zurückgreifen (Kuhn et al., 2017). Solche Strategien sind deutlich zeitaufwendiger und mit höheren Anforderungen an das Arbeitsgedächtnis verbunden (Meier et al., 2021). Entsprechend schneiden Kinder mit Rechenstörung bei der Bearbeitung von Rechenaufgaben schlechter ab als unbeeinträchtigte gleichaltrige Kinder (z. B. Raddatz et al., 2017).

Neben der Anwendung ineffizienter Problemlösungsstrategien manifestieren sich die Schwierigkeiten häufig in Problemen beim Zehnerübergang oder in einem beeinträchtigten Faktenabruf (Busch et al., 2018). Dies zeigt sich darin, dass selbst bei einfachen Rechenaufgaben wie 10 + 2 nicht auf Faktenwissen zurückgegriffen werden kann, sondern die Aufgabe immer wieder neu berechnet werden muss. Defizite im Faktenwissen können sich sowohl in einer erhöhten Fehlerquote (Haberstroh & Schulte-Körne, 2022) als auch in einer verlängerten Bearbeitungszeit widerspiegeln (Endlich et al., 2024).

Resümierend lässt sich festhalten, dass sich anfängliche Schwierigkeiten in den basisnumerischen Fertigkeiten nicht auflösen, sondern fortsetzen. Dies unterstreicht erneut die Notwendigkeit der Früherkennung von Kindern mit einem Risiko für Rechenstörung in der Grundschule.

3 Früherkennung

Für die Identifikation von Kindern mit Anzeichen einer Rechenstörung beziehungsweise die Feststellung von „besonderen Schwierigkeiten im Rechnen“ (Kultusministerkonferenz [KMK], 2007) ist die Schule verantwortlich (Kuhn, 2017). In der Konsequenz bedeutet dies, dass diese Aufgabe bei den Lehrkräften liegt. Diagnostische Verfahren können Lehrkräfte dabei unterstützen (Sikora & Voß, 2018). Allerdings hängt die *Akzeptanz* und *Nutzung* von diagnostischen Testverfahren im schulischen Kontext maßgeblich von der *ökonomischen Effizienz* und von der wahrgenommenen *Nützlichkeit* durch die Lehrkräfte ab (Marx & Lenhard, 2011). Eine höhere ökonomische Effizienz geht tendenziell mit einer geringeren diagnostischen Genauigkeit einher, während genauere und differenziertere Ergebnisse in der Regel mit einem höheren diagnostischen Aufwand verbunden sind (Voß, 2017). Der praktische Einsatz diagnostischer Verfahren in der Unterrichtspraxis bewegt sich somit „[…] in einem Spannungsfeld zwischen Ökonomie und Präzision“ (Voß, 2017, S. 36).

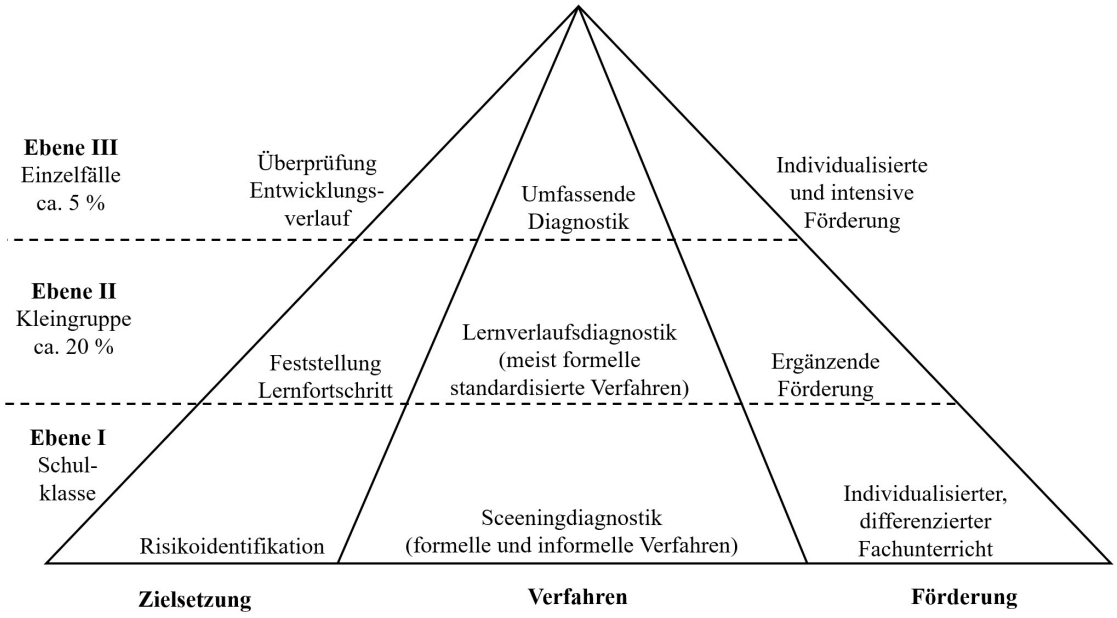
Screeningverfahren sind in der Regel zeitökonomisch anwendbar und können den Lehrkräften dabei helfen, diejenigen Schüler*innen zu identifizieren, die Schwierigkeiten aufweisen und somit Gefahr laufen, schulische Minderleistungen zu entwickeln (Hartung et al., 2021; Voß, 2017). Screeningverfahren zielen darauf ab, Auffälligkeiten frühzeitig zu erkennen, um nach einer Risikobestimmung eine genauere und weiterführende diagnostische Abklärung zu ermöglichen (Breitenbach, 2020). Das bedeutet, dass das Ergebnis eines Screenings keine Grundlage für eine diagnostische Entscheidung darstellt, sondern darauf abzielt, Risikokinder herauszufiltern, um sie daraufhin einer gezielten weiteren Diagnostik zuzuführen (Tröster, 2009). Aus diagnostischer Sicht ergibt sich daraus der ressourcenökonomische Vorteil, dass nicht alle Kinder an einer aufwendigeren diagnostischen Testung teilnehmen müssen. Dies ist vor allem im schulischen Kontext

relevant. Damit bietet das Ergebnis eines Screenings trotz seiner begrenzten Aussagekraft einen zentralen diagnostischen Hinweis (Voß, 2017).

Im *Response-to-Intervention-Ansatz* (RTI; Blumenthal et al., 2014; Fuchs et al., 2018; Jordan et al., 2018) wird dem Einsatz von Screeningverfahren ebenfalls eine zentrale Bedeutung zugeschrieben (Sikora & Voß, 2018). Der RTI-Ansatz ist ein präventiver, dreistufiger Förderansatz (Blumenthal et al., 2014; s. Abbildung 5). Die Kernidee besteht darin, Kinder mit anfänglichen Lernschwierigkeiten durch regelmäßige diagnostische Screenings frühzeitig zu erkennen und so zu fördern, dass sie möglichst schnell den Anschluss an das Klassenniveau finden (z. B. Hartung et al., 2021; Sikora & Voß, 2018). Dies ist auch für Kinder mit Anzeichen einer Rechenstörung zentral. Denn je früher die Schwierigkeiten erkannt werden, desto stärker profitieren die Kinder von mathematikspezifischen Interventionen (z. B. Chodura et al., 2015; Kaufmann & Wessolowski, 2021). Im Rahmen des RTI-Ansatzes werden *informelle* (z. B. Beobachtungen) und *formelle* (z. B. standardisierte Testverfahren) diagnostische Verfahren zu Screeningzwecken eingesetzt (Vossen & Krizan, 2021).

Abbildung 5

Der Einsatz diagnostischer Verfahren im RTI-Modell



Anmerkungen. Eigene modifizierte Darstellung des Einsatzes diagnostischer Verfahren im Rahmen des RTI-Modells nach Hartung et al. (2021, S. 37).

Im folgenden Kapitel 3.1 werden verschiedene diagnostische Verfahren zur Früherkennung von Kindern mit dem Risiko einer Rechenstörung vorgestellt, um deren Potenziale und Herausforderungen vor dem Hintergrund des skizzierten Spannungsfeldes für den Einsatz als Screeningverfahren im schulischen Setting zu diskutieren.

3.1 Diagnostische Verfahren – Chancen und Herausforderungen

Zur Risikoidentifikation von Kindern mit Anzeichen für eine Rechenstörung können verschiedene formelle und informelle diagnostische Verfahren eingesetzt werden. Diese Verfahren lassen sich in direkte und indirekte Beurteilungsverfahren einteilen (s. Abbildung 6).

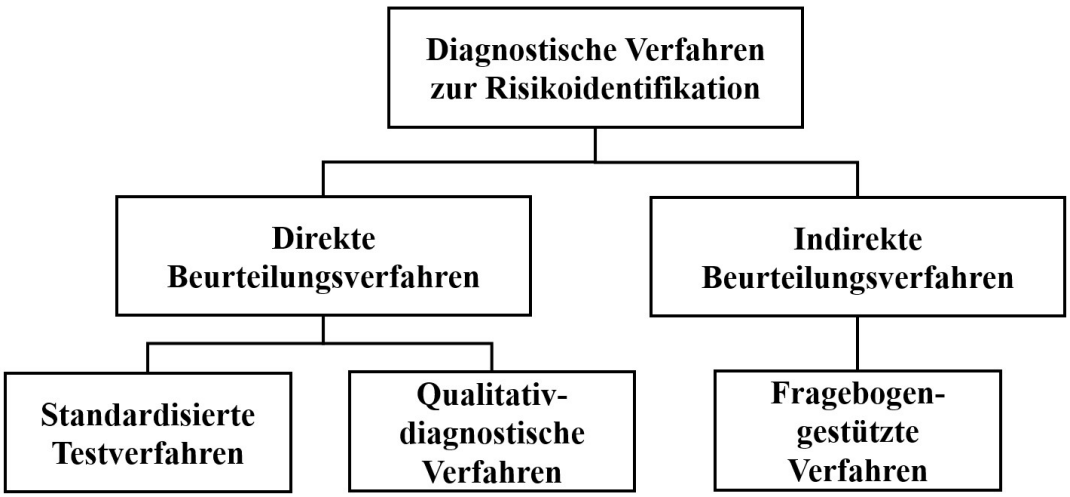
Direkte Beurteilungen werden verwendet, um die Fertigkeiten der Kinder durch eine direkte Demonstration dieser zu ermitteln (Kilday et al., 2012). Dazu können formelle standardisierte Tests, die für die klinische Diagnostik von Rechenstörungen entwickelt wurden, curricular-orientierte Instrumente und qualitative Verfahrensansätze verwendet

werden (Fischer et al., 2017; Kuhn & Schwenk, 2018; Vossen & Krizan, 2021). *Indirekte Beurteilungen* basieren hingegen auf den Einschätzungen Dritter, beispielsweise Lehrkräften oder Eltern. Diese werden in der Regel über Fragebögen erfasst (Kilday et al., 2012; s. Abbildung 6).

Aus den jeweiligen Schwerpunkten und Vorgehensweisen dieser Verfahren ergeben sich Vor- und Nachteile für den Einsatz in der Unterrichtspraxis. Diese werden nachfolgend entlang unterschiedlicher Instrumente skizziert.

Abbildung 6

Diagnostische Verfahren zur Risikoidentifikation



Direkte Beurteilungsverfahren

Direkte Beurteilungen, die auf *standardisiert gemessenen Testleistungen* basieren, ermöglichen, sofern Normwerte vorliegen, eine genaue Einschätzung der Leistung im Vergleich zur altersgleichen Referenzgruppe (Kilday et al., 2012). So kann beispielsweise ermittelt werden, ob ein Kind zu den leistungsschwächsten 10 % (PR ≤ 10) der gleichen Altersgruppe gehört und somit Anzeichen für eine Rechenstörung aufweist. Testverfahren, die zu diesem Zweck eingesetzt werden, unterscheiden sich unter anderem hinsichtlich der inhaltlichen Schwerpunktsetzung aber auch in Bezug auf die Durchführung (Kuhn & Schwenk, 2018). Im folgenden Abschnitt werden standardisierte

Testverfahren, die zur Früherkennung oder Feststellung von Kindern mit Anzeichen einer Rechenstörung eingesetzt werden können, exemplarisch vorgestellt, um deren Potenziale und Herausforderungen für den Einsatz als Screeningverfahren im schulischen Setting zu diskutieren

Standardisierte Testverfahren

Der *Eggenberger Rechentest* (ERT) ist ein lehrplanorientiertes Diagnostikum für Rechenstörungen. Das Verfahren differenziert insbesondere im unteren Leistungsbereich. Da der Test für unterschiedliche Klassenstufen in verschiedenen Versionen mit variierenden inhaltlichen Schwerpunkten vorliegt, kann die Testreihe von der ersten (ERT 0+; Lenart et al., 2013) bis zur achten Klassenstufe (ERT JE; Holzer et al., 2017) eingesetzt werden. Da die vorliegende Dissertation Grundschulkinder der Klassenstufen zwei bis vier untersucht, wird beispielhaft der ERT 3+ (Holzer et al., 2010) vorgestellt, der am Ende der dritten bis zur Mitte der vierten Klassenstufe eingesetzt werden.

Der Test ermöglicht eine detaillierte Erfassung der mathematischen Kompetenzen und eignet sich zur Diagnostik aber auch zur Prozess- und Qualitätsdokumentation (Holzer et al., 2010). Allerdings fokussiert der ERT 3+ die Rechenfertigkeiten, obwohl basalere basisnumerische Defizite ebenfalls indikativ für Rechenstörungen sind (vgl. Kapitel 2.2). Der Test wird für den Einsatz als Klassenscreening empfohlen, für eine Gruppentestung mit Auswertung müssen jedoch bis zu zwei Schulstunden eingeplant werden. Für das letzte Kindergartenjahr bis zur Mitte der ersten Klassenstufe liegt ein Screeningtest zur frühen Identifizierung von Risikokindern vor, der die Basisnumerik berücksichtigt und zeitlich ressourcenschonender eingesetzt werden kann (Testdurchführung ca. 20 Minuten, ERT 0+; Lenart et al. 2013). Für die Klassenstufen zwei bis vier fehlt ein solches Instrument allerdings.

Die *Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern* (ZAREKI-R) ist für den Einsatz in der ersten bis zu vierten Klassenstufe geeignet (von Aster et al., 2013). Der standardisierte Test erfasst vorrangig basisnumerische Fertigkeiten, beinhaltet aber auch komplexere Aufgaben. Der ZAREKI-R eignet sich für die Diagnostik und bietet Hinweise für die Förderplanung (von Aster et al., 2013). Nachteilig ist, dass die Kinder die Aufgaben durch motorische, mündliche oder schriftliche Reaktionen beantworten müssen. Damit ist der ZAREKI-R eher für die Individualdiagnostik, aber nicht als Screening einer Schulklasse geeignet.

Der *Kettenrechner für dritte und vierte Klassen* (KR 3 – 4; Roick et al., 2011) kann effizient und zeitökonomisch (Testdurchführung ca. 25 Minuten) als Leistungsscreening eingesetzt werden. Inhaltlich fokussiert sich der KR 3 – 4 auf die Grundrechenarten und den mathematischen Faktenabruf. Damit berücksichtigt der Test, dass Kinder mit Rechenstörung Schwierigkeiten beim zeitabhängigen Lösen komplexer arithmetischer Faktenaufgaben haben (Roick et al., 2011). Basisnumerische Fertigkeiten werden jedoch nicht erfasst.

Computeradministrierte Testverfahren wie der *CODY-Mathetest für die 2. – 4. Klasse* (CODY-M 2 – 4; Kuhn et al., 2017) haben im Vergleich zu papierbasierten Instrumenten den ressourcenökonomischen Vorteil, dass die Testergebnisse automatisch vom Programm generiert werden. Inhaltlich erfasst der CODY-M 2 – 4 sowohl die basale und komplexe Zahlenverarbeitung als auch die Rechenfertigkeiten. Der Test kann als Einzel- oder Gruppentest administriert werden. Die Bearbeitungszeit variiert je nach individueller Lösungsgeschwindigkeit zwischen 25 und 35 Minuten. Die Testergebnisse werden in Form von individuellen Testprofilen dargestellt, die als Grundlage für die Förderplanung genutzt werden können. Allerdings setzt der Test eine digitale Infrastruktur in den Schulen voraus, die nicht immer flächendeckend gegeben ist.

Neben diesen beispielhaft skizzierten standardisierten Instrumenten (für eine umfassende Übersicht s. S3-Leitlinie; Schulte-Körne & Haberstroh, 2018), die vorrangig zur Statusdiagnostik eingesetzt werden, gibt es auch *qualitativ-diagnostische Verfahren*.

Qualitativ-diagnostische Verfahren

Qualitativ-diagnostische Instrumente zählen ebenfalls zu den direkten Beurteilungsverfahren, allerdings werden die Informationen über die Leistung des Kindes hier meist in einer bilateralen Interviewsituation zwischen Testleitung und Kind erhoben. Dabei werden die Kinder gebeten, ihre Fertigkeiten zu demonstrieren, indem sie beispielsweise die Anzahl von Gegenständen zählen und der Testleitung die Gesamtzahl mitteilen (Kilday et al., 2012). Qualitativ-diagnostische Verfahren sind im Vergleich zu Leistungstests weniger stark standardisiert, da die Ergebnisse in hohem Maße von der Interpretation der Testleitung abhängig sind. Die Durchführung solcher Verfahren erfordert entsprechend besonders viel Erfahrung und diagnostisches Wissen seitens der Testleitung (Faix et al., 2023). Zudem werden geringere Anforderungen an die klassischen Gütekriterien Objektivität, Reliabilität und Validität gestellt (Benz et al., 2015; Kuhn & Schwenk, 2018).

Ein solches Instrument ist das *Elementar Mathematische Basisinterview* (EMBI; Peter-Koop et al., 2007). Dieses dient einer qualitativen-förderdiagnostischen Einschätzung und soll eine gezielte Betrachtung der Entwicklung mathematischer Fähigkeiten (z. B. in den Bereichen Zählen oder Rechnen) ermöglichen. Allerdings ist das EMBI zeitaufwendig in der Durchführung und nicht als Klassenscreening geeignet. Ähnlich komplex in der Anwendung ist die *Qualitative Diagnostik von Rechenschwierigkeiten im Grundlagenbereich Arithmetik* (QUADRIGA; Wehrmann, 2003) oder das *Informelle Verfahren* von Kaufmann und Wessolowski (2021).

Wenn bei einem Kind bereits eine Rechenstörung festgestellt wurde, kann in einem weiteren diagnostischen Schritt ein qualitatives Verfahren dazu beitragen, die Lösungsstrategien von Kindern mit Rechenstörung aufzudecken, da sich diese Verfahren auf die Denkschritte oder Rechenoperationen des jeweiligen Kindes fokussieren und so detaillierte Hinweise auf förderrelevante Problembereiche liefern können (Faix et al., 2023; Fischer et al., 2017). Für den Einsatz als Screeninginstrument sind diese Verfahren aufgrund der geringen Standardisierung sowie des hohen Aufwands und der Komplexität jedoch nicht geeignet.

Eine weitere Möglichkeit zur Identifikation von Kindern mit Anzeichen für eine Rechenstörung bieten indirekte Beurteilungsverfahren, zu denen fragebogengestützte Instrumente zählen. Fragebögen sind im Vergleich zu den oben beschrieben Instrumenten relativ einfach zu handhaben (Reid et al., 2014) und erfordern keine Testung des Kindes (Cabell et al., 2009).

Indirekte Beurteilungsverfahren

Um Kinder mit Anzeichen einer Rechenstörung zu identifizieren, wurden zahlreiche indirekte Beurteilungsverfahren für Lehrkräfte veröffentlicht. Dabei handelt es sich um *Fragebögen* und *Symptom-Checklisten*. Ein indirektes Beurteilungsverfahren, das basierend auf wissenschaftlichen Standards der Test- und Fragebogenkonstruktion konzipiert und hinsichtlich klassischer (z. B. Objektivität, Reliabilität und Validität) und screeningrelevanter Güteeigenschaften (z. B. Sensitivität, Spezifität, RATZ-Index, positive und negative Korrektheit; Breitenbach, 2020; Marx & Lenhard, 2011) überprüft wurde, konnte für den deutschsprachigen Raum nicht identifiziert werden. International veröffentliche Fragebögen und Symptom-Checklisten weisen ähnliche Mängel auf. Daher werden nachfolgend fragebogengestützte Verfahren beispielhaft vorgestellt, die

für den Einsatz im Unterricht konzipiert wurden, aber aus wissenschaftlicher Perspektive Schwächen aufweisen.

Jacobs und Petermann (2012) entwickelten eine Symptom- beziehungsweise Checkliste für Lehrkräfte. Diese beschreibt 20 spezifische Schwierigkeiten, die charakteristisch für Kinder mit einer Rechenstörung sind (z. B. Fehler beim Abzählen konkreter Objekte oder das Überspringen von Objekten oder Zahlen beim Zählen). Die Lehrkräfte sollen diese Schwierigkeiten hinsichtlich ihrer Auftretenshäufigkeit auf einer vierstufigen Skala („sehr häufig, oft, gelegentlich, gar nicht“) einschätzen. Treten mehr als die Hälfte der Fehler sehr häufig oder oft auf, kann dies nach Angabe der Autor*innen als Hinweis auf eine Rechenstörung interpretiert werden. Unklar bleibt, auf welche Altersgruppe sich die Checkliste bezieht.

Der Deutsche Bildungsserver des Leibniz-Instituts für Bildungsforschung und Bildungsinformation (DIPF, 2024) verweist auf den *symptomorientierten Kriterienkatalog – Rechenschwäche* von Wieneke (o. J.), herausgegeben vom Zentrum zur Therapie der Rechenschwäche. Der Kriterienkatalog dient in erster Linie der Sensibilisierung von Lehrkräften für die mathematischen Schwierigkeiten von Kindern und Jugendlichen mit einer Rechenstörung. Der Kriterienkatalog richtet sich an drei Zielgruppen: 1) Kinder der ersten Klassenstufe (22 Items), 2) Kinder der zweiten und dritten Klassenstufe (22 Items) sowie 3) Kinder ab der vierten Klassenstufe bis hin zur Berufsschule (49 Items). Die zu beurteilenden Schwierigkeiten variieren je nach adressierter Zielgruppe und umfassen Fertigkeiten der Basisnumerik aber auch komplexere mathematische Kompetenzen. Die Einschätzung durch die Lehrkräfte erfolgt auf einer fünfstufigen Skala, die die Auftretenshäufigkeit der Schwierigkeiten von „nie“ bis „100 %“ abbildet (Wieneke, o. J.). Obwohl der Kriterienkatalog eine breite Zielgruppe adressiert und ein breites Spektrum unterschiedlicher mathematikbezogener Schwierigkeiten abbildet, weist das Instrument mehrere methodische Mängel auf. So sind

beispielsweise die meisten Items nicht trennscharf formuliert und auch das Antwortformat („nie – selten – häufig – fast immer – 100 %“) erscheint durch die Angabe „100 %“ aus testtheoretischer Perspektive eher ungeeignet. Unklar bleibt, ab welchem Summenscore eine weitere diagnostische Abklärung erfolgen sollte.

Der vom Zentrum für Rechentherapie (o. J.) entwickelte *Symptomfragebogen für Eltern und Lehrkräfte* setzt sich aus 17 zu beurteilenden Schwierigkeiten zusammen, die typischerweise bei Rechenstörungen auftreten (z. B. „Treten besondere Schwierigkeiten bei sogenannten Platzhalteraufgaben [ ] – 5 = 3 auf?“). Ein Item wird dann angekreuzt, wenn die Aussage auf das zu beurteilende Kind zutrifft. Treten fünf oder mehr der genannten Schwierigkeiten gehäuft auf, wird eine diagnostische Abklärung des Vorliegens einer Rechenstörung empfohlen. Unklar ist, welche Altersgruppe der Fragebogen adressiert.

Auch vom Mathematischen Institut zur Behandlung der Rechenschwäche/Dyskalkulie wurde ein *Symptomfragebogen zum Erkennen von elementaren Lernschwierigkeiten im Grundlagenbereich der Mathematik (Rechenschwäche/Dyskalkulie)* entwickelt und veröffentlicht (o. J.). Der Fragebogen richtet sich an unterschiedliche Zielgruppen: Eltern, Lehrkräfte sowie Psycholog*innen und adressiert Kinder der Grund-, Förder- und weiterführenden Schulen bis zur fünften Klasse. Der Symptomfragebogen gliedert sich in drei Teile: Mathematischer Bereich (24 Items), Auffälligkeiten im Lernverhalten (9 Items) und Alltäglichen Bereichen (7 Items). Die 24 Items des mathematischen Bereichs beschreiben mathematikbezogene Schwierigkeiten (z. B. „Das Kind verwechselt Vorgänger und Nachfolger einer Zahl.“), die hinsichtlich ihrer Auftretenshäufigkeit auf einer vierstufigen Skala („fast immer“, „oft“, „gelegentlich“, „nie“, Ausweichkategorie: „nicht bekannt“) eingeschätzt werden sollen. Inhalte, die noch nicht im Unterricht vermittelt wurden, sind von der Beurteilung auszuschließen. Treffen nur wenige der Items zu, wird zu innerschulischer Förderung

geraten. Bei anhaltenden Schwierigkeiten trotz Intervention wird eine förderdiagnostische Untersuchung empfohlen. Kinder, die in mehreren Bereichen Schwierigkeiten aufweisen, sollten einer individuellen Diagnostik zugewiesen werden und entsprechende Förderung erhalten. Doch auch zu diesem Screeninginstrument liegen keine Angaben zu den Gütekriterien oder Cut-Off-Werten vor.

Auch international wurde eine Vielzahl fragebogengestützter Screeninginstrumente für Lehrkräfte veröffentlicht. Allerdings weisen auch diese Verfahren ähnliche Schwächen auf, wie die Instrumente aus dem deutschsprachigen Raum. Beispiele hierfür sind: die *Checklist for Dyscalculia* von Chinn (2023), die *Possible Indicators of Dyscalculia Teacher Checklist* (Education Scotland, o. J.) oder die *Dyscalculia: Checklist*, welche vom Special Educational Needs and Disability Independent Support Service (SENDISS) (o. J.) herausgegeben wurde.

Für psychometrisch untersuchte Screeninginstrumente Instrumente wie dem *Colorado Learning Difficulties Questionnaire* (CLDQ; Willcutt et al., 2011) liegen zwar Angaben zur Sensitivität (85 bis 89 %) und Spezifität (47 bis 48 %) vor (Koriakin et al., 2019), doch inhaltlich betrachtet bildet der Fragebogen das Konstrukt Rechenstörung nicht differenziert ab. In dem Instrument wird eher global erfasst, ob Rechenschwierigkeiten vorliegen. Zwischen Schwierigkeiten in der Basisnumerik und den Rechenfertigkeiten wird nicht ausreichend differenziert.

Neben den Fragebögen und Checklisten, die für Praktiker*innen entwickelt wurden, existieren weitere Fragebögen, die zu Forschungszwecken konzipiert wurden. Beispielswiese entwickelte Dögnitz (2022) einen Fragebogen zur Bestimmung der kriterialen Validität zwischen einem standardisierten Leistungstest und der Lehrkräfteeinschätzung. Der Fragebogen erfasst 17 Beeinträchtigungen, die im Zusammenhang mit Rechenschwierigkeiten auftreten können. Die meisten Items beziehen sich auf mathematikbezogene Schwierigkeiten, unter anderem die Anwendung

wenig ausgereifter Rechenstrategien (z. B. Fingerzählen), willkürliches Vermischen von Regeln oder Zahlendreher. Die übrigen Inhalte adressieren emotional-soziale Verhaltensweisen (z. B. Konzentrationsschwierigkeiten, Ängste etc.), externale Einflussfaktoren sowie Fragen zum schulischen und außerschulischen Umfeld. Bei der Beantwortung der Fragen können die Lehrkräfte zwischen drei kategorialen Antworten auswählen: „ja“, „nein“ oder „keine Angabe“. Doch auch zu diesem Instrument liegen keine Angaben zu den psychometrischen Eigenschaften und Gütekriterien vor, sodass dessen potenzielle Eignung für den Einsatz als Screeningverfahren nicht beurteilt werden kann.

Zwischenfazit

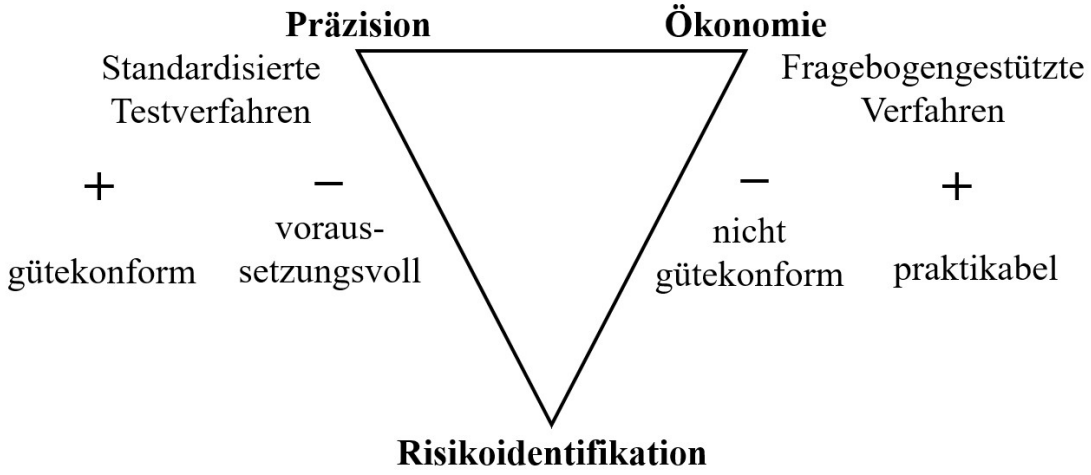
Formelle standardisierte Testverfahren bieten im Vergleich zu den vorgestellten Checklisten und Fragebögen den Vorteil, dass sie üblicherweise theoretisch fundiert konstruiert wurden und eine objektiv-reliable Diagnostik mit normbasierter Einordnung der Leistung ermöglichen (z. B. CODY-M 2 – 4; Kuhn et al., 2017). Für die Feststellung von Rechenstörungen sind diese Instrumente unverzichtbar. Allerdings ist ihr Einsatz zu Screeningzwecken in der schulischen Praxis mit Einschränkungen hinsichtlich der Ökonomie und Praktikabilität verbunden (Wagner & Ehlert, 2017). So erfordert die Anwendung in der Regel eine zeitintensive Einarbeitung (z. B. ZAREKI-R; von Aster et al., 2013) oder es wird eine technische Infrastruktur vorausgesetzt (z. B. CODY-M 2 – 4; Kuhn et al., 2017). Andere Screeningverfahren, die sich zeiteffizienter in die Unterrichtspraxis integrieren lassen, fokussieren sich hingegen eher auf komplexere Rechenfertigkeiten (z. B. KR 3 – 4; Roick et al., 2011) und vernachlässigen die Basisnumerik, obwohl Defizite in diesem Bereich typisch für Rechenstörungen sind. Qualitativ-diagnostische Verfahren (z. B. EMBI; Peter-Koop et al., 2007) sind besonders zeitaufwendig und nicht als Screeningverfahren geeignet.

Fragebögen sind im Vergleich dazu relativ einfach zu handhaben (Reid et al., 2014) und erfordern keine Testung des Kindes (Cabell et al., 2009). Dies ist nicht nur ressourcenökonomisch vorteilhaft, sondern auch für die Kinder selbst von Vorteil. Denn Kinder mit Rechenschwierigkeiten leiden häufig unter Mathematikangst (z. B. Rubinsten & Tannock, 2010; Meier et al., 2021). Somit kann eine Testung für diese Kinder eine Belastungssituation darstellen. Inhaltlich berücksichtigen die meisten der beispielhaft skizzierten Fragebögen und Checklisten wichtige Leitsymptome der Rechenstörung, doch auf welcher theoretischen Basis die Instrumente konstruiert wurden, bleibt weitgehend offen. Zudem liegen keine Angaben zu den psychometrischen Eigenschaften und Gütekriterien (z. B. Objektivität, Reliabilität oder Validität; Moosbrugger & Kelava, 2020) vor, sodass die Eignung der Instrumente für den Einsatz in der Praxis nicht beurteilt werden kann.

Resümierend lässt sich festhalten, dass die beispielhaft skizzierten direkten (z. B. standardisierte Testverfahren) und indirekten Beurteilungsverfahren (z. B. Checklisten und Fragebögen) einerseits Chancen, aber auch Herausforderungen für den Einsatz in der Unterrichtspraxis mit sich bringen. Daher ist es notwendig, alternative Screeninginstrumente für Lehrkräfte zu konstruieren, die dem „Spannungsfeld zwischen Präzision und Ökonomie“ (Voß, 2017, S. 36) begegnen (s. Abbildung 7). Im nachfolgenden Abschnitt wird formuliert, welche Anforderungen ein solches Screeninginstrument erfüllen muss.

Abbildung 7

Diagnostische Verfahren im Spannungsfeld zwischen Präzision und Ökonomie



Anforderungen an einen Screeningfragebogen für Lehrkräfte

In Anlehnung an Moser Opitz und Bern (2008) werden folgende Anforderungen an ein diagnostisches Instrument für Lehrkräfte gestellt: 1) Das Instrument sollte die zentralen mathematischen Kernkompetenzen (vgl. Kapitel 2.4) und häufig beobachtbare Schwierigkeiten, die im Zusammenhang mit Rechenstörungen im Grundschulalter auftreten (vgl. Kapitel 2.1 und 2.4) abbilden. Daher sollten *Defizite in der Basisnumerik* Bestandteil eines Fragebogens zur Risikoidentifikation von Rechenstörungen durch Lehrkräfte sein. Denn wenn Lehrkräfte ihre Einschätzungen ausschließlich anhand der schulischen Leistungen der Kinder vornehmen, wird vernachlässigt, dass Kinder mit Rechenstörungen nicht nur Schwierigkeiten in den Rechenfertigkeiten haben (Fischer et al., 2015). Da sich Defizite in den Kernsystemen der Zahlenverarbeitung negativ auf die weitere mathematische Kompetenzentwicklung auswirken, ist es darüber hinaus wichtig, dass Schwierigkeiten in den *Rechenfertigkeiten* berücksichtigt werden (vgl. Kapitel 2.4).

Die Ergebnisse aus *Studie I* (Lamb et al., 2024a) deuten darauf hin, dass die basisnumerischen Fertigkeiten von Kindern mit Rechenstörungen zeitlich verzögert sind. Dies impliziert, dass Kinder mit Rechenstörungen qualitativ die gleichen

Entwicklungsschritte der Zahlenverarbeitung durchlaufen wie Kinder ohne Rechenstörung, jedoch zeitlich verzögert. Daher ist es sinnvoll, sich bei der Konstruktion eines Instruments zur Risikoidentifikation von Rechenstörungen auf Entwicklungsmodelle der Zahlenverarbeitung zu stützen (z. B. Fischer et al., 2017; vgl. Kapitel 2.4).

Um die Akzeptanz und die wahrgenommene Nützlichkeit des Instruments maximal zu gestalten (Marx & Lenhard, 2011, Voß, 2017, Walter, 2020), sollten 2) Lehrkräfte den Screeningfragebogen ohne umfangreiche Schulung selbstständig ausfüllen, auswerten und interpretieren können. 3) Der Fragebogen sollte wenige, aber inhaltlich und testtheoretisch informative Items enthalten. 4) Es sollten Testergebnisse generiert werden, die leicht interpretierbar sind und klare Handlungsempfehlungen ermöglichen (in Anlehnung an Moser Opitz & Bern, 2008). Darüber hinaus müssen 5) allgemein anerkannte Gütekriterien (z. B. Reliabilität, Objektivität, Validität; AERA, APA, NCME, 2014) und screeningrelevante Güteeigenschaften (z. B. Sensitivität, Spezifität, Positive Korrektheit, Negative Korrektheit, Youden-Index, Relativer Anstieg der Trefferquote gegenüber der Zufallstrefferquote [RATZ-Index]; Tröster, 2009) erfüllt sein, die es ermöglichen, Kinder mit einem Risiko für eine Rechenstörung mit hinreichender Genauigkeit zu identifizieren.

In *Studie II* (Lamb et al., 2024b) wurde ein Lehrkräftefragebogen zur Früherkennung von Grundschulkindern mit Anzeichen für eine Rechenstörung, der diese Anforderungen erfüllen soll, vorgestellt und psychometrisch validiert und hinsichtlich seiner Gütekonformität überprüft.

3.2 Studie II: Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule

In *Studie II* (Lamb et al., 2024b) wurde ein Lehrkräftefragebogen zur Früherkennung von Grundschulkindern mit Anzeichen für eine Rechenstörung vorgestellt und hinsichtlich seiner psychometrischen Eigenschaften untersucht. Dazu wurden drei Forschungsfragen untersucht: (a) Lässt sich die theoriebasierte zweifaktorielle Struktur (Basisnumerik und Rechenfertigkeiten) des Fragebogens empirisch bestätigen? (b) Entspricht das psychometrische Messmodell auf Itemebene einem einparametrigen oder einem zweiparametrigen logistischen Testmodell? (c) Wie zuverlässig werden Kinder mit und ohne Anzeichen für eine Rechenstörung durch den Fragebogen identifiziert?

Zur Beantwortung der Forschungsfragen wurden die Daten von N = 377 Grundschulkindern (Klassenstufe zwei bis vier) erhoben. (a) Die Ergebnisse einer konfirmatorischen Faktorenanalyse bestätigten die theoretische zweidimensionale Struktur des Fragebogens. Aufgrund der hohen Korrelationen beider Faktoren r = .93 [.87; .98] wurden die Subskalen (Basisnumerik und Rechenfertigkeiten) sowie die Gesamtskala untersucht. (b) Der Likelihood-Ratio-Test des Vergleichs eines einparametrigen (1PL-Modell) und zweiparametrigen logistischen Testmodells (2PL-Modell) ergab eine bessere Passung für das 2PL-Modell. Da das informationstheoretische Fitmaß des BIC für den FERMAT-Gesamtscore und die Subskala Rechenfertigkeiten eher eine bessere Anpassung an das 1PL-Modell nahelegte und die Schätzungen der Personenparameter verschiedener unidimensionaler IRT-Modelle (Items-Response-Theorie) sehr hoch korrelieren, wurde der Fragebogen Rasch-konform ausgewertet. (c) Mittels Receiver-Operating-Characteristics-Analysen (ROC-Analysen) wurden verschiedene screeningrelevante Gütekriterien ermittelt. Da die Kriterien zur Klassifikation der Rechenstörung sowohl in der Wissenschaft als auch in der Praxis variieren und weiterhin uneinheitlich verwendet werden (vgl. Kapitel 2.1), wurde die

Treffsicherheit des Fragebogens zur Erfassung mathematischer Fertigkeiten (FERMAT) anhand zwei gebräuchlicher Cut-Off-Kriterien (PR ≤ 7 und PR ≤ 10) geprüft. *Studie II* zeigte, dass der FERMAT über zufriedenstellende Screeningeigenschaften verfügt. Die Sensitivität und Spezifität variierten je nach Kriterium zwischen 57.6 % und 69.8 % beziehungsweise 82.5 % und 91.4 %. Die RATZ-Indizes lagen zwischen .511 und .613. Die Screeningeigenschaften der beiden Teilbereiche (basisnumerische Fertigkeiten und Rechenfertigkeiten), in die sich der FERMAT gliedert, waren ebenfalls überwiegend zufriedenstellend. Damit identifiziert der FERMAT Kinder mit und ohne testdiagnostische Anzeichen für eine Rechenstörung mit angemessener Treffsicherheit. Der vollständige Artikel ist dem Anhang (B) zu entnehmen.

4 Diagnostische Einschätzungen von Lehrkräften

Wenn Lehrkräfte einschätzen sollen, ob die mathematischen Fertigkeiten ihrer Schüler*innen beeinträchtigt sind, spielt ihre diagnostische Kompetenz eine zentrale Rolle. Diese Schlüsselkompetenz (z. B. Helmke 2017), die sich auf alle Schüler*innen und nicht nur auf diejenigen mit Lernschwierigkeiten bezieht (Wittich & Kuhl, 2021), wird nachfolgend beschrieben und operationalisiert. Entlang des heuristischen Modells zur Akkuratheit diagnostischer Urteile von Lehrkräften nach Südkamp et al. (2012) werden Merkmale skizziert, die die diagnostische Einschätzung der Lehrkräfte beeinflussen können. Da es bisher nur wenige Studien gibt, die die diagnostische Kompetenz von Lehrkräften in Bezug auf die Identifikation von Grundschulkindern mit mathematikbezogenen Schwierigkeiten untersucht haben, werden ergänzend Studienergebnisse dargestellt, die die Mathematikleistung im Allgemeinen oder andere Domänen (z. B. Leseleistung) in den Blick genommen haben.

Forschungsinteresse an der Diagnosekompetenz

Leistungsbeurteilungen gehören zum Alltag von Lehrkräften (Hesse & Latzko, 2017; Karst et al., 2014). Lehrkräfte treffen diagnostische Urteile zu verschiedenen Zwecken und mit unterschiedlichem Formalisierungsgrad (White & Gunstone, 1992). Im engeren Sinne beschreibt die diagnostische Kompetenz von Lehrkräften die auf das fachspezifische unterrichtliche Handeln bezogene Fähigkeit, Schüler*innen zutreffend zu beurteilen und Lernanforderungen korrekt einzuschätzen (Karst, 2012; Schrader & Helmke, 1987).

Diagnostische Kompetenzen sind ein elementarer Teilbereich des professionellen Handelns von Lehrkräften (z. B. Baumert & Kunter, 2006) und entscheidend für die Lernentwicklung einzelner Schüler*innen (z. B. Förster & Souvignier, 2017), aber auch für das Leistungsniveau ganzer Schulklassen (Anders et al., 2010). Das diagnostische

Urteil der Lehrkraft entscheidet darüber, ob Schüler*innen gezielte Fördermaßnahmen erhalten oder nicht (z. B. Fischer et al., 2015; Schmitterer & Brod, 2021). Gerade in relativ kohärent aufbauenden Unterrichtsfächern wie der Mathematik ist dies besonders relevant, da unentdeckte Lernschwierigkeiten negative Folgen für den Lern- und Entwicklungsverlauf der Schüler*innen haben (Schrader, 2013; Sikora & Voß, 2018; Tröster, 2009). Eine genaue Diagnosekompetenz ist somit eine Voraussetzung dafür, dass Lehrkräfte den Unterricht unter Berücksichtigung der heterogenen Lernvoraussetzungen ihrer Schüler*innen adäquat gestalten können (Gebhardt et al., 2018; Rogalla & Vogt, 2008; Schrader et al., 2006).

Im deutschsprachigen Raum ist das forschungs- und bildungspolitische Interesse an der diagnostischen Kompetenz von Lehrkräften seit der Veröffentlichung der PISA-Studie deutlich gestiegen (Hesse & Latzko, 2017). Die PISA-Studie (2000) zeigte, dass Hauptschullehrkräfte große Schwierigkeiten bei der Identifizierung von Schüler*innen mit geringer Lesekompetenz hatten (Artelt et al., 2001). Angeregt durch die Ergebnisse der PISA-Studie wurden verschiedene Maßnahmen formuliert und Projekte initiiert, die dazu beitragen sollen, die diagnostische Kompetenz der Lehrkräfte besser zu verstehen und zu stärken (z. B. Artelt & Gräsel, 2009; Hosenfeld et al., 2002). Ein Beispiel hierfür ist die Einführung von Standards in der Lehramtsausbildung, durch die der Aspekt der Diagnosekompetenz fester Bestandteil des Curriculums aller Schulformen wurde (KMK, 2004).

Diagnosekompetenz als mehrdimensionales Konstrukt

Es existieren verschiedene Begriffe, die synonym zur diagnostischen Kompetenz verwendet werden: „Diagnosekompetenz“, „diagnostische Expertise“ oder „diagnostisches Wissen“ (Lorenz, 2011, S. 16). Zudem wird das Konstrukt nicht einheitlich definiert (Praetorius et al., 2012). Zwar ähneln sich die unterschiedlichen

Definitionen, doch im Detail unterscheiden sie sich. Im deutschsprachigen Raum existiert zudem eine begriffliche Ungenauigkeit in der synonymen Verwendung von Urteilsgenauigkeit und diagnostischer Kompetenz. Die diagnostische Kompetenz bildet die Voraussetzung dafür, dass Lehrkräfte die Leistungen ihrer Schüler*innen zutreffend einschätzen können. Die Urteilsgenauigkeit ist somit eher ein Ergebnis der diagnostischen Kompetenz (Lorenz, 2011).

Die Diagnostische Kompetenz als Bestandteil der pädagogischen Diagnostik bezieht sich nach Ingenkamp und Lissmann (2008, S. 13) auf alle diagnostischen Tätigkeiten, die darauf abzielen Lehr- und Lernprozesse so zu gestalten, dass individuelles Lernen optimiert wird. Diese Diagnosetätigkeiten setzen verschiedene Kompetenzen voraus.

Nach Helmke (2017, S. 119) setzt sich die „diagnostische Expertise“ aus dem methodischen und prozeduralen Wissen (z. B. Einsatz von Methoden, um Schüler*innenleistungen einschätzen zu können) sowie dem konzeptuellen Wissen (z. B. Kenntnisse über Urteilsfehler) zusammen.

Die Diagnosekompetenz von Mathematiklehrkräften umfasst nach Brunner et al. (2011) das *fachdidaktische* und *pädagogisch-psychologische Wissen*. Beide Kompetenzbereiche fächern sich in einzelne Facetten auf, wobei die diagnostische Kompetenz im Wesentlichen aus a) dem Wissen über das mathematische Denken von Schüler*innen, b) dem Wissen über die mathematischen Aufgaben und c) dem Wissen über die Leistungsbeurteilung besteht (Brunner et al., 2011, S. 217).

In Anlehnung an Brunner et al. (2011) sowie Ingenkamp und Lissmann (2008) bezieht sich die diagnostische Kompetenz der Lehrkräfte auf a) das Wissen über das diagnostische Potenzial von Aufgaben, b) die Einschätzung des Vorwissens und des Entwicklungsstandes der Schüler*innen, c) das Erkennen von Schwierigkeiten, d) die Beurteilung von Lösungsstrategien und -prozessen sowie e) die Kenntnis verschiedener standardisierter Testverfahren, die für diagnostische Zwecke eingesetzt werden können.

Wenn Lehrkräfte die mathematischen Schwierigkeiten ihrer Schüler*innen mittels des FERMAT einschätzen sollen (Lamb et al., 2024b), müssen sie den Entwicklungsstand der Schüler*innen beurteilen, Schwierigkeiten erkennen und Lösungsstrategien und -prozesse einschätzen können. Damit spiegelt sich ein Teil des mehrdimensionalen Konstrukts auch im FERMAT wider.

Der FERMAT zielt darauf ab, Kinder mit Anzeichen für eine Rechenstörung zu identifizieren. Daher benötigen die Lehrkräfte darüber hinaus auch spezifisches Wissen über die typischen Schwierigkeiten von Kindern mit einer Rechenstörung. Dieses Wissen kann im Kompetenzmodell von Baumert und Kunter (2006) dem *pädagogisch-psychologischen Wissen* zugeordnet werden. Es ist allerdings nicht klar definiert, welches Professionswissen Lehrkräfte im Hinblick auf Rechenstörungen erwerben sollen (Landesregierung Nordrhein-Westfalen, 2020). Im Rahmen der Entwicklung eines Wissenstests für Lehrkräfte unternahmen Bender et al. (2024) den Versuch, zentrale Wissensbereiche zu identifizieren und zu definieren, über die Lehrkräfte im Hinblick auf Rechenstörungen verfügen sollten. Nach den Einschätzungen der Autor*innen zählt dazu unter anderem das Wissen über die Ursachen, Entstehung, Klassifikation, Charakteristika und Identifikation der Rechenstörung. Dies deckt sich in Teilen auch mit den Inhalten der Subskala Teilleistungsstörungen des Tests zum Wissen über verschiedene Diversitätsbereiche von (angehenden) Lehrkräften (DiWi; Steinmayr et al., 2022), der das professionsübergreifende pädagogisch-psychologische Wissen erfasst.

4.1 Urteilsakkuratheit

Dass unter dem Oberbegriff der diagnostischen Kompetenz verschiedene Facetten zusammengefasst werden und es sich um einen hochkomplexen Prozess handelt, spiegelt sich auch in der Forschung wider. Die Forschung hat eine Vielzahl unterschiedlicher Aspekte der Diagnosekompetenz untersucht, unter anderem das diagnostische Denken

von Lehrkräften und informationsverarbeitende Prozesse, die dem Diagnostizieren im Unterricht zugrunde liegen (Loibl et al., 2020).

Die meisten Studien untersuchten allerdings die Genauigkeit diagnostischer Urteile (Schrader & Helmke, 1987), also die Fähigkeit, Lernprozesse und -ergebnisse von Schüler*innen akkurat einzuschätzen (Artelt & Gräsel, 2009). In der Forschung werden vorrangig drei verschiedene Urteilskomponenten unterschieden, die drei Facetten der *Urteilsakkuratheit* abbilden: die Niveau-, Differenzierungs- und Rangordnungskomponente (Schrader & Helmke, 1987).

Die *Niveaukomponente* veranschaulicht, ob Lehrkräfte dazu tendieren, die zu beurteilenden Schüler*innenmerkmale zu über- oder zu unterschätzen (Praetorius et al., 2012; s. Abbildung 8). Dazu wird die Differenz zwischen dem mittleren Urteil der Lehrkraft und der mittleren Schüler*innenleistung berechnet (Karst & Bonefeld, 2020). Die Differenz gibt Auskunft darüber, wie weit die Einschätzung der Lehrkraft von der tatsächlich gezeigten Schüler*innenleistung abweicht. Idealtypisch sollte die Differenz den Wert 0 annehmen (Hesse & Latzko, 2017). Werte über 0 weisen auf eine Überschätzung der Leistung durch die Lehrkraft hin. Werte unter 0 zeigen eine Unterschätzung der Leistung an (Karst & Bonefeld, 2020). Studien belegen, dass die Leistungen der Schüler*innen tendenziell eher überschätzt werden (z. B. Hosenfeld et al., 2002; Kaiser et al., 2015).

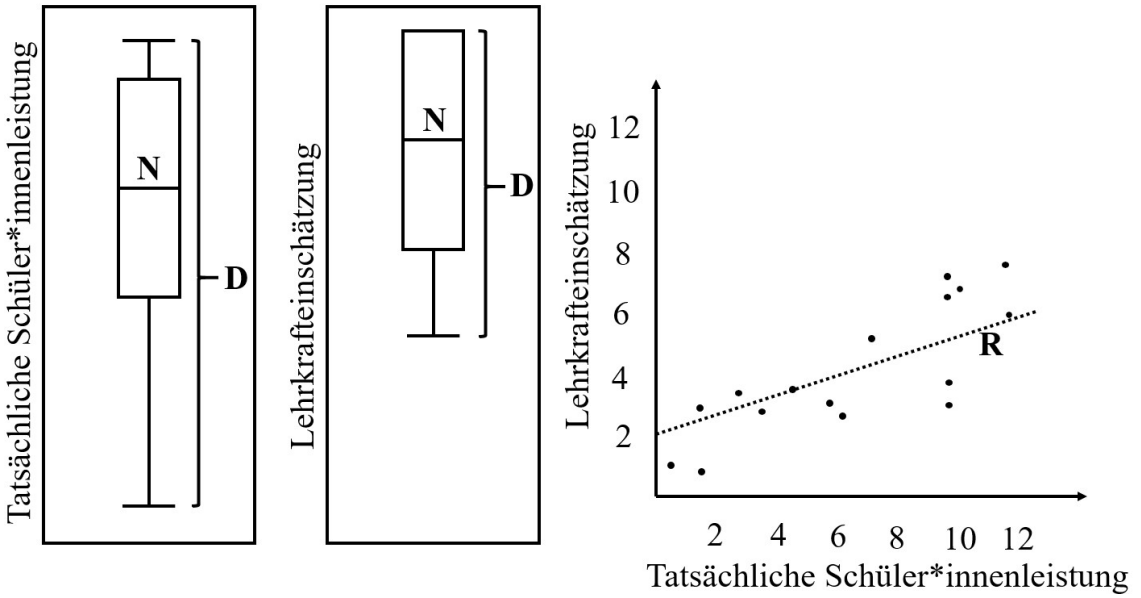
Die *Differenzierungs- oder Streuungskomponente* zeigt, ob Lehrkräfte dazu neigen, die Streuung des zu beurteilenden Schüler*innenmerkmals zu über- oder zu unterschätzen (Helmke, 2017; Praetorius et al., 2012; s. Abbildung 8). Diese wird durch das Verhältnis der Streuungen zwischen den Lehrkräfteeinschätzungen und den Schüler*innenmerkmalen bestimmt (Hesse & Latzko, 2017). Diese sollte im besten Fall den Wert 1 annehmen. Werte über 1 zeigen eine Überschätzung der Leistungsstreuung an. Werte unter 1 zeigen eine Unterschätzung an (Karst & Bonefeld, 2020). Eine

unzureichende Berücksichtigung der Merkmalsstreuung deutet auf eine mangelnde Sensibilität der Lehrkraft für die Unterschiede zwischen den Lernenden hin (Tröster, 2018). Eine Überschätzung der Merkmalsstreuung indiziert demgegenüber eine Überbetonung der Unterschiede zwischen den Lernenden. Dies bedeutet, dass die Lehrkraft größere Unterschiede zwischen den Schüler*innen wahrnimmt, als tatsächlich vorhanden sind (Tröster, 2018). Kaiser et al. (2015) zeigten, dass Lehrkräfte eher dazu tendieren, die Streuung der Schüler*innenleistung zu unterschätzen. Somit wurden die Leistungen der Schüler*innen homogener beurteilt als dies tatsächlich der Fall war.

Während die ersten beiden Komponenten eher Urteilstendenzen beschreiben, betrifft die *Rangordnungskomponente* das Erkennen von Fähigkeitsabstufungen zwischen den Schüler*innen. Diese Urteilskomponente wird als originärer Kennwert der diagnostischen Kompetenz betrachtet (Schrader, 1989; Schrader et al., 2006; s. Abbildung 8). Diese Komponente ist unter verschiedenen Begriffen bekannt: Korrelations-, Vergleichs- oder Rangkomponente oder auch als diagnostische Sensibilität beziehungsweise Sensitivität (Brunner et al., 2011; Hosenfeld et al., 2002; Schrader, 1989; Spinath, 2005) und beschreibt, wie präzise die Lehrkraft die Schüler*innenleistungen in ihrer Klasse einschätzen kann (Karst & Bonefeld, 2020). Dies bedeutet, dass die Übereinstimmung zwischen den Einschätzungen der Lehrkräfte (z. B. Mathematikleistung) und den in der Realität tatsächlich vorhandenen Merkmalsausprägungen (z. B. gemessen durch standardisierte Leistungstests) dargestellt wird (z. B. Hosenfeld et al., 2002; Schrader & Praetorius 2018). Zur Quantifizierung wird üblicherweise der statistische Zusammenhang (Produkt-Moment-Korrelation) zwischen den Einschätzungen der Lehrkräfte und den tatsächlichen Schüler*innenmerkmalen berechnet (Hesse & Latzko, 2017; Hosenfeld et al., 2002; Südkamp et al., 2012; Tröster, 2018). Entsprechend kann der Korrelationskoeffizient Werte zwischen –1 und 1 annehmen.

Abbildung 8

Schematische Darstellung der drei Komponenten der Urteilsakkuratheit



Anmerkung. Eigene modifizierte Darstellung nach Karst und Bonefeld (2020, S. 269). N = Niveaukomponente: Differenz zwischen mittlerem Lehrkräfteurteil und mittlerer Schüler*innenleistung; D = Differenzierungskomponente: Verhältnis der Streuungen ($s_{Lehrkräfte}$ / $s_{Schüler*innen}$). R = Rangkomponente abgebildet durch die gestrichelt dargestellte Gerade.

Neben der Berechnung der statistischen Korrelation kann die Übereinstimmung zwischen den Lehrkräfteeinschätzungen und den Schüler*innenleistungen auch anhand von Übereinstimmungskoeffizienten quantifiziert werden (z. B. Cohen‘s Kappa; Cohen, 1960). Jedoch steht die Berechnung von Übereinstimmungskoeffizienten und Korrelationsanalysen aufgrund ihrer geringen Aussagekraft seit geraumer Zeit in der Kritik (Karst & Bonefeld, 2020; Karst et al., 2017). Unter anderem deswegen, weil in diesen Ansätzen die Unterschiede zwischen den Schüler*innen (Ebene 1) und Lehrkräften beziehungsweise Schulklassen (Ebene 2) nicht getrennt voneinander modelliert werden können. Für eine methodisch adäquatere Analyse solcher hierarchisch strukturierten Daten werden Multilevel-Analysen empfohlen (Karst et al., 2017). Diese erlauben nicht nur die simultane Berücksichtigung unterschiedlicher Analyseebenen, sondern ermöglichen auch die Integration zusätzlicher Einflussfaktoren – etwa

Schüler*innenmerkmale wie Geschlecht oder Intelligenz –, welche die Einschätzung der Lehrkräfte potenziell beeinflussen (Karst & Bonefeld, 2020). Dennoch basieren die Ergebnisse vieler Studien auf korrelativen Zusammenhangsmaßen oder Übereinstimmungskoeffizienten.

Forschungsstand

Meta-Analysen und Übersichtsarbeiten, die die Urteilsgenauigkeit von Lehrkräften anhand korrelativer Übereinstimmungen untersuchten, deuten darauf hin, dass die Einschätzungen der Lehrkräfte über die Fähigkeiten ihrer Schüler*innen im Allgemeinen recht präzise sind (Übersichtsarbeiten: Hoge & Coladarci, 1989; Kaufmann, 2020; Südkamp et al., 2012). In einem systematischen Literaturreview berichten Hoge und Coladarci (1989), dass der Median der Korrelationen bei r = .66 lag (N = 16 Studien). Kaufmann (2020) replizierte die Studie von Hoge und Coladarci (1989) mit einem meta-analytischen Ansatz. Die Ergebnisse zeigten, dass die Urteilsgenauigkeit von Hoge und Coladarci (1989) sogar unterschätzt wurde (r = .80). Die Meta-Analyse von Südkamp et al. (2012) kam zu ähnlichen Ergebnissen mit einer durchschnittlichen Korrelation von r = .63 (N = 75 Studien).

Fischer et al. (2015) untersuchten, ob die Einschätzungen der Lehrkräfte über das Vorliegen oder Nicht-Vorliegen von mathematikbezogenen Schwierigkeiten mit den Ergebnissen objektiver Leistungstests übereinstimmen. Dazu schätzten die Lehrkräfte auf einer dreistufigen Skala ein, ob die Kinder ihrer Klassen a) unbedingt, b) eventuell oder c) keinen Förderbedarf in Mathematik haben. Anschließend wurde die Leistung der Schüler*innen objektiv über einen Test erfasst. Basierend auf den Testergebnissen wurden die Kinder in drei Gruppen eingeteilt: a) PR ≤ 10 unbedingt Förderbedarf, b) 10 < PR ≤ 15 eventuell Förderbedarf und c) PR > 15 kein Förderbedarf. Um die Übereinstimmung zwischen der Lehrkräfteeinschätzung und den Ergebnissen der

Klassentestung zu prüfen, wurden Prozentwerte und gewichtete Übereinstimmungskoeffizienten (Cohen‘s Kappa; Cohen, 1960, 1968) berechnet. Die Ergebnisse zeigten, dass die Einschätzung der Lehrkräfte und die Ergebnisse der Testung bei 170 der 222 Kinder (77 %) übereinstimmte. Eine nach Klassenstufen separierte Auswertung ergab, dass die Übereinstimmung in der ersten Klasse eher gering ausfiel (59 %; $\kappa_w = 0.18$) und in den Klassenstufen zwei bis vier mittelmäßig übereinstimmte (ca. 80 %; $\kappa_w = 0.49$ bis 0.53; für eine Interpretation der Koeffizienten s. Landis & Koch, 1977).

Schmitterer und Brod (2021) erforschten, wie akkurat Lehrkräfte einschätzen, ob ihre Schüler*innen eine Leseförderung benötigen. In der Studie gaben $N = 64$ Lehrkräfte von $N = 697$ Drittklässler*innen an, ob eine Leseförderung für ihre Schüler*innen: a) nicht, b) möglicherweise oder c) auf jeden Fall notwendig sei. Anschließend wurden die Schüler*innen basierend auf ihren objektiv gemessenen Testleistungen ebenfalls in drei Gruppen eingeteilt: a) T-Wert < 35: Kinder, die eine Leseförderung erhalten sollten, b) 35 ≤ T-Wert < 40: Kinder, die möglicherweise eine Förderung erhalten sollten und c) T-Wert ≥ 40: Kinder, die keine Förderung benötigen. Die Ergebnisse zeigten, dass die Lehrkräfte etwa 50 % der Schüler*innen für potenziell förderbedürftig einschätzen, davon erzielten allerdings 25 % bis 40 % der Kinder unauffällige Testergebnisse (T-Wert > 40).

Auch wenn die Ergebnisse insgesamt zeigen, dass es große Übereinstimmungen zwischen den Einschätzungen der Lehrkräfte und den objektiven Testergebnissen der Kinder gibt, berichten Studien übereinstimmend von deutlichen Unterschieden in der Genauigkeit, mit der Lehrkräfte die Leistungen ihrer Schüler*innen einschätzen (Hoge & Coladarci, 1989; Hosenfeld et al., 2002; Lorenz 2011; McElvany et al., 2009; Schrader et al., 2006; Spinath 2005; Südkamp et al., 2012). Die Ergebnisse von Hoge und Coladarci (1989) zeigten, dass die Korrelationen von geringer Übereinstimmung bis hin zu fast perfekter Genauigkeit reichen ($r = .28$ und $r = .92$).

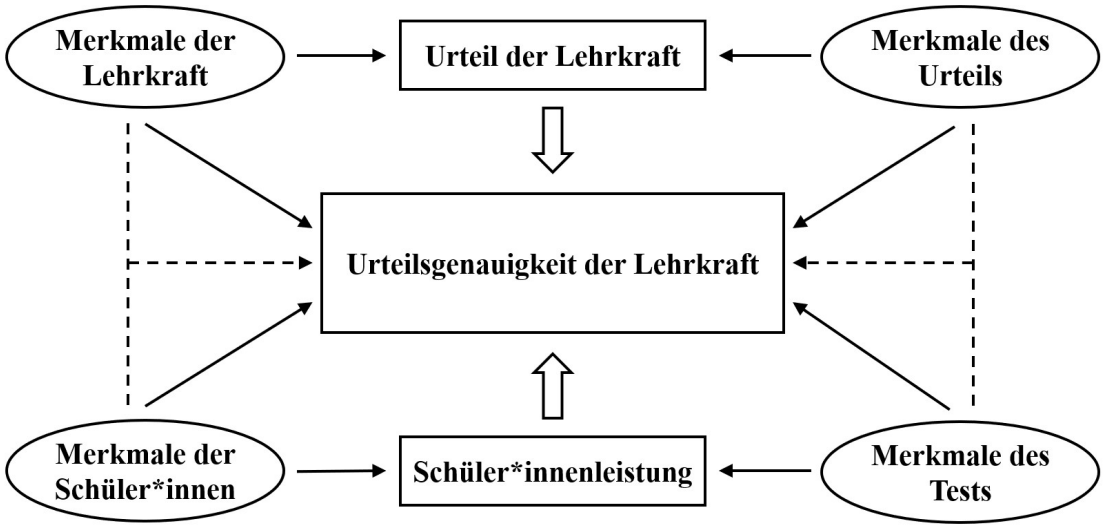
Dies wirft die Frage auf, welche Faktoren die Einschätzungen der Lehrkräfte beeinflussen. Im nachfolgenden Kapitel 4.2 werden verschiedene Einflussfaktoren entlang des heuristischen Modells der Akkuratheit diagnostischer Urteile von Lehrkräften (Südkamp et al., 2012) dargestellt.

4.2 Einflussfaktoren auf die diagnostischen Einschätzungen von Lehrkräften

In dem *heuristischen Modell der Akkuratheit diagnostischer Urteile von Lehrkräften* (Südkamp et al., 2012) werden die Einflussfaktoren auf vier Bereiche zurückgeführt: Auf Merkmale, die sich auf 1) die *Lehrkraft* (z. B. Stereotype), 2) die *Schüler*innen* (z. B. Verhalten), 3) die Merkmale des *Tests* (z. B. Länge des Tests) und 4) das abzugebende *Urteil* (z. B. global oder spezifisch) beziehen (s. Abbildung 9). Diese Einflussfaktoren können sich gegenseitig beeinflussen, aber auch direkt auf die Urteilsgenauigkeit wirken. Im Fokus der vorliegenden Dissertation stehen die Merkmale der Schüler*innen.

Abbildung 9

Heuristisches Modell der Akkuratheit diagnostischer Urteile von Lehrkräften nach Südkamp et al. (2012)



Anmerkungen. Eigene Darstellung des heuristischen Modells der Akkuratheit diagnostischer Urteile von Lehrkräften nach Südkamp et al. (2012, S. 756).

Merkmale der Lehrkraft

Beurteilungsunterschiede zwischen den Lehrkräften können unter anderem auf Urteilstendenzen oder Beurteilungsfehler zurückgeführt werden (Helmke, 2017; Hesse & Latzko, 2017). Beim *Strenge-Effekt* bewerten Lehrkräfte die Leistungen systematisch strenger beziehungsweise gewichten Minderleistungen stärker als andere Lehrkräfte, während beim *Milde-Effekt* die Leistungen der Schüler*innen tendenziell weniger streng bewertet werden (z. B. Helmke, 2017). Der Milde-Effekt kann möglicherweise darauf zurückgeführt werden, dass Lehrkräfte befürchten, dass ihre pädagogische Kompetenz bei schlechten Beurteilungen infrage gestellt wird oder sie ihre Beliebtheit bei Schüler*innen wahren wollen (Hesse & Latzko, 2017). Der Strenge-Effekt könnte darin begründet sein, dass Lehrkräfte ihre Fachkompetenz betonen wollen oder Schüler*innen für die Notwendigkeit einer intensiven Auseinandersetzung mit dem Fach sensibilisieren möchten (Hesse & Latzko, 2017). Andere Urteilstendenzen beziehen sich eher auf die Merkmale der Schüler*innen. Auf diese fokussiert sich die vorliegende Arbeit.

*Merkmale der Schüler*innen*

Schüler*innenmerkmale können zu kognitiven Verzerrungen im Sinne eines logischen Fehlschlusses führen. Beim *logischen Fehlschluss* wird von einem Schüler*innenmerkmal auf ein anderes Merkmal geschlossen, ohne dass dies empirisch begründet wäre (Helmke, 2017).

Kaiser et al. (2015) erforschten in einer Experimentalstudie, ob Lehrkräfte bei der Vergabe von Mathematiknoten ihr Urteil auf mathematikspezifische urteilsrelevante (schriftliche und mündliche Mathematikleistung) Informationen stützen oder ob auch mathematikunspezifische urteilsirrelevante Schüler*innenmerkmale (z. B. Leistung im Unterrichtsfach Deutsch und Intelligenz) in die Benotung einfließen. Auch wenn die tatsächliche Mathematikleistung der Schüler*innen am stärksten in die Benotung

einfloss, wurde die Beurteilung der Lehrkräfte zusätzlich durch urteilsirrelevante Schüler*innenmerkmale beeinflusst. So führte eine höhere Intelligenz der Schüler*innen dazu, dass die Lehrkräfte die mathematische Leistung der Schüler*innen besser einschätzten. Eine niedrigere Intelligenz der Schüler*innen wiederum bewirkte das Gegenteil. Ähnlich verhielt es sich mit der Leistung im Unterrichtsfach Deutsch. Schüler*innen mit einer guten Note im Unterrichtfach Deutsch erhielten eine bessere Mathematiknote und Schüler*innen mit einer schlechteren Deutschnote wurden auch im Unterrichtsfach Mathematik schlechter bewertet. Dieser Effekt wird durch kognitive Verzerrungen wie dem *logischen Fehlschluss* erklärt (z. B. Helmke, 2017) und weist darauf hin, dass die Leistungen der Schüler*innen in einem Bereich, die Beurteilung in anderen Bereichen beeinflussen (Kaiser et al., 2015; Mack et al., 2023).

Auch die Ergebnisse von Fischer et al. (2015) deuten darauf hin, dass die Lehrkräfte bei der Einschätzung mathematischer Schwierigkeiten ihre Einschätzung nicht ausschließlich auf Basis mathematikspezifischer Informationen vornehmen. So gab es Hinweise darauf, dass die Lehrkräfte unter anderem sprachliche Schwierigkeiten als Indikator für mathematikspezifische Schwierigkeiten heranziehen.

Mathematikunspezifische Fertigkeiten wie die Sprache (Peng et al., 2020), Intelligenz (z. B. Peng et al., 2019; Wyschkon et al., 2009) und Lesefertigkeiten (Akin, 2022; Singer & Strasser, 2017) sind mit der mathematischen Leistung korreliert. Zudem werden sie zur Konstruktvalidierung mathematischer Screeninginstrumente (z. B. KR 3 – 4) herangezogen (z. B. Roick & Hasselhorn, 2005). Vor diesem Hintergrund könnte angenommen werden, dass es plausibel ist, bei der Einschätzung mathematischer Fertigkeiten auch die Leistung in anderen Domänen mit einzubeziehen. Doch mehrere Aspekte sprechen dagegen.

So sind mathematikspezifische Fertigkeiten weitaus bedeutsamer für die Vorhersage der mathematischen Leistung als beispielsweise die Intelligenz (z. B. Gallit

et al., 2018; Stern, 2003). Krajewski (2008) zeigte, dass die Intelligenz zu Beginn der Grundschulzeit die mathematischen Fertigkeiten am Ende der Grundschulzeit nicht direkt, sondern nur indirekt über mathematische Basiskompetenzen vorhersagte. Lerkkanen et al. (2005) belegten in einer Längsschnittstudie, dass die Leistung in Mathematik und das Leseverständnis in den ersten beiden Schuljahren zwar eng miteinander verbunden waren, die mathematische Leistung aber das spätere Leseverständnis voraussagte. Umgekehrt war dies nicht der Fall. Auch Duncan et al. (2007) stellten fest, dass die Vorhersage späterer Lesefertigkeiten durch frühe mathematische Fertigkeiten besser gelingt als andersherum. Vor diesem Hintergrund sollten mathematikspezifische Fertigkeiten, die für die Vorhersage der mathematischen Leistung insgesamt bedeutsamer sind (z. B. Chen & Li, 2014; Gallit et al., 2018; Kuhn et al., 2019), die Lehrkräfteeinschätzung über mathematikbezogene Schwierigkeiten bestimmen.

Hinzu kommt, dass die Intelligenz ein latentes Merkmal ist und damit nicht direkt beobachtbar ist. Um die Intelligenz der Schüler*innen einzuschätzen zu können, muss die Lehrkraft beobachtbare, proximale Merkmale heranziehen (Helmke, 2017). Daher überrascht es nicht, dass es Lehrkräften nur dann gut gelingt die Intelligenz ihrer Schüler*innen einzuschätzen, wenn die Leistung und Intelligenz der zu beurteilenden Schüler*innen übereinstimmen (Spinath, 2005). Die Autorin interpretiert dies als Beleg dafür, dass sich die Lehrkräfte bei der Einschätzung der Intelligenz vor allem an den Leistungen der Schüler*innen orientieren. Divergieren Intelligenz und Leistung (z. B. in Mathematik), wie es bei Kindern mit einer Rechenstörung der Fall ist, könnte dies im Umkehrschluss zu Fehleinschätzungen der Schüler*innen führen. Denn so würden mathematikbezogene Schwierigkeiten geringeren intellektuellen Fähigkeiten zugeschrieben werden.

Lorenz und Artelt (2009) stellten zudem fest, dass Grundschullehrkräfte, die die sprachlichen Leistungen ihrer Schüler*innen genau einschätzen können, nicht automatisch auch gute Diagnostiker*innen für Schulleistungen in Mathematik und umgekehrt sind. Zudem gibt es Studien, die darauf hinweisen, dass Lehrkräfte die mathematischen Fertigkeiten weniger präzise einschätzen als die sprachbezogenen Fertigkeiten (z. B. Demaray & Elliot, 1998; Eckert et al., 2006; Mack et al., 2023). Auch wenn dieser Effekt in der Meta-Analyse von Kaufmann (2020) nur deskriptiv und nicht signifikant bestätigt werden konnte, impliziert das Ergebnis, dass eine Lehrkraft, die die mathematischen Fertigkeiten ihrer Schüler*innen ungenau, die Lesefertigkeiten jedoch genau einschätzt und diese Fertigkeit als Indikator für die mathematische Leistung einbezieht, zu einer ungenaueren Einschätzung der mathematischen Fertigkeiten gelangen würde (Tröster, 2018). Daraus resultiert die Gefahr, dass Förderbedarfe oder Lernprobleme übersehen werden.

Logische Fehlschlüsse werden auch innerhalb einer Domäne beobachtet. Schmitterer und Brod (2021) zeigten, dass die Einschätzung der Lehrkräfte über die Notwendigkeit einer Leseförderung vorrangig auf der Rechtschreibleistung der Kinder basierte und nicht auf deren Leseleistung. Wird die Einschätzung der Lehrkräfte nicht systematisch überprüft (z. B. über einen Test), kann dies in der Konsequenz dazu führen, dass Kinder mit isolierten Leseschwierigkeiten eine geringere Wahrscheinlichkeit auf eine Leseförderung haben. Ähnliche Effekte sind für den mathematischen Bereich zu erwarten.

Darüber hinaus berichtet die Forschung von *Referenzfehlern*. Dabei orientiert sich die Lehrkraft bei ihrer Einschätzung nicht an objektiven Kriterien (z. B. mathematische Kernkompetenzen), sondern an anderen Bezugsnormen, beispielsweise am Leistungsniveau der Schulklasse (soziale Bezugsnorm; Helmke, 2017). Dies wurde unter anderem von Schmitterer und Brod (2021) bestätigt. Die Studie zeigte, dass die

Einschätzung der Lehrkräfte durch das durchschnittliche Leistungsniveau der Klasse beeinflusst wurde. Schüler*innen mit den gleichen Lesefertigkeiten wurden je nach Durchschnittsniveau der Klasse als (nicht) förderbedürftig eingeschätzt. Kinder in einer Klasse mit einem hohen durchschnittlichen Leseniveau wurden mit größerer Wahrscheinlichkeit als förderbedürftig eingestuft, während das Gegenteil für Kinder in einer Klasse mit einem niedrigen durchschnittlichen Leseniveau galt.

Nicht nur das Leistungsniveau der Klasse, sondern auch die individuelle Leistung der Schüler*innen kann einen Einfluss auf die Lehrkräfteeinschätzung haben. Studien zeigen, dass Lehrkräfte Kinder mit niedrigem Leistungsniveau weniger akkurat beurteilen (Begeny et al., 2008; Coladarci, 1986; Wagner, 2024). Lorenz (2011) zeigt, dass leistungsstarke Schüler*innen häufiger unterschätzt und leistungsschwache Schüler*innen häufiger überschätzt werden. Wagner und Ehlert (2017, 2019) untersuchten im Rahmen einer universitären Lehrveranstaltung, ob angehende Grundschullehrkräfte anhand von Videovignetten die mathematischen Kompetenzen von Schüler*innen hinsichtlich ihres Entwicklungsstandes und einer eventuell notwendigen Förderung einschätzen können. Die Studierenden konnten die mathematischen Kompetenzen sicherer beurteilen als den Entwicklungsstand des Kindes. Dieser wurde tendenziell eher überschätzt. Somit wurde der bestehende Förderbedarf nicht erkannt. Diese Ergebnisse haben wichtige Implikationen, da solche Beurteilungsverzerrungen dazu beitragen können, dass sich in der Konsequenz die Unterschiede zwischen leistungsstarken und -schwachen Schüler*innen verstärken.

Beim *Halo-Effekt* werden basierend auf (äußeren) Hinweisreizen (z. B. Sprachherkunft oder Geschlecht) Rückschlüsse auf „[…] globale Merkmale der [Schüler*innen]persönlichkeit geschlossen“ (Helmke, 2017, S. 136). So gibt es Studien, die auf eine geschlechtsspezifische Voreingenommenheit von Lehrkräften hinweisen, die sich im Fach Mathematik zum Nachteil von Mädchen auswirken kann (vgl. Kapitel 2.1).

Solche negativen Stereotype zu Ungunsten von *Schülerinnen* könnten unter anderem dazu führen, dass *Schüler* von Lehrkräften als mathematisch begabter eingeschätzt werden (Holder & Kessels, 2017; Lorenz, 2011; Mack et al., 2023; Robinson-Cimpian et al., 2014) ohne, dass entsprechende Unterschiede in den tatsächlichen Leistungen der Schüler*innen vorliegen (Lorenz 2011; Mack et al., 2023). Aktuelle Ergebnisse aus dem Hochschulbereich deuten ebenfalls daraufhin, dass geschlechtsspezifische Stereotype („Jungen sind gut in Mathematik, Mädchen sind gut im Lesen‘) auch bei angehenden Lehrkräften immer noch verbreitet sind (Klapproth & von der Lippe, 2024). Robinson-Cimpian et al. (2014) stellten fest, dass die Einschätzungen der Lehrkräfte langfristig betrachtet dazu führen können, dass sich die geschlechterbezogenen Unterschiede auch in der tatsächlichen Leistung widerspiegeln. Auch Muntoni et al. (2020) zeigten, dass die geschlechterspezifischen Erwartungen von Grundschullehrkräften die Unterschiede in der mathematischen Leistung zwischen Jungen und Mädchen erklären konnten (für weitere Informationen s. Kapitel 2.1).

Somit wäre denkbar, dass Lehrkräfte die mathematischen Schwierigkeiten von Jungen eher übersehen, da sie diese mathematisch begabter einschätzen. Bei Mädchen mit mathematikbezogenen Schwierigkeiten könnten sich die Leistungsunterschiede durch die Erwartungen der Lehrkräfte sogar noch verstärken. Auch wenn einige Studien geschlechtsspezifische Urteilstendenzen identifizieren, besteht weiterer Forschungsbedarf, um eindeutige Aussagen treffen zu können. Denn es gibt auch Studien, die zeigen, dass Mädchen unabhängig von der tatsächlichen Mathematikleistung eine bessere Mathematiknote erhalten als Jungen (Kaiser et al., 2015). Auch Wagner (2024) stellt in einer Literaturübersicht dar, dass die Studienlage insgesamt keine einheitlichen Ergebnisse hinsichtlich eines Zusammenhangs zwischen Geschlecht und der diagnostischen Einschätzung von Lehrkräften zulässt.

Test- und Urteilsmerkmale

Ein weiterer Teil der Variabilität der Urteilsgenauigkeit ist auf Unterschiede in der Methodik zurückzuführen. Dies bezieht sich vor allem auf die Informiert- und Differenziertheit des zu treffenden Urteils. Die Literatur unterscheidet zwischen informierten und uninformierten Urteilen. Bei der *uninformierten Beurteilung* (Südkamp et al., 2012) gibt die Lehrkraft in der Regel ein eher globales Urteil ab (z. B. eine Gesamteinschätzung über die allgemeine Mathematikleistung; Praetorius et al., 2012). Bei *informierten Urteilen* ist der Lehrkraft der Vergleichsmaßstab bekannt (z. B. Schulleistungstest; Praetorius & Südkamp, 2017). Informierte Einschätzungen sind tendenziell präziser als uninformierte (z. B. Demaray & Elliott, 1998; Feinberg & Shapiro, 2009), da die einschätzende Person Informationen darüber erhält, welche konkrete Fähigkeit eingeschätzt werden soll. Dies bestätigt sich auch in der Meta-Analyse von Südkamp et al. (2012).

Weiter wird zwischen direkten und indirekten Urteilen unterschieden. Bei *direkten Urteilen* wird die Leistung der Schüler*innen im Hinblick auf bestimmte Fähigkeiten oder Aufgaben eingeschätzt (z. B. Feinberg & Shapiro, 2009). So werden den Lehrkräften beispielsweise dieselben Aufgaben oder Items vorgelegt, die auch die Schüler*innen lösen beziehungsweise beantworten müssen (Praetorius et al., 2012, S. 118). Dabei schätzen die Lehrkräfte ein, wie viele Aufgaben von den einzelnen Schüler*innen jeweils korrekt gelöst werden (Praetorius & Südkamp, 2017). *Bei indirekten Urteilen* wird ein weniger spezifisches Urteil abgegeben, beispielsweise auf einer mehrstufigen Ratingskala (Praetorius & Südkamp, 2017). Die Übersichtsarbeit von Hoge und Coladarci (1989) zeigte, dass die Range der Korrelationen für die indirekten Einschätzungen zwischen $r = .28$ und $r = .86$ (mit einem Median von $r = .62$) variierte, während die direkten Beurteilungen in einen Bereich von $r = .48$ bis $r = .92$, mit einem Median von r

= .69 lagen. Die Meta-Analyse von Südkamp et al. (2012) fand keinen Effekt im Hinblick auf die Spezifität des Urteils.

Zwischenfazit

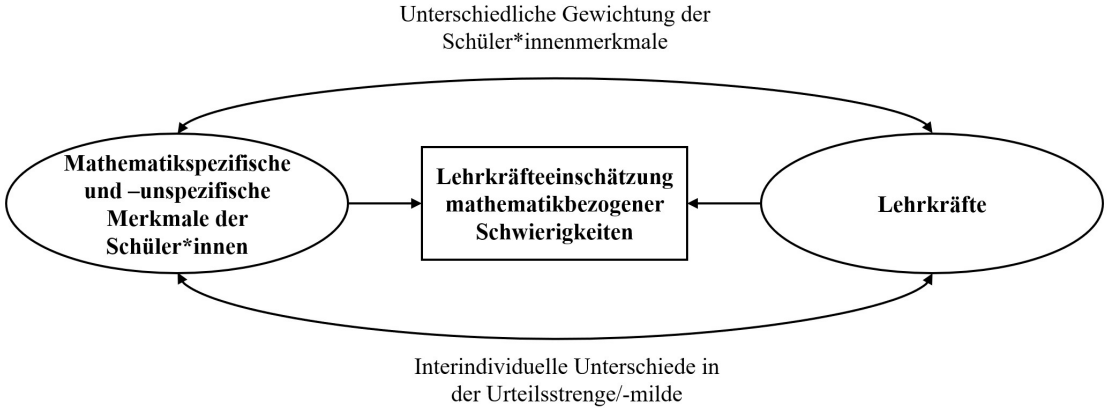
Die Forschungsergebnisse, die entlang des heuristischen Modells von Südkamp et al. (2012) verortet wurden, zeigen, dass die Einschätzungen der Lehrkräfte aus einer Kombination von urteilsrelevanten und -irrelevanten Informationen bestehen, die nicht unbedingt in direktem Zusammenhang mit der Leistung der Schüler*innen stehen müssen. Dies kann zu Fehleinschätzungen führen. Besonders problematisch ist dies, wenn Schüler*innen, die Lernschwierigkeiten haben, dadurch übersehen werden oder wenn aus inadäquaten Einschätzungen negative Erwartungen an die Leistungsentwicklung der Schüler*innen resultieren. Solche Fehleinschätzungen können im Sinne einer sich selbst erfüllenden Prophezeiung die tatsächliche Leistung der Schüler*innen negativ beeinflussen (z. B. de Boer et al., 2010; Jussim & Eccles, 1992; Jussim & Harber, 2005; Olczyk et al., 2023).

Bislang fehlen Meta-Analysen, die die empirische Evidenz des heuristischen Modells von Südkamp et al. (2012) umfassend bestätigen. Kaufmann (2020) konnte beispielsweise keine Evidenz dafür liefern, dass die Beurteilungsgenauigkeit der Lehrkräfte von den Lernschwierigkeiten der Schüler*innen oder der Klassenstufe abhängt. Wagner (2024) kommt in einem systematischen Literaturüberblick zu dem Schluss, dass das Urteil der Lehrkräfte von einer Reihe verschiedener Faktoren beeinflusst wird, die Evidenz jedoch überwiegend heterogen ist. Forschungsarbeiten, die untersuchen, welche Faktoren das Urteil von Lehrkräften bei der Einschätzung mathematikbezogener Schwierigkeiten beeinflussen, sind bislang nur begrenzt vorhanden.

Werden die skizzierten Studienergebnisse auf die diagnostische Einschätzung von Lehrkräften hinsichtlich mathematischer Schwierigkeiten übertragen, ist anzunehmen, dass diese Beurteilung ebenfalls sowohl auf urteilsrelevanten (mathematikspezifischen) als auch auf urteilsirrelevanten (mathematikunspezifischen) Schüler*innenmerkmalen basiert. Darüber hinaus erscheint es denkbar, dass sich Lehrkräfte hinsichtlich ihrer diagnostischen Urteilsstrenge beziehungsweise -milde unterscheiden, sodass objektiv identische Schüler*innenleistungen je nach Lehrkraft unterschiedlich bewertet werden. Zudem könnten die Schüler*innenmerkmale mit unterschiedlicher Gewichtung in das Urteil der Lehrkräfte einfließen (s. Abbildung 10).

Abbildung 10

Potenzielle Einflussfaktoren auf die Lehrkräfteeinschätzung bei der Beurteilung mathematikbezogener Schwierigkeiten



Die potenziellen Einflussfaktoren auf die Lehrkräfteeinschätzung bei der Beurteilung mathematikbezogener Schwierigkeiten von Grundschulkindern wurden daher in *Studie III* (Lamb et al., 2025) untersucht. Dazu wurden in Anlehnung an die Empfehlungen von Karst et al. (2017) Mehrebenenmodelle spezifiziert.

4.3 Studie III: Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule

Studie III (Lamb et al., 2025) untersuchte basierend auf den Daten von N = 377 Grundschüler*innen der Klassenstufe zwei bis vier und N = 33 Lehrkräften (a), ob das Lehrkräfteurteil über Schwierigkeiten in den mathematischen Fertigkeiten ihrer Schüler*innen vorrangig von deren tatsächlichen mathematikspezifischen Leistungen (Basisnumerik und Rechenfertigkeiten) abhängt, oder ob auch mathematikunspezifische Schüler*innenmerkmale wie die Intelligenz, Lesefertigkeit und das Geschlecht[1] in das Urteil einfließen. Zudem wurde geprüft (b), ob das Ausmaß, in dem die Schüler*innenmerkmale herangezogen werden, lehrkräfteabhängig ist und ob die Urteilsstrenge beziehungsweise -milde zwischen den Lehrkräften variiert.

Die Ergebnisse der Multilevel-Analysen zeigten (a), dass die Lehrkräfte die Schwierigkeiten ihrer Schüler*innen in den *basisnumerischen Fertigkeiten* basierend auf deren tatsächlichen mathematikspezifischen Leistungen einschätzten. Mathematikunspezifische Merkmale (z. B. das Geschlecht oder die Intelligenz der Schüler*innen) flossen nicht in das Urteil der Lehrkräfte ein. Bei der Beurteilung von Schwierigkeiten in den *Rechenfertigkeiten* bezogen die Lehrkräfte neben der mathematikspezifischen Leistungen auch die Intelligenz und die Lesefertigkeit ihrer Schüler*innen mit in ihre Einschätzung ein. Ergänzend wurde geprüft, ob die Klassenstufe einen Effekt hat. Nach Einschätzung der Lehrkräfte hatten Kinder der dritten und vierten Klassenstufe weniger Schwierigkeiten als Zweitklässler*innen. Ein tendenziell ähnlicher Effekt war auch zwischen den Schüler*innen der dritten und vierten Klassenstufe zu beobachten.

[1] In dieser Studie wurden die geschlechtsspezifischen Effekte auf männliche und weibliche Personen beschränkt, da die Anzahl der Personen ohne Geschlechtszugehörigkeit und diverser Personen zu gering für statistische Analysen ist. Weniger als 0.01 % der Allgemeinbevölkerung ordnen sich offiziell dem Geschlecht divers oder der Kategorie ‚keine Angabe' zu (Statistisches Bundesamt, 2022).

(b) Darüber hinaus gab es Hinweise auf den Strenge- beziehungsweise Milde-Effekt. Bei gleicher objektiver Testleistung in den mathematikspezifischen und -unspezifischen Fertigkeiten gaben einige Lehrkräfte Schwierigkeiten in den *Rechenfertigkeiten* an, während andere Lehrkräfte keine Schwierigkeiten sahen. Dies deutet darauf hin, dass die gleiche Schüler*innenleistung von verschiedenen Lehrkräften unterschiedlich bewertet wird. Es gab keinen Hinweis darauf, dass das Ausmaß, in dem Schüler*innenmerkmale in das Urteil einfließen, lehrkräfteabhängig ist. Für die Einschätzung von Schwierigkeiten in der *Basisnumerik* wurden keine Unterschiede zwischen den Lehrkräften festgestellt. Damit sprechen die Ergebnisse insgesamt dafür, dass die diagnostische Einschätzung der Lehrkräfte über die mathematischen Schwierigkeiten ihrer Schüler*innen in erster Linie von deren tatsächlichen mathematischen Leistungen bestimmt wird (Lamb et al., 2025). Der vollständige Artikel ist dem Anhang (C) zu entnehmen.

5. Diskussion

Werden Rechenstörungen nicht frühzeitig erkannt und entsprechende Interventionen daher nicht eingeleitet, steigt das Risiko für einen persistierenden Entwicklungsverlauf mit erheblichen Einschränkungen für die psychische Gesundheit und gesellschaftliche Teilhabe (z. B. Esser, 1992; Kohn et al., 2013; Parsons & Bynner, 2005; Saga et al., 2022; Schulz et al., 2018; vgl. Kapitel 1). Diese potenziellen Langzeitfolgen unterstreichen die Notwendigkeit, Kinder mit Anzeichen einer Rechenstörung frühzeitig zu identifizieren. Die Hauptverantwortung für diese Aufgabe liegt in erster Linie bei den Bildungsinstitutionen, insbesondere der Schule und damit bei den Lehrkräften (Hesse & Latzko, 2017; Kuhn, 2017; vgl. Kapitel 3.1). Daher untersucht die vorliegende Dissertation, *wie Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren können.* Diese Ausgangsfrage wurde in drei Studien sukzessive beforscht (s. Tabelle 1). Im Folgenden werden inhaltliche Bezüge zwischen den Teilstudien herausgestellt, um die zentralen Ergebnisse vor dem Hintergrund der übergeordneten Fragestellung zu diskutieren und praxis- sowie forschungsrelevante Implikationen abzuleiten.

Tabelle 1

Übersicht über die drei promotionsrelevanten Studien

Studie	Titel	Fragestellung	Methodik	Kernergebnisse
Studie I: Lamb et al., 2024a (Anhang A)	Delayed development of basic numerical skills in children with developmental dyscalculia	(a) Deuten die Defizite in den basisnumerischen Fertigkeiten von Grundschulkindern mit Rechenstörung auf eine Entwicklungsverzögerung oder auf eine spezifische qualitative Abweichung hin? (b) Sind diese Defizite auf eine (selektive) Beeinträchtigung in einem der zentralen Kernsysteme der Zahlenverarbeitung zurückzuführen?	Multilevel-Analysen	(a) Die Beeinträchtigungen in der Basisnumerik deuten auf eine Entwicklungsverzögerung hin. (b) Die Defizite resultieren eher aus einer Beeinträchtigung im ANS und sprechen weniger für das AD. Es gab keine Hinweise auf ein beeinträchtigtes OTS.
Studie II: Lamb et al., 2024b (Anhang B)	Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule	(a) Lässt sich die theoriebasierte zweifaktorielle Struktur (Basisnumerik und Rechenfertigkeiten) des Fragebogens empirisch bestätigen? (b) Entspricht das psychometrische Messmodell auf Itemebene einem einparametrischen oder zweiparametrischen logistischen Testmodell? (c) Wie zuverlässig werden Kinder mit und ohne Anzeichen für eine Rechenstörung durch den Fragebogen identifiziert?	Konfirmatorische Faktorenanalyse Item Response Theorie Receiver-Operating-Characteristic-Curve-Analysen	(a) Die zweidimensionale Struktur des Fragebogens wurde empirisch bestätigt. (b) Der Fragebogen kann Rasch-konform ausgewertet werden. (c) Der Fragebogen verfügt über gute bis überwiegend zufriedenstellende Screeningeigenschaften und identifiziert Kinder mit und ohne Anzeichen für eine Rechenstörung ökonomisch, valide und reliabel.
Studie III: Lamb et al., 2025 (Anhang C)	Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule	(a) Inwieweit berücksichtigen Lehrkräfte mathematikspezifische und –unspezifische Schüler*innenmerkmale bei der Einschätzung mathematischer Schwierigkeiten ihrer Schüler*innen? (b) Variiert die Urteilsstrenge beziehungsweise –milde zwischen den Lehrkräften und ist das Ausmaß, in dem mathematikspezifische und –unspezifische Schüler*innenmerkmale herangezogen werden lehrkräfteabhängig?	Multilevel-Analysen	(a) Die Lehrkräfteeinschätzung über die mathematischen Schwierigkeiten ihrer Schüler*innen wurde vorrangig von deren tatsächlichen mathematischen Leistungen bestimmt. (b) Variationen zwischen den Lehrkräften deuten auf interindividuelle Unterschiede in der Urteilsstrenge hin; die Gewichtung verschiedener Schüler*innenmerkmale variierte nicht.

5.1 Diagnostische Kriterien und Messinstrumente

Subgruppen

Die Forschung zeigt, dass die Basisnumerik von Kindern mit Rechenstörungen beeinträchtigt ist (vgl. Kapitel 2.1). Dies wurde auch in *Studie I* (Lamb et al., 2024a) bestätigt. Daher ist es wichtig, dass Lehrkräfte bei der Identifikation von Kindern mit Anzeichen einer Rechenstörung nicht nur auf die schulische Leistung der Kinder (z. B. Rechenfertigkeiten), sondern ebenso auf ihre Schwierigkeiten in den basisnumerischen Fertigkeiten achten (z. B. Fischer et al., 2015; vgl. Kapitel 2.4). Ein geeignetes deutschsprachiges fragebogengestütztes Screeninginstrument, das die Lehrkräfte dabei präzise und ökonomisch unterstützt, konnte nicht identifiziert werden (vgl. Kapitel 3.1). Daher wurde in *Studie II* (Lamb et al., 2024b) ein theoriegeleitet entwickelter Lehrkräftefragebogen (FERMAT) vorgestellt und hinsichtlich seiner Güte- und Screeningeigenschaften untersucht.

Der RATZ-Index stellt ein Maß für die Leistungsfähigkeit eines Screenings im Vergleich zur Zufallsanordnung dar und ist geeignet, um verschiedene Screeningverfahren hinsichtlich ihrer Vorhersagegenauigkeit miteinander zu vergleichen (Tröster, 2009, S. 143). Nach Jansen et al. (1999) gelten Werte ab .34 als gut, jedoch noch unspezifisch, während Werte ab .66 als sehr gut einzustufen sind. Die RATZ-Indizes des FERMAT variieren in Abhängigkeit von Subskala (Basisnumerik und Rechenfertigkeiten) und gewähltem Cut-Off-Wert (PR ≤ 7; PR ≤ 10) zwischen .445 und .795. Damit erreicht der FERMAT gemäß der Klassifikation von Jansen et al. (1999) gute bis sehr gute diagnostische Kennwerte. Auch im Vergleich zu anderen Screeninginstrumenten schneidet der FERMAT gut ab: So berichten Roick und Hasselhorn (2005) für den KR 3 – 4 je nach Subskala (Faktenabruf und Kettenaufgaben) RATZ-Indizes zwischen .31 und .64 (PR < 10) beziehungsweise zwischen .35 und .53 (PR < 25).

Die *Positive und Negative Korrektheit* stellen keine Eigenschaften des Tests dar, sondern kennzeichnen die Sicherheit eines positiven beziehungsweise negativen Screeningbefundes in einer spezifischen Population (Tröster, 2009, S. 129f.). Die Untersuchung der Screeningeigenschaften des FERMAT ergab, dass maximal 52 % der Kinder, die im Teilbereich „FERMAT Basisnumerische Fertigkeiten“ durch die Lehrkräfte als *gefährdet* identifiziert wurden, tatsächlich substanzielle Defizite in der Basisnumerik aufweisen (CODY-M 2 – 4; PR ≤ 7). Durch die Subskala „FERMAT Rechenfertigkeiten“ wurden maximal 41 % der Kinder mit Defiziten in den Rechenfertigkeiten (Heidelberger Rechentest [HRT 1 – 4]; Haffner et al., 2005; PR ≤ 10) korrekt klassifiziert

Auf den ersten Blick erscheinen die Werte gering. Allerdings schwanken der positive prädiktive Wert und die Anzahl falsch-positiver Klassifikationen je nach Prävalenz (WHO, Regionalbüro für Europa, 2020). In *Studie II* (Lamb et al., 2024b) wurden konservative Cut-Off-Werte (PR ≤ 7 und PR ≤ 10) zur Differenzierung zwischen Kindern mit und ohne Anzeichen einer Rechenstörung verwendet. Diese Cut-Off-Werte wirken sich auf die Grundquote (Anteil der Kinder mit testdiagnostischen Anzeichen für eine Rechenstörung) und damit auf die positive und negative Korrektheit aus: „Je höher die Grundquote ist, desto größer ist auch die Wahrscheinlichkeit, dass ein positiver [Screeningbefund] zutrifft“ (Tröster, 2009, S. 130). Ein konservativer Grenzwert führt somit zu einer Reduktion der Anzahl von Kindern mit bestätigten Minderleistungen.

Dies wiederum bedeutet, dass eine Reihe von Kindern, die über den FERMAT falsch-positiv klassifiziert wurden, möglicherweise testdiagnostische Schwierigkeiten im Rechnen (z. B. 10 < PR ≤ 25) aufweisen, diese aber aufgrund der strengen Kriterien nicht erkannt wurden. Denn in *Studie II* (Lamb et al., 2024b) wurden Kinder, deren Leistung einem PR > 10 entsprach, als unbeeinträchtigt eingestuft. Dies könnte erklären, weshalb über den FERMAT deutlich mehr Kinder falsch-positiv (z. B. 99 Kinder) als falsch-

negativ klassifiziert (z. B. 4) wurden. Anders formuliert bedeutet dies, dass die Lehrkräfte (mithilfe des FERMAT) möglicherweise bereits leichtere mathematikbezogene Schwierigkeiten erkannt haben, diese Kinder aber aufgrund der konservativen Cut-Off-Werte als falsch-positiv klassifiziert wurden.

Vor dem Hintergrund inklusiver Schul- und Unterrichtsprozesse sollten alle leistungsschwächeren Schüler*innen frühzeitig identifiziert und in ihren mathematikspezifischen Fertigkeiten gefördert werden und nicht nur diejenigen, die substanzielle Schwierigkeiten im Sinne der dritten Förderstufe des RTI-Modells aufweisen (z. B. Blumenthal et al., 2014; vgl. Kapitel 3). Gersten et al. (2005) sind beispielsweise der Ansicht, dass zu Screeningzwecken ein PR < 35 verwendet werden könnte, um möglichst alle Kinder zu identifizieren. In einer zukünftigen Studie sollte daher untersucht werden, wie zuverlässig der FERMAT Kinder mit leichteren mathematikbezogenen Schwierigkeiten identifiziert, die im RTI-Modell eher auf der zweiten Stufe zu verorten sind (z. B. Blumenthal et al., 2014; vgl. Kapitel 3).

Gemäß der im vorherigen Abschnitt skizzierten Argumentation wäre zu erwarten, dass liberalere Cut-Off-Kriterien (z. B. PR ≤ 25) in verbesserten Screeningeigenschaften münden. Es ist aber ebenfalls wichtig zu prüfen, inwieweit der FERMAT inhaltlich die Schwierigkeiten von Kindern mit leichteren Rechenschwierigkeiten abbildet. Denn für diese Teilgruppe ist ein anderes Ursachen- und Defizitprofil zu erwarten als für Kinder mit schwereren Defiziten (Gallit et al., 2017; vgl. Kapitel 2.1). So sind beispielsweise die basisnumerischen Fertigkeiten in dieser Subgruppe nicht zwingend beeinträchtigt. Busch et al. (2018) zeigten, dass Kinder mit leichteren Rechenschwierigkeiten (35 < T-Wert < 40) im Vergleich zu nicht beeinträchtigten Kindern keine basisnumerischen Defizite aufwiesen, Kinder mit stärkeren Beeinträchtigungen (T-Wert ≤ 35) hingegen schon.

Eine zukünftige Studie, die die Ursachen- und Defizitprofile von Kindern mit schwereren (PR ≤ 10) und leichteren Beeinträchtigungen im Rechnen (z. B. 10 < PR ≤

25) untersucht, sollte aus methodischer Perspektive zudem ergänzend *Bottom-up-Ansätze* verwenden (z. B. Bartelet et al., 2014; Kißler et al., 2021; Salvador et al., 2019). Denn Kißler et al. (2021) stellten mittels einer *Mixture-Model-Analyse* fest, dass selbst die stärker beeinträchtigte Subgruppe (PR ≤ 10) kein homogenes Defizitprofil aufweist. Vielmehr existieren innerhalb dieser Gruppe zwei unterschiedliche Subtypen, die sich in der Ausprägung ihrer mathematischen Defizite unterscheiden. Bartelet et al. (2014) identifizierten sogar sechs verschiedene Subtypen. Ähnliche Ergebnisse liegen für Kinder mit weniger ausgeprägten Rechenschwierigkeiten vor: Salvador et al. (2019) identifizierten mithilfe einer *Clusteranalyse* einen Subtyp mit domänenspezifischen Kerndefiziten in der symbolischen und non-symbolischen Mengenverarbeitung und einen weiteren Subtyp mit domänengenerellen Beeinträchtigungen. Daher sollten zukünftig auch domänengenerelle Fertigkeiten (z. B. das visuell-räumliche Arbeitsgedächtnis) berücksichtigt werden, da diese bei Kindern mit Rechenstörung beeinträchtigt (vgl. Kapitel 2.2) und für die Unterscheidung zwischen Subtypen entscheidend sein können (z. B. Busch et al., 2018; Salvador et al., 2019). Möglicherweise lassen sich über Bottom-up-Ansätze (z. B Kißler et al., 2021; Salvador et al., 2019) nicht nur verschiedene Subtypen mit unterschiedlich ausgeprägten Defiziten in den Kernsystemen der Zahlenverarbeitung identifizieren (ähnlich zu Bartelet et al., 2014), sondern auch einerseits ein *entwicklungsverzögerter* und andererseits ein *qualitativ beeinträchtigter* Subtyp.

Darüber hinaus ist es sinnvoll, zukünftig klassische frequentistische Analysen durch bayesianische Methoden zu ergänzen (z. B. Decarli et al., 2023; Kißler et al., 2020; Mammarella et al., 2021). Denn die bayesianische Statistik ermöglicht im Vergleich zu frequentistischen Ansätzen präzisere Aussagen über *Nulleffekte*. Denn „[…] die Abwesenheit einer Signifikanz [stellt] keine Evidenz für die Nicht-Existenz eines Effektes dar“ (Kißler et al., 2020, S. 90). Werden die Ergebnisse aus *Studie I* (Lamb et

al., 2024a) aus dieser Perspektive betrachtet, kann aus der fehlenden statistischen Signifikanz für die Simultanerfassung in der Aufgabe „Punkte zählen“ (CODY-M 2 – 4; Kuhn et al., 2017) nicht per se geschlossen werden, dass das OTS der Kinder mit einer Rechenstörung unbeeinträchtigt ist. Der Bayes-Faktor (BF_{10}) gibt das Verhältnis der Evidenz für das Alternativmodell (H_1) im Vergleich zum Nullmodell (H_0) an. Ein $BF_{10} > 3$ bedeutet, dass die Daten H_1 stärker unterstützen als H_0, während ein $BF_{10} < 0.33$ eher für H_0 spricht. Werte zwischen 0.33 und 3 sprechen eher für unbestimmte Evidenz (Wagenmakers et al., 2018). Die bayesianische Statistik ermöglicht somit eine präzisere Einschätzung darüber, welche der drei Defizithypothesen am besten mit den Daten übereinstimmt. Zudem kann die bayesianische Statistik kann die Robustheit der frequentistischen Ergebnisse entweder bestätigen oder ergänzen, wenn bayesianische und frequentistische Analysen zu unterschiedlichen Ergebnissen führen (z. B. Kißler et al., 2020). Dies ist auch mit Blick auf die Ergebnisse in der Simultanerfassung zentral, denn deskriptiv fielen die Reaktionszeiten der Kinder mit einer Rechenstörung im Mittel höher aus als die Reaktionszeiten unbeeinträchtigter Kinder (Lamb et al., 2024a).

Klassenstufenspezifische Betrachtung

In *Studie II* (Lamb et al., 2024b) erreichten der FERMAT Gesamtscore beziehungsweise die Subskalen „Basisnumerische Fertigkeiten“ und „Rechenfertigkeiten“, je nach Kriterium, das zur Klassifikation des Risikos einer Rechenstörung (HRT 1 – 4, CODY-M 2 – 4; PR ≤ 7, PR ≤ 10) herangezogen wurde, ihre maximale Trennschärfe bei Cut-Off-Werten von ein bis vier Testpunkten. Auf dieser Grundlage wurde empfohlen, dass Kinder mit Schwierigkeiten in zwei Bereichen des FERMAT einer umfassenderen Diagnostik zugeteilt werden sollten (Lamb et al., 2024b). Hier besteht weiteres Optimierungspotenzial, um die diagnostische Genauigkeit gezielt zu erhöhen.

Zu diesem Zweck ist es sinnvoll zu untersuchen, ob klassenstufenspezifische Cut-Off-Werte für den FERMAT formuliert werden sollten (Lamb et al., 2024b). Denn *Studie I* (Lamb et al., 2024a) zeigte, dass die Defizite in der Basisnumerik auf eine Entwicklungsverzögerung hindeuten. So gab es beispielsweise keinen Hinweis darauf, dass Viertklässler*innen mit Rechenstörung Zahlen auf dem Zahlenstrahl (Zahlenraum bis 100) weniger präzise verorten als Kinder ohne Rechenstörung. Allgemein waren die Defizite in der Basisnumerik bei Kindern der zweiten Klassenstufe besonders ausgeprägt. Auch die Lehrkräfte identifizierten bei Kindern der dritten und vierten Klassenstufe weniger Schwierigkeiten als bei den Kindern der zweiten Klassenstufe. Ein tendenziell ähnlicher Effekt war auch zwischen den Schüler*innen der dritten und vierten Klassenstufe zu beobachten (Lamb et al., 2025).

Daher wäre es denkbar, dass der FERMAT je nach Klassenstufe seine maximale Trennschärfe bei einer unterschiedlichen Anzahl von Testpunkten erreicht. Durch klassenstufenspezifische Analysen könnten somit möglicherweise Informationen generiert werden, die eine noch spezifischere Einschätzung ermöglichen. So könnte beispielsweise für Zweitklässler*innen, die in vier Bereichen des FERMAT Auffälligkeiten zeigen, eine weiterführende Diagnostik empfohlen werden, wohingegen bei Viertklässler*innen bereits bei Defiziten in einem oder zwei Fertigkeiten zu einer standardisierten Testung geraten werden könnte.

Da zur weiteren diagnostischen Abklärung eine große Auswahl potenzieller Testverfahren zur Verfügung steht (für eine Übersicht s. S3-Leitlinie; Schulte-Körne & Haberstroh, 2018), ist es wichtig, Lehrkräfte bei der Auswahl geeigneter diagnostischer Instrumente zu unterstützen. Eine Hilfestellung bietet das webbasierte Tool des *LONDI-Hilfssystems* (Schulte-Körne & Hasselhorn, 2025). Das Tool stellt evidenzbasierte, gütekonforme Verfahren für verschiedene Klassenstufen niedrigschwellig zusammen, sodass die Auswahl für die Lehrkräfte vereinfacht wird. Um die diagnostischen Prozesse

von der Früherkennung über die Diagnostik bis hin zur Förderung (vgl. Kapitel 3) zu optimieren, könnte es sinnvoll sein, auf Plattformen wie dem LONDI-Hilfssystem auch fragebogengestützte Screeninginstrumente (z. B. den FERMAT) aufzulisten. Zudem sollten (angehende) Lehrkräfte für die Schwächen nicht gütekonformer diagnostischer Verfahren sensibilisiert werden. Denn bei der Mehrzahl der publizierten Symptom-Checklisten und Fragebögen handelt es sich um informelle Verfahren, für die keine Güteangaben vorliegen (vgl. Kapitel 3.1).

Messinstrumente

Auch die als Kriterium herangezogenen Testinstrumente (CODY-M 2 – 4; Kuhn et al., 2017 und HRT 1 – 4; Haffner et al., 2005) beeinflussen das Ergebnis der kriterialen Validitätsprüfung (Lamb et al., 2024b). Obwohl die im FERMAT erfassten Fertigkeiten inhaltlich weitgehend mit den Testverfahren übereinstimmen, die zur Bestimmung der objektiven Testleistung der Kinder verwendet wurden, gibt es auch Unterschiede zwischen den Verfahren (Lamb et al., 2024b). So erfasst der CODY-M 2 – 4 (Kuhn et al., 2017) teils basalere basisnumerische Fertigkeiten (z. B. Subitizing oder einstelliger Zahlenvergleich) als die Subskala „FERMAT Basisnumerische Fertigkeiten“ (Lamb et al., 2024b). Auch die Subskala „FERMAT Rechenfertigkeiten“ beinhaltet zum Teil Bereiche (z. B. Text- oder Sachaufgaben), die nicht durch die ausgewählten Subskalen des HRT 1 – 4 (Haffner et al., 2005) erfasst wurden. Es ist daher zu erwarten, dass die psychometrischen Güteeigenschaften des FERMAT (Lamb et al., 2024b) verbessert werden können, wenn zur Bestimmung der kriterialen Validität weitere Testverfahren herangezogen werden, die eine noch höhere inhaltliche Übereinstimmung mit dem FERMAT aufweisen.

Zukünftig könnte es potenziell aufschlussreich sein, zu untersuchen, ob das von Richter et al. (2018) geplante Online-Screening zur Identifikation von Grundschulkindern

mit Anzeichen einer Lernstörung zu ähnlichen Ergebnissen führt wie die Lehrkräfteeinschätzung im FERMAT. Das geplante Screening soll einheitlich skalierte Aufgaben für die erste bis vierte Klassenstufe beinhalten und inhaltlich ähnliche Bereiche abdecken wie der FERMAT, darunter Numerositäten (z. B. Zahlenstrahlaufgaben), Faktenabruf und Kopfrechnen (z. B. Grundrechenarten für verschiedene Zahlenräume) sowie mathematisches Problemlösen (z. B. Textaufgaben mit Additions- und Subtraktionsaufgaben).

Ähnliche Effekte sind für *Studie III* (Lamb et al., 2025) zu erwarten. Würden Testinstrumente verwendet, die exakt die im FERMAT zu beurteilenden mathematischen Fertigkeiten erfassen, könnten die mathematikspezifischen Schüler*innenmerkmale, die in das Urteil der Lehrkräfte einfließen, eine noch größere Bedeutung erlangen. Denn wenn Lehrkräfte die Leistung von Schüler*innen im Hinblick auf eine konkrete Aufgabe einschätzen (z. B. Feinberg & Shapiro, 2009), die die Schüler*innen anschließend bearbeiten, ist ihre Einschätzung präziser (z. B. Hoge & Coladarci, 1989).

Auch in *Studie I* (Lamb et al., 2024a) wurde die Basisnumerik über den CODY-M 2 – 4 (Kuhn et al., 2017) erfasst, da der Test die klassischen Aufgabenparadigmen (z. B. symbolischer Mengenvergleich, Zahlenstrahlaufgaben) beinhaltet (vgl. Kapitel 2.2). Diese werden in der Forschung üblicherweise verwendet, um Rückschlüsse auf die Kernmechanismen der Zahlenverarbeitung zu ziehen (vgl. Kapitel 2.2). Die Paradigmen werden jedoch je nach Studie unterschiedlich konzipiert und ausgewertet. Dies wirkt sich auf die Testergebnisse aus und könnte unter anderem erklären, warum einige Studien bestimmte Defizite bei Kindern mit Rechenstörung finden und andere nicht (vgl. Kapitel 2.2).

So zeigten beispielsweise in *Studie I* (Lamb et al., 2024a) Viertklässler*innen mit einer Rechenstörung keine Defizite in der Zahlenstrahlaufgabe. Landerl (2013) stellte dagegen fest, dass Kinder mit einer Rechenstörung die Zahlen auf dem Zahlenstrahl

weniger genau lokalisieren als Kinder ohne Rechenstörung in der vierten Klassenstufe. Allerdings wurde in *Studie I* (Lamb et al., 2024a) der Zahlenraum von 0 bis 100 betrachtet, wohingegen Landerl (2013) einen Zahlenstrahl mit den Endpunkten 0 und 1000 verwendete.

In *Studie I* (Lamb et al., 2024a) fanden sich – im Gegensatz zu früheren Untersuchungen (z. B. Kuhn et al., 2013) – keine signifikanten Hinweise auf eine Beeinträchtigung des Subitizing-Prozesses bei Kindern mit einer Rechenstörung. Die Ergebnisse deuten jedoch auf einen qualitativ abweichenden Zählprozess hin. Während in *Studie I* für jedes Kind individuell bestimmt wurde, ob der Zählprozess bei vier oder fünf Objekten beginnt, wurde der Subitizing-Bereich in anderen Studien (z. B. Kuhn et al., 2013) a priori auf vier Objekte festgelegt. Dies könnte zu einer Vermischung von Zähl- und Subitizing-Fähigkeiten geführt haben, was niedrigere Reaktionszeiten im Zählbereich und höhere im Simultanbereich erklären würde.

Eine weitere häufig untersuchte Aufgabe ist das non-symbolische Mengenvergleichsparadigma. Dabei besteht die Aufgabe darin, so schnell und so genau wie möglich zu entscheiden, welche von zwei nebeneinander visuell dargestellten Mengen quantitativ größer ist (vgl. Kapitel 2.1). Dabei können verschiedene visuelle Parameter kontrolliert werden: z. B. konvexe Hülle, Gesamtfläche, mittlerer Durchmesser und Gesamtumfang der dargestellten Objekte (z. B. Brambrink et al., 2024). In *Studie I* (Lamb et al., 2024a) wurden verschiedene Parameter kontrolliert. So konnten aus der Größe der dargestellten Punkte keine Rückschlüsse auf die Menge der dargestellten Punkte gezogen werden. Außerdem wurden unterschiedliche Ratios verwendet und die Anzahl der Punkte systematisch variiert. Die konvexe Hülle wurde nicht kontrolliert. Die Kontrolle dieser Stimuli kann sich auf die Ergebnisse auswirken (z. B. Gebuis & Reynvoet, 2011) und könnte ein Grund dafür sein, dass einige Studien ein Defizit in der

Verarbeitung non-symbolischer Mengen fanden (z. B. Decarli et al., 2020), andere jedoch nicht (z. B. Lamb et al., 2024a).

Variierende Kriterien

Diese Beispiele zeigen: Wenn in der Forschung von Rechenstörungen gesprochen wird, verbergen sich hinter diesem Konstrukt unterschiedliche Messinstrumente, Aufgabenbedingungen und Auswertungsmethoden. Darüber hinaus variieren die Kriterien, die zur Operationalisierung von Rechenstörungen herangezogen werden. Auch in der vorliegenden Dissertation wurden je nach Zielsetzung unterschiedliche Klassifikationskriterien beziehungsweise Ein- und Ausschlusskriterien verwendet.

Studie I (Lamb et al., 2024a) untersuchte die *domänenspezifischen* Kerndefizite und -ursachen der Rechenstörung. Dazu wurde die Rechenstörung anhand einer signifikanten Abweichung (T-Wert ≤ 38) im HRT 1 – 4 (Haffner et al., 2005) bei unbeeinträchtigter Intelligenz (IQ > 70) operationalisiert, um eine Gruppe mit primär kognitiven Defiziten zu erfassen (z. B. Mazzocco et al., 2011; vgl. Kapitel 2.1). Das doppelte Diskrepanzkriterium wurde nicht angewandt, da die Sinnhaftigkeit zunehmend hinterfragt wird (vgl. Kapitel 2.1). Die Leseflüssigkeit wurde zusätzlich zur Intelligenz als Kovariate, jedoch nicht als Ausschlusskriterium berücksichtigt, da Kinder mit Leseschwierigkeiten nur selektive Schwierigkeiten beim Bearbeiten numerischer Aufgaben haben, wenn diese die verbal-phonologische Zahlenverarbeitung betreffen (Raddatz et al., 2017).

Studie II (Lamb et al., 2024b) zielte darauf ab, Kinder mit Anzeichen einer Rechenstörung mittels eines Lehrkräftefragebogens zu identifizieren. Angesichts der starken Variabilität der diagnostischen Kriterien und Testverfahren, die zur Klassifikation von Rechenstörungen verwendet werden (vgl. Kapitel 2.1), wurden zwei unterschiedliche Cut-Off-Werte (PR ≤ 10 und PR ≤ 7) und zwei verschiedene Testverfahren mit

unterschiedlichen Schwerpunkten, Rechenfertigkeiten (HRT 1 – 4; Haffner et al., 2005) und basisnumerische Fertigkeiten (CODY-M 2 – 4; Kuhn et al., 2017) verwendet. Eine Festlegung von Ausschlusskriterien bezüglich Intelligenz oder Leseleistung wurde entsprechend dem Vorgehen anderer Autor*innen (z. B. Fischer et al., 2015) nicht vorgenommen. Denn ein Screening ist nur der erste Schritt eines diagnostischen Prozesses, der darauf abzielt, diejenigen Kinder herauszufiltern, die einer umfassenderen diagnostischen Untersuchung zugewiesen werden sollten (vgl. Kapitel 3).

In *Studie III* (Lamb et al., 2025) wurde untersucht welche Schüler*innenmerkmale das Lehrkräfteurteil im FERMAT beeinflussen. Da diese Studie den Fokus auf die Einflussfaktoren der diagnostischen Einschätzung und Variationen zwischen den Lehrkräften legte, wurden keine spezifischen Ein- und Ausschlusskriterien formuliert.

Dennoch sind diese Vorgehensweisen zu diskutieren: Bei ausreichend großen Stichprobengrößen hätten Kinder mit Schwierigkeiten im Lesen aus den statistischen Analysen ausgeschlossen werden können, um potenzielle Einflussfaktoren auf die Ergebnisse zu minimieren. Zudem wäre ein *Propensity Score Matching* (z. B. An & Bai, 2015) sinnvoll gewesen, um eine möglichst unverzerrte Vergleichsbasis zwischen Kindern mit und ohne Rechenstörung zu schaffen. Denn in *Studie I* (Lamb et al., 2024a) schnitten die Kinder mit einer Rechenstörung in einem allgemeinen Intelligenztest und in einem Lesetest schlechter ab als Kinder ohne Rechenstörung. Auch in der Stichprobe, die in den *Studien II* (Lamb et al., 2024b) *III* (Lamb et al., 2025) untersucht wurde, wiesen die Kinder mit testdiagnostischen Anzeichen einer Rechenstörung häufiger zusätzliche Leseschwierigkeiten auf als Kinder ohne eine Rechenstörung.

Die unterschiedliche Handhabung der Cut-Off-Kriterien in den drei promotionsrelevanten Studien verdeutlicht, dass verschiedene Forschungsinteressen, wie zum Beispiel die Sicherstellung klinisch valider Daten und die Gewährleistung ausreichender statistischer Power, in Einklang gebracht werden müssen. Im Hinblick auf

die jeweiligen Zielsetzungen der Studien kann die klinische und ökologische Validität der erhobenen Daten weitgehend als gegeben angesehen werden. Dennoch können die Ergebnisse nicht uneingeschränkt auf alle Kinder mit Anzeichen einer Rechenstörung generalisiert werden, da das Profil der Rechenstörung heterogen ist (z. B. Kißler et al., 2020). Zudem sollte kritisch reflektiert werden, dass die Festlegung von Cut-Off-Kriterien zwangsläufig zu ‚blinden Flecken' führt, da bestimmte Kinder, wie etwa solche mit leichteren Rechenschwierigkeiten, nicht berücksichtigt werden.

Für die klinische Diagnostik sind kategoriale Ansätze als Legitimationsgrundlage unverzichtbar. Dennoch sollte die Festlegung der Cut-Off-Kriterien, insbesondere in Bezug auf die Anwendung der doppelten Diskrepanz, hinterfragt werden (vgl. Kapitel 2.1). Die Forschung sollte kategoriale Gruppeneinteilungen durch dimensionale Ansätze ergänzen, die es ermöglichen, Kinder basierend auf ihren individuellen Fertigkeiten auf einem Kontinuum zu verorten (z. B. Dowker, 2024; Mammarella et al., 2021).

Angesichts der weit verbreiteten und heterogenen Anwendung unterschiedlicher Kriterien und Messinstrumente, die wahrscheinlich nicht aufgelöst werden kann, ist es für die pädagogische Praxis von entscheidender Bedeutung, (angehende) Lehrkräfte für die Auswirkungen unterschiedlicher Definitionen und Kriterien zu sensibilisieren. Denn die Variabilität und die damit in Zusammenhang stehende heterogene Evidenz (z. B. unterschiedliche Defizitprofile und Prävalenzen), könnte ein Grund dafür sein, dass bei angehenden Grundschullehrkräften erhebliche Unsicherheiten in Bezug auf Rechenstörungen bestehen. Beispielsweise konnten angehende Grundschullehrkräfte in einem Wissenstest zu Teilleistungsstörungen im Durchschnitt nur 34 % der Fragen korrekt beantworten (Lamb et al., 2023a). Bender et al. (2024) befragten angehende Grundschullehrkräfte und stellten fest, dass etwa 50 % der Befragten nicht wussten, wie häufig Rechenstörungen auftreten und nicht sicher sagen konnten, ob es geschlechtsspezifische Unterschiede in der Prävalenz gibt (Bender et al., 2024).

Da universitäre Lerngelegenheiten prädiktiv für das pädagogisch-psychologische Wissen angehender Lehrkräfte sind (Lamb et al., 2023b), sollte das Wissen über die Auswirkungen unterschiedlicher Definitionen und Kriterien in der universitären Lehramtsausbildung vermittelt werden. Hierzu könnte eine stärkere Verzahnung von psychologischen und mathematikdidaktischen Inhalten sinnvoll sein (z. B. Faix et al., 2023). Wichtig ist auch, dass klar definiert wird, welches Professionswissen Lehrkräfte in Bezug auf Rechenstörungen erwerben sollen. Denn dies ist bisher noch nicht klar definiert worden (Landesregierung Nordrhein-Westfalen, 2020).

5.2 Merkmale der Schüler*innen und Lehrkräfte

Mathematikspezifische und -unspezifische Einflussfaktoren

Studie III (Lamb et al., 2025) zeigte, dass die Lehrkräfte die Schwierigkeiten in den *basisnumerischen Fertigkeiten* ihrer Schüler*innen basierend auf deren mathematikspezifischen Leistungen beurteilen. Weder das Geschlecht der Schüler*innen noch die mathematikunspezifische Leistung – Intelligenz und Lesefertigkeit – beeinflussten die Einschätzung der Lehrkräfte im FERMAT. Auch bei der Beurteilung von *Schwierigkeiten in den Rechenfertigkeiten* stützten sich die Lehrkräfte vorrangig auf die mathematikspezifischen Leistungen ihrer Schüler*innen (Lamb et al., 2025). Damit basieren die Einschätzungen der Lehrkräfte, konform zu Kaiser et al. (2015), primär auf mathematikspezifischen Informationen.

Es ist denkbar, dass der FERMAT dazu beiträgt, dass sich die Lehrkräfte bei ihrer Einschätzung stärker auf die mathematikspezifischen Leistungen ihrer Schüler*innen und weniger auf mathematikunspezifische Merkmale fokussieren. Denn im FERMAT (Lamb et al., 2024b) werden alle einzuschätzenden Fertigkeiten beziehungsweise Schwierigkeiten anhand von Beispielen erläutert. Dies soll unter anderem die Objektivität erhöhen, was besonders im Bereich der Basisnumerik von Bedeutung ist. Denn der

Begriff *Basisnumerik* wird synonym zu anderen Begriffen wie *pränumerische Grundfertigkeiten*, *mathematische Vorläuferfertigkeiten* und *mathematische Basiskompetenzen* verwendet und unterschiedlich definiert. Während sich die Psychologie – in Übereinstimmung mit der vorliegenden Dissertation – bei der Betrachtung basisnumerischer Fertigkeiten eher auf fundamentale Prozesse wie die Mengenverarbeitung fokussiert, umfasst der Begriff je nach Operationalisierung auch weiter gefasste mathematische Fertigkeiten. So sprechen beispielsweise Haffner et al. (2005) im HRT 1 – 4 von *mathematischen Basiskompetenzen*, fassen darunter aber auch die Beherrschung der Grundrechenarten (Addition, Subtraktion, Multiplikation und Division). Diese werden in der vorliegenden Dissertation, übereinstimmend mit dem CODY-M 2 – 4 Test (Kuhn et al., 2017), zu den Rechenfertigkeiten gezählt und von den basisnumerischen Fertigkeiten abgegrenzt.

Die Operationalisierung des Beurteilungsgegenstandes im FERMAT könnte somit zu einer stärkeren Annäherung des Konzeptverständnisses und damit zu konsistenteren Beurteilungen durch die Lehrkräfte führen. Diese Hypothese sollte in einer zukünftigen Untersuchung überprüft werden. Dazu könnten Lehrkräfte im ersten Schritt global einschätzen, welche Schüler*innen ihrer Klasse mathematikbezogene Schwierigkeiten haben (z. B. der*die Schüler*in hat Schwierigkeiten in der Basisnumerik: ja/nein; Schwierigkeiten in den Rechenfertigkeiten: ja/nein). In einem zweiten Schritt könnten die Lehrkräfte den FERMAT für alle Schüler*innen ausfüllen. Sollten Lehrkräfte bei der globalen Einschätzung mathematikunspezifische Schüler*innenmerkmale (z. B. Lesen, Intelligenz und Geschlecht) der Schüler*innen in die Beurteilung einbeziehen, bei der Anwendung des FERMAT jedoch nicht, könnte dies darauf hinweisen, dass der FERMAT die Fokussierung auf mathematikspezifische Merkmale unterstützt. Um die praktische Brauchbarkeit des FERMAT abschließend beurteilen zu können, sollte in diesem Zusammenhang auch untersucht werden, ob über

die Beurteilung im FERMAT mehr Kinder mit und ohne Anzeichen einer Rechenstörung korrekt identifiziert werden als über eine globale unstandardisierte Ja/Nein-Einschätzung.

Während die Einschätzung von Schwierigkeiten in der Basisnumerik nicht durch mathematikunspezifische Merkmale beeinflusst wurde, fielen die Ergebnisse für die Einschätzung von *Schwierigkeiten in den Rechenfertigkeiten* anders aus. Hier zeigte sich, dass die Lehrkräfte in die Einschätzung von Schwierigkeiten in den Rechenfertigkeiten auch mathematikunspezifische Schüler*innenmerkmale einbeziehen (Lamb et al., 2025). So flossen, ähnlich zu den Ergebnissen von Kaiser et al. (2015), die allgemeinen kognitiven Fähigkeiten (Intelligenz) und Lesefertigkeiten (Leseflüssigkeit) der Schüler*innen in die Beurteilung ein. Das Geschlecht der Schüler*innen beeinflusste die Einschätzung nicht.

Ein möglicher Grund hierfür könnte sein, dass die basisnumerischen Fertigkeiten im FERMAT (z. B. Zahlen ordnen und vergleichen) von den Lehrkräften als sprachunabhängiger und kognitiv weniger anspruchsvoll wahrgenommen wurden als die zu beurteilenden Rechenfertigkeiten (z. B. Textaufgaben oder Subtraktionsaufgaben). Dies wiederum würde bedeuten, dass die Lehrkräfte die Lesefertigkeiten und die allgemeinen kognitiven Fähigkeiten ihrer Schüler*innen bewusst in ihr Urteil einbeziehen (Lamb et al., 2025).

Fischer et al. (2015) fanden Hinweise darauf, dass Lehrkräfte bei der Beurteilung mathematischer Schwierigkeiten auch sprachliche Faktoren berücksichtigen. Bender et al. (2024) zeigten, dass sich knapp 16 % der befragten angehenden Grundschullehrkräfte nicht sicher sind, ob Rechenstörungen durch unterdurchschnittliche intellektuelle Fähigkeiten verursacht werden. Etwa 5 % waren sich sogar sicher, dass Kinder mit Rechenstörungen eine unterdurchschnittliche Intelligenz aufweisen.

Somit wäre denkbar, dass Lehrkräfte im FERMAT Schwierigkeiten markieren, diese aber nicht auf Defizite in den mathematischen Fertigkeiten zurückführen, sondern

auf sprachliche Barrieren, Leseschwierigkeiten oder geringere allgemeine kognitive Fähigkeiten. Zukünftige Studien sollten daher systematisch untersuchen, unter welchen Bedingungen Lehrkräfte die Lesefertigkeit und die Intelligenz ihrer Schüler*innen in ihre Beurteilungen einbeziehen und ob sie sich dessen bewusst sind. Für eine erste Exploration könnte anhand der vorliegenden Daten untersucht werden, ob Leseschwierigkeiten (z. B. PR $\leq$ 10) oder eine niedrigere Intelligenz (z. B. 70 < IQ < 85) systematische Fehleinschätzungen begünstigen oder ob Lehrkräfte diese mathematikunspezifischen Faktoren unabhängig von ihrer Ausprägung in ihr Urteil einfließen lassen.

Auch das Merkmal *Deutsch als Zweitsprache* sollte zukünftig als Einflussfaktor auf die Lehrkräfteeinschätzung und möglicher Grund für Fehlklassifikationen im FERMAT berücksichtigt werden. Denn Studien zeigen, dass Schüler*innen mit Einwanderungsgeschichte von Lehrkräften weniger akkurat eingeschätzt und tendenziell schlechter beurteilt werden als Schüler*innen ohne Einwanderungsgeschichte (z. B. Karst & Bonefeld, 2020; Wagner, 2024).

Um potenziellen Fehleinschätzungen im FERMAT proaktiv entgegenzuwirken, könnte es sinnvoll sein, weitere Items zu ergänzen, z. B.: *Die Rechenschwierigkeiten des*der Schüler*in sind nicht auf Lese-/Rechtschreibschwierigkeiten oder sprachliche Barrieren, z. B. Deutsch als Zweitsprache, zurückzuführen*. Um den praktischen Nutzen solcher ergänzenden Items zu bewerten, könnten, ähnlich wie in anderen Studien (z. B. Dögnitz, 2022; Bender et al., 2024), Lerntherapeut*innen als Expert*innen hinzugezogen werden.

*Verhalten der Schüler*innen*

Kinder mit einer Rechenstörung sind häufiger von internalisierenden und externalisierenden Auffälligkeiten betroffen als Kinder ohne Lernstörung (Aro et al., 2021; Auerbach et al., 2008; Fischbach et al., 2010; Kohn et al., 2013). Dies könnte die Einschätzungen der Lehrkräfte im FERMAT beeinflussen. Denn Studien belegen, dass Verhaltensauffälligkeiten der Schüler*innen die Leistungsbeurteilung und Urteilsgenauigkeit von Lehrkräften negativ beeinflussen können (z. B. Krämer & Zimmermann 2020; Schabmann & Schmidt 2009). Auch in der Studie von Fischer et al. (2015) gab es einen Hinweis darauf, dass Kinder, die den Lehrkräften im Unterricht auffällig erschienen, fälschlicherweise als rechenschwach eingestuft wurden. Möglicherweise sind sich Lehrkräfte dieses Einflusses bewusst. Denn in einer Studie von Schabmann und Schmidt (2009) zeigte sich, dass die Lehrkräfte ihre Einschätzung im Hinblick auf die Lesekompetenz umso stärker anhoben beziehungsweise nach oben korrigierten, je störender sie das Verhalten der Schüler*innen (oppositionelles Verhalten) wahrnahmen.

Somit wäre denkbar, dass Lehrkräfte Kinder mit einer Rechenstörung weniger präzise einschätzen oder die Schwierigkeiten im Rechnen auf verhaltensbezogene Merkmale zurückführen. In der Konsequenz könnte dies dazu führen, dass diese Kinder geringere Chancen auf eine angemessene Förderung haben. Zukünftige Studien sollten daher auch das Verhalten der Schüler*innen als potenziellen Einflussfaktor auf die diagnostische Einschätzung der Lehrkräfte untersuchen.

Dies gilt auch mit Blick auf die Ergebnisse aus *Studie I* (Lamb et al., 2024a). Studien, die die Ursachen und Symptome der Rechenstörung untersuchen, sollten komorbide, affektive und verhaltensbezogene Auffälligkeiten (z. B. Angst, Depression, ADHS und oppositionelles Verhalten) als Kovariaten berücksichtigen, da komorbide Auffälligkeiten die mathematische Leistung negativ beeinflussen können. Ein Beispiel

hierfür ist die Mathematikangst. Kinder und Jugendliche mit Rechenstörungen leiden häufiger unter Mathematikangst als nicht betroffene Kinder, was sich wiederum negativ auf die mathematische Leistung auswirken kann (Barroso et al., 2021; Semeraro et al., 2020).

Lehrkraftbezogene Einflussfaktoren

Studie III (Lamb et al., 2025) zeigte, dass bei gleicher objektiver Testleistung einige Lehrkräfte Schwierigkeiten in den Rechenfertigkeiten ihrer Schüler*innen sahen, andere hingegen nicht. Dies ist ein Hinweis auf den Strenge- beziehungsweise Milde-Effekt (Helmke, 2017; vgl. Kapitel 4.2). Auch in *Studie II* (Lamb et al., 2024b) wurden einige Kinder mit testdiagnostischen Anzeichen einer Rechenstörung von den Lehrkräften nicht erkannt, während andere ohne Anzeichen fälschlicherweise als rechenschwach eingestuft wurden. Diese Fehlklassifikationen und Variationen zwischen den Lehrkräften können verschiedene Ursachen haben.

Es ist möglich, dass Lehrkräfte bei der Einschätzung von Schwierigkeiten im FERMAT auf die *soziale Bezugsnorm* zurückgreifen und sich am Leistungsniveau der Klasse orientieren (Helmke, 2017). Studien haben festgestellt, dass Schüler*innen in leistungsstarken Klassen negativer beurteilt werden als vergleichbare Schüler*innenleistungen in leistungsschwächeren Klassen (z. B. Gnas et al., 2022; Schmitterer & Brod, 2021). In der Konsequenz bedeutet dies, dass Kinder in leistungsschwachen Klassen geringere Chancen auf eine frühzeitige Risikoerkennung und Förderung haben. Zukünftige Studien sollten daher untersuchen, ob Kinder in leistungsstarken Klassen häufiger als rechenschwach eingeschätzt werden und ob die Schüler*innenmerkmale je nach Leistungsdurchschnitt der Klasse unterschiedlich stark in die Lehrkräfteeinschätzung einfließen.

Daran anknüpfend sollten auch die Leistungserwartungen der Lehrkräfte untersucht werden, da sich diese negativ auf die Leistungen der Schüler*innen auswirken können (z. B. de Boer et al., 2010). Jussim und Eccles (1992) zeigten, dass die Erwartungen der Lehrkräfte die Schüler*innenleistungen sogar stärker vorhersagten als deren vorherige Leistung und Motivation.

Berufserfahrung und Wissen der Lehrkräfte

Die Unterschiede in den Einschätzungen der Lehrkräfte (Lamb et al., 2025) und Fehlklassifikationen im FERMAT (Lamb et al., 2024b) könnten mit der Berufs- und Unterrichtserfahrung der Lehrkräfte in Zusammenhang stehen. Denn erfahrene Lehrkräfte erreichen tendenziell eine höhere Urteilsgenauigkeit als unerfahrene Lehrkräfte (Kosel et al., 2024; Wagner, 2024). Auch, wenn einige Studien keinen oder nur einen schwachen Zusammenhang zwischen der Urteilsgenauigkeit und der Berufserfahrung der Lehrkräfte finden (z. B. Anders et al., 2010; McElvany et al., 2009), wäre es denkbar, dass die Genauigkeit der Risikoidentifikation mittels des FERMAT davon abhängt, ob erfahrene oder unerfahrene Lehrkräfte den FERMAT ausfüllen.

Wenn Lehrkräfte einschätzen sollen, ob ihre Schüler*innen Anzeichen einer Rechenstörung aufweisen, spielt das pädagogisch-psychologische Wissen der Lehrkräfte über Rechenstörungen und ihre diagnostische Kompetenz eine zentrale Rolle (vgl. Kapitel 4.1). Zwar verfügen (angehende) Grundschullehrkräfte überwiegend über Wissen zu den frühen Anzeichen und typischen Symptomen von Rechenstörungen, dennoch bestehen Unsicherheiten. So gaben in einer Studie von Bender et al. (2024) etwa 33 % der Befragten an, unsicher zu sein, ob Kinder mit Rechenstörungen Schwierigkeiten haben, mathematische Fakten abzurufen. Etwa 17 % waren sich nicht sicher, ob unreife Rechenstrategien typisch für Rechenstörungen sind, knapp 30 % konnten nicht eindeutig sagen, ob die Mengenverarbeitung bei Kindern mit Rechenstörungen beeinträchtigt ist,

und rund 18 % wussten nicht genau, ob Kinder mit Rechenstörungen zu Transkodierfehlern neigen. Studien aus dem internationalen Raum kommen zu ähnlichen Ergebnissen und zeigen, dass Lehrkräfte verschiedener Schulformen nicht über ausreichende Kenntnisse und Erfahrungen in Bezug auf die Bedeutung, Merkmale, Folgen, Ursachen und Interventionsstrategien von Rechenstörungen verfügen (Butterworth et al., 2011; Chideridou-Mandari et al., 2016; Mutlu et al., 2022; Sousa et al., 2017; Tennant & Tennant, 2010). Fehlendes Wissen über Rechenstörungen könnte somit ein Grund für Fehlklassifikationen und variierende Beurteilungen zwischen den Lehrkräften sein.

Um potenzielle Gründe für Fehlklassifikationen (falsch-positive und falsch-negative Ergebnisse im FERMAT) und Unterschiede zwischen den Einschätzungen der Lehrkräfte aufzudecken, eignet sich ein *Mixed-Methods-Design* (z. B. Buchholtz 2021; Kuckartz et al., 2014). Dabei kann der FERMAT im Rahmen eines *Expert*inneninterviews* als Gesprächsgrundlage dienen. Die Lehrkräfte könnten gebeten werden, ihre Einschätzungen im FERMAT inhaltlich zu begründen. Über eine *qualitative Analyse* (Mayring & Fenzel, 2019) der Antworten könnte offengelegt werden, ob die Lehrkräfte ihre Einschätzungen im FERMAT plausibel begründen können.

Über eine *quantitative Befragung* könnte erfasst werden, ob Lehrkräfte über ausreichendes diagnostisches und pädagogisch-psychologisches Wissen verfügen. Dabei sollte auch berücksichtigt werden, dass die diagnostische Expertise der Lehrkräfte ebenso das methodische und prozedurale Wissen über diagnostische Methoden und Verfahren sowie die Beherrschung dieser umfasst (Helmke, 2017). Denn obwohl die Lehrkräfte vor dem Ausfüllen des FERMAT umfassende Informationen über den Fragebogen erhielten und sich mit dem Fragebogen vertraut machten, hatte keine Lehrkraft Erfahrung mit dem FERMAT, da es sich zum Erhebungszeitpunkt um ein noch unveröffentlichtes Instrument

handelte. Daher sollte in Zukunft zumindest die Vorerfahrung im Umgang mit fragebogengestützten Beurteilungsskalen erhoben werden.

Die diagnostische Einschätzung von Lehrkräften stellt ein multifaktorielles Konstrukt dar, das von einer Vielzahl zum Teil miteinander interagierender Variablen beeinflusst wird. Um die Rolle der einzelnen Einflussfaktoren (z. B. pädagogisch-psychologisches Wissen zu Rechenstörungen, diagnostischen Verfahren, Verhaltensauffälligkeiten der Schüler*innen etc.) klarer zu definieren, könnte es sinnvoll sein, mithilfe von Moderations- und Mediationsanalysen oder Strukturgleichungsmodellen zu untersuchen, wie die verschiedenen Einflussfaktoren zusammenhängen. So könnte untersucht werden, welche Einflussfaktoren direkt in die Einschätzung der Lehrkräfte einfließen und welche Faktoren einen indirekten Einfluss haben. Zudem sollte geprüft werden, ob die Einflussfaktoren in Abhängigkeit von der zu beurteilenden Schwierigkeit im FERMAT variieren.

Dabei ist es wichtig, dass die potenziellen Einflussfaktoren (z. B. Wissen über Rechenstörung) mit validierten Instrumenten erhoben werden. Denn in aktuellen Forschungsarbeiten werden unterschiedliche Fragebögen eingesetzt, für die zum Teil (noch) keine Gütekriterien vorliegen (z. B. Bender et al., 2024; Dögnitz, 2022; Wagner & Ehlert, 2016). Um die Einschätzungen der Lehrkräfte zu erheben, könnte der FERMAT (Lamb et al., 2024b) eingesetzt werden. Zur Erfassung des Wissens könnte beispielsweise die Subskala Teilleistungsstörungen des DiWi genutzt werden (Steinmayr et al., 2022).

Die weitere Erforschung von Faktoren, die die diagnostische Einschätzung der Lehrkräfte beeinflussen können, ist nicht nur für die Identifikation von Kindern mit Anzeichen einer Rechenstörung im schulischen Setting relevant, sondern auch für die Forschung. Denn es ist gängige Praxis, Lehrkrafturteile zur Überprüfung der kriterialen Validität diagnostischer Verfahren heranzuziehen (Dögnitz, 2022). Beispiele hierfür sind der in Kapitel 3.1 vorgestellte ERT 3+ (Holzer et al., 2010) sowie der CODY-M 2 – 4

Test (Kuhn et al., 2017). In beiden Testverfahren wurde die Beurteilung der Lehrkraft (rechenschwach: ja/nein) als kriteriales Maß zur Bestimmung der Validität herangezogen. Dabei bleibt jedoch häufig unberücksichtigt, dass Lehrkrafturteile selbst von einer Vielzahl an Faktoren beeinflusst werden können (vgl. Kapitel 4.2). Vor diesem Hintergrund ist die Verwendung von Lehrkrafturteilen als *objektives Kriterium* kritisch zu hinterfragen.

5.3 Methodische Limitationen

Um die Frage zu beantworten, ob sich die basisnumerischen Fertigkeiten von Grundschulkindern mit einer Rechenstörung qualitativ von denen unbeeinträchtigter Kinder unterscheiden oder ob die basisnumerische Entwicklung von Kindern mit Rechenstörung verzögert ist, wurden unter anderem die basisnumerischen Profile von Viertklässler*innen mit Rechenstörungen und Zweitklässler*innen ohne Rechenstörungen verglichen (Lamb et al., 2024a). Dazu wurde in *Studie I* (Lamb et al., 2024a) eine unselektierte Stichprobe rekrutiert. Der Anteil der Kinder mit einer Rechenstörung beträgt in etwa 14 % (n = 68). Die Prävalenz der Stichprobe ist vergleichbar mit anderen Studien, die ebenfalls unselektierte Stichproben untersuchten (z. B. ca. 16 %; Fischer et al., 2015). Dennoch ist die Anzahl von n = 68 Kindern mit einer Rechenstörung für statistische Analysen eher klein. Besonders deutlich wird dies mit Blick auf den Profilvergleich von Viertklässler*innen mit Rechenstörung (n = 17) und unbeeinträchtigten Zweitklässler*innen (n = 177). Daraus ergeben sich Limitationen hinsichtlich der Power und Aussagekraft der statistischen Analysen. Um die Prävalenz der Rechenstörungen in der Stichprobe gezielt zu erhöhen und eine robustere Analyse zu ermöglichen, wäre die Untersuchung einer klinischen Stichprobe in einer zukünftigen Forschungsarbeit denkbar. Dies hätte zudem den Vorteil, dass die Ergebnisse auch für Kinder mit einer diagnostizierten Rechenstörung gelten würden. Denn eine

uneingeschränkte Generealisierung ist basierend auf den vorliegenden Daten nicht möglich.

Für zukünftige Arbeiten wäre es außerdem wichtig, Längsschnittstudien durchzuführen. Dabei sollte die Forschung auch ältere Kinder (z. B. fünfte bis siebte Klassenstufe) berücksichtigen. Denn Studien zeigen, dass sich die mathematischen Leistungen von Kindern mit Rechenstörungen in höheren Klassenstufen im Laufe der Zeit verbessern, sie aber nicht auf das Niveau gleichaltriger unbeeinträchtigter Kinder aufschließen (McCaskey et al., 2017; Meier et al., 2021; Schulz et al., 2018). Darüber hinaus zeigen Studien, dass auch Erwachsene mit Rechenstörung teils basisnumerische Defizite aufweisen (z. B. Bulthé et al., 2019; Gliksman & Henik, 2019; de Visscher et al., 2018). Somit ist davon auszugehen, dass nicht alle Personen mit einer Rechenstörung auf das basisnumerische Niveau gleichaltriger aufschließen.

Für ein ganzheitliches Bild müssten zusätzliche Informationen auf neuronaler Ebene gewonnen werden, zum Beispiel über funktionelle Magnetresonanztomografie oder funktionelle Nahinfrarotspektroskopie (Landerl et al., 2022). Denn die Beeinträchtigungen sind nicht nur auf behavioraler und kognitiver, sondern auch auf neuronaler Ebene beobachtbar (z. B. Pielsticker et al., 2024).

Ob die Defizite, die auf unterschiedlichen Ebenen beobachtbar sind, die Ursache der Rechenstörung oder die Folge eines beeinträchtigten Lernprozesses sind, ist ebenfalls noch nicht abschließend beantwortet. Zur Beantwortung dieser Frage, müssten Kinder mit einem genetischen Risiko für Rechenstörungen bereits vor dem Erwerb von Rechenfertigkeiten längsschnittlich untersucht werden, um zu prüfen, ob die Beeinträchtigungen auf neuronaler Ebene bereits vor dem Erwerb von Rechenfertigkeiten vorhanden sind oder erst später als Folge eines beeinträchtigten Lernprozesses auftreten (Vogel & De Smedt, 2021). Solche Studien sind jedoch aus verschiedenen Gründen (finanzielle Ressourcen, Abbrecherquoten) schwer durchführbar.

In *Studie II* (Lamb et al., 2024b) wurden unter anderem verschiedene screeningrelevante Güteindizes für den FERMAT berechnet, um diesen hinsichtlich seiner Gütekonformität beurteilen zu können. Die Gruppengrößen, der Kinder mit Anzeichen einer Störung, variierten zwischen $n = 31$ (CODY-M 2 – 4 PR ≤ 7) und $n = 45$ (HRT 1 – 4 PR ≤ 10). Durch die geringen Fallzahlen können insbesondere die Sensitivität und Spezifität verzerrt sein. Zukünftige Studien sollten daher größere Stichproben einbeziehen, um robustere Schlussfolgerungen und klassenstufenspezifische Analysen zu ermöglichen.

Obwohl bei der Risikoidentifikation im FERMAT der Status quo im Vordergrund steht (Lamb et al., 2024b) und weniger die prognostische Validität, sollte diese zusätzlich zur kriterialen Validität durch zwei unabhängige Messzeitpunkte in einer Längsschnittstudie untersucht werden. Dies gilt nicht nur für den FERMAT, sondern auch für andere Screeninginstrumente. Denn Angaben zur prognostischen Validität fehlen meist (Walter, 2020) und wenn sie vorliegen, handelt es sich oft nur um statistische Korrelationen zwischen der Ausgangsleistung und der späteren Mathematikleistung, sodass die Aussagen zur prognostischen Validität eher explorativen Charakter haben (Gloor, 2023).

Um verschiedene Einflussfaktoren auf die Einschätzung der Lehrkräfte im FERMAT zu untersuchen, wurden in *Studie III* (Lamb et al., 2025) Multilevel-Analysen durchgeführt. Dieser Mehrebenen-Ansatz ist im Forschungskontext der diagnostischen Kompetenz ein vergleichsweise selten angewandter statistischer Ansatz, der zukünftig aber vorzugsweise verwendet werden sollte (Karst et al., 2017). Allerdings ist die Stichprobengröße auf der zweiten Ebene ($N = 33$ Lehrkräfte) in *Studie III* (Lamb et al., 2025) eher gering einzuschätzen, da eine Stichprobengröße von $N = 50$ oder weniger zu Verzerrungen der Ergebnisse führen kann (Maas & Hox, 2005). Daher sollte zukünftig eine größere Stichprobe rekrutiert werden.

Da empirische Befunde (z. B. Cai et al., 2018; Krajewski, 2008) zeigen, dass domänenspezifische und -generelle Fertigkeiten je nach Altersstufe unterschiedliche Relevanz für die Vorhersage der Mathematikleistung haben, könnten sich ähnliche Effekte auch in den Lehrkräfteeinschätzungen widerspiegeln. Um die Veränderungen der Einflussfaktoren auf die Einschätzungen der Lehrkräfte über die Zeit zu untersuchen, wäre eine Längsschnittstudie sinnvoll.

Für zukünftige Forschungen ist es ratsam, a priori Poweranalysen durchzuführen, um sicherzustellen, dass die Stichprobengröße für die geplanten Analysen ausreichend ist. Für Multilevel-Analysen können Simulationen mit unterschiedlichen Stichprobengrößen durchgeführt werden (z. B. mittels des R-Paktes „smir“ von Green & MacLeod, 2016). Für die ROC-Analysen könnte das „pROC“ R-Paket (Robin et al., 2011) genutzt werden.

6 Fazit und Ausblick

Die vorliegende Dissertation untersuchte, *wie Lehrkräfte Grundschulkinder mit Anzeichen für eine Rechenstörung frühzeitig identifizieren können.* Dazu wurden drei Studien durchgeführt, die an bisherige Forschungsergebnisse anknüpfen.

Die Ergebnisse aus *Studie I* (Lamb et al., 2024a) zeigten, dass Grundschulkinder mit einer Rechenstörung Defizite in der Basisnumerik aufweisen, die auf eine Entwicklungsverzögerung hindeuten. Ursächlich könnte eine Beeinträchtigung des ANS und eine gestörte Interaktion zwischen dem ANS und dem visuell-arabischen Modul sein. Signifikante Hinweise auf eine Beeinträchtigung im OTS wurden nicht gefunden.

Für die pädagogische Praxis bedeutet dies, dass basisnumerische Fertigkeiten im gesamten diagnostischen Prozess von der Früherkennung über die Diagnostik bis hin zur Förderung eine Rolle spielen sollten (z. B. Fischer et al., 2017; Kuhn & Schwenk, 2018). Dass die Defizite in der Basisnumerik auf eine Entwicklungsverzögerung hindeuten, impliziert, dass Kinder mit Rechenstörungen ähnliche Entwicklungsschritte durchlaufen wie Kinder ohne Beeinträchtigungen, allerdings zeitlich verzögert. Daraus lässt sich ableiten, dass allgemeine Kompetenzentwicklungsmodelle der Zahlenverarbeitung (z. B. von Aster & Shalev, 2007; Fischer et al., 2017) einen geeigneten theoretischen Rahmen bieten, um die einzelnen diagnostischen Schritte inhaltlich zu verzahnen und aufeinander abzustimmen.

Der in *Studie II* (Lamb et al., 2024b) vorgestellte Screeningfragebogen (FERMAT) identifiziert Kinder mit und ohne testdiagnostische Anzeichen einer Rechenstörung ökonomisch, reliabel und mit angemessener Genauigkeit. Durch das einfache und handhabbare Format des vorgestellten Lehrkräftefragebogens kann dieser gut in den Unterricht integriert werden. Aufgrund seiner wissenschaftlichen Fundierung und Gütekonformität eignet sich der FERMAT auch als Legitimationsgrundlage für

Elterngespräche, beispielsweise wenn Kinder einer weiteren Diagnostik zugeführt werden sollen.

Die Ergebnisse aus *Studie III* (Lamb et al., 2025) sprechen insgesamt dafür, dass die diagnostische Einschätzung der Lehrkräfte im FERMAT in erster Linie von den tatsächlichen mathematischen Leistungen der Schüler*innen bestimmt wird (Lamb et al., 2025). Dies ist ein wichtiges und positives Ergebnis, da die Lehrkräfte eine zentrale Rolle bei der Identifikation von Kindern mit Anzeichen einer Rechenstörung einnehmen.

Dennoch besteht weiterhin Optimierungsbedarf, um die frühzeitige Risikoidentifikation zu verbessern. Im Hinblick auf die Zielgruppe der Kinder mit Anzeichen einer Rechenstörung ist es im Sinne inklusiver Bildungsprozesse wichtig, auch Kinder mit weniger ausgeprägten, aber dennoch relevanten Rechenschwierigkeiten zu berücksichtigen. Zudem sollten Schüler*innenmerkmale, die im Zusammenhang mit Rechenschwierigkeiten auftreten, beispielswiese Verhaltensauffälligkeiten, stärker berücksichtigt werden, da diese nicht nur das Wohlbefinden der Kinder, sondern auch die diagnostischen Einschätzungen der Lehrkräfte beeinflussen können.

Auf der Ebene der Lehrkräfte sollten weitere Aspekte untersucht werden, die die diagnostischen Einschätzungen der Lehrkräfte und damit den Prozess der Früherkennung beeinflussen können. Dazu zählt unter anderem das pädagogisch-psychologische Wissen der Lehrkräfte, welches auch in der Hochschullehre stärker verankert werden sollte.

Mit Blick auf den in dieser Dissertation vorgestellten Screeningfragebogen ist es wichtig, die diagnostische Genauigkeit des Instruments weiter gezielt zu erhöhen, indem sowohl die inhaltlichen als auch die psychometrischen Eigenschaften des Screeningfragebogens vor dem Hintergrund subgruppen- und klassenstufenspezifischer Analysen untersucht werden.

Damit ist die Forschung jedoch noch nicht abgeschlossen. Denn die Früherkennung stellt lediglich den ersten Schritt auf dem Weg zu einer geeigneten Intervention dar

(Tröster, 2009; Voß, 2017; vgl. Kapitel 3). Die Identifikation von Schüler*innen mit einem Risiko für eine Rechenstörung führt nicht kausal zu adäquatem didaktischem Handeln. Schüler*innen profitieren nur dann von einer frühzeitigen Risikoidentifikation, wenn die Lehrkraft aus den Ergebnissen *weiterer standardisierter Tests* (für eine Übersicht s. Kuhn & Schwenk, 2018 oder S3-Leitlinie; Schulte-Körne & Haberstroh, 2018) die richtigen Schlüsse zieht und diese im Unterricht umsetzt (Tröster, 2018).

Dazu zählt unter anderem die Bereitstellung geeigneter Fördermaterialien. Lehrkräfte müssen daher über Kenntnisse zur Qualität und Wirksamkeit standardisierter Förderprogramme verfügen (Faix et al., 2023). Bei der Auswahl evidenzbasierter Fördermaßnahmen sollten Lehrkräfte unterstützt werden (Wittich & Kuhl, 2021), beispielsweise über das LONDI-Hilfssystem (Schulte-Körne & Hasselhorn, 2025). Dieses bündelt wissenschaftlich fundierte Informationen für Lehrkräfte und unterstützt bei der Auswahl geeigneter Testverfahren und Interventionsprogramme durch ein webbasiertes Tool. Damit solche Informationen alle Lehrkräfte erreichen, sollten diese systematisch in die Schulen getragen werden.

Da Kinder mit Rechenstörungen vor allem von einer nicht-curricularen Intervention im Einzelsetting profitieren (z. B. Meta-Analyse: Ise et al., 2012), sollten Kinder mit einer Rechenstörung neben der Förderung im regulären Mathematikunterricht begleitende professionelle Unterstützung erhalten (Faix et al., 2023). Um Eltern dahingehend beraten zu können, müssen Lehrkräfte wissen, welche außerschulischen Angebote (z. B. Lerntherapie; Bender et al., 2017) in der Region zur Verfügung stehen (Faix et al., 2023).

Vor dem Hintergrund inklusiver Schulentwicklung sollte der gezielte Einsatz von Lehrtherapeut*innen an Schulen diskutiert werden (z. B. Ricken, 2014). Von einer schulintegrierten Lerntherapie (z. B. Balke-Melcher et al., 2016; Hilkenmeier et al., 2020) würden besonders die Kinder profitieren, die das doppelte Diskrepanzkriterium, das heißt die Diagnose einer Rechenstörung nach ICD-10/-11, nicht erfüllen (Hilkenmeier et al.,

2020; Mähler, 2021) und somit keine finanzielle Unterstützung für eine außerschulische Fördermaßnahme erhalten (Busch et al., 2018; Mähler, 2021). Darüber hinaus würden auch Kinder profitieren, die die Diagnose einer Rechenstörung nach ICD-10/-11 erfüllen. Denn bisher ist die Finanzierung einer Lerntherapie an die *Eingliederungshilfe* nach § 35a Kinder- und Jugendhilfegesetz (KJHG; Bundesministerium der Justiz, o. J.) gekoppelt. Die Eingliederungshilfe kann gewährt werden, wenn aufgrund der Lernstörung eine seelische Behinderung droht oder bereits eingetreten ist (Mähler, 2021). Im Sinne des *Wait-to-Fail-Ansatzes* (z. B. Ricken, 2014) ist eine finanzielle Unterstützung also auch für Kinder mit einer Rechenstörung nach ICD-10/-11 erst dann möglich, wenn bereits gravierende Einschränkungen vorliegen. Dem könnte durch eine vom § 35a losgelöste schulintegrierte Lerntherapie entgegengewirkt werden.

Darüber hinaus können Lerntherapeut*innen als Expert*innen für Lernstörungen eine weitere wichtige Rolle im schulischen Kontext einnehmen. Denn neben der Förderung der betroffenen Kinder ist es ebenso wichtig, dass im Unterricht angemessen über Rechenstörungen informiert wird. Denn Missverständnisse und fehlendes Wissen über Lernstörungen unter den Schüler*innen können zur Stigmatisierung und Ausgrenzung der betroffenen Kinder führen (z. B. Laursen et al., 2021; Schuchardt et al., 2021). Aus diesem Grund ist es wichtig, dass sich die Forschung mit dem Thema der *Psychoedukation* auseinandersetzt und bereits bestehende Ansätze (Gabriel et al., 2021) in den Schulen implementiert werden. Lerntherapeut*innen als Expert*innen könnten die psychoedukativen Unterrichtseinheiten im Klassenverband durchführen.

Auf schulrechtlicher Ebene ist es von zentraler Bedeutung, dass Rechenstörungen den gleichen Stellenwert erhalten wie Lese-Rechtschreibstörungen. Denn für „Schüler*innen mit besonderen Schwierigkeiten im Rechnen“ (KMK, 2007) gelten nach wie vor nachteilige Regelungen, beispielsweise im Hinblick auf den Nachteilsausgleich.

Die Ergebnisse der vorliegenden Dissertation weisen darauf hin, dass Kinder mit Rechenstörungen entwicklungsverzögerte Defizite in der Basisnumerik aufweisen, die durch eine Beeinträchtigung des ANS und eine gestörte Interaktion zwischen dem ANS und dem visuell-arabischen Modul verursacht werden (Lamb et al., 2024a). Dieser Befund unterstützt die Annahme, dass Rechenstörungen nicht plötzlich auftreten, sondern ihren Ursprung in vorschulischen Entwicklungsprozessen haben (Krajewski, 2008). Dies wird auch dadurch bestätigt, dass sich erste Schwierigkeiten in basalen Vorläuferfertigkeiten bereits im Vorschulalter zeigen (z. B. Aunio & Niemivirta, 2011; Gallit et al., 2017) und Rechenstörungen schon im Vorschulalter basierend auf basalen basisnumerischen Fertigkeiten vorhergesagt werden können (Krajewski, 2008).

Zukünftige Forschungsarbeiten sollten daher auch frühpädagogische Fachkräfte und Eltern stärker in den Prozess der Früherkennung und Förderung einbeziehen. Denn auch wenn rechenstörungsspezifische Defizite primär mit den Lernvoraussetzungen (z. B. domänenspezifische Kognition) des Kindes in Zusammenhang stehen (vgl. Kap. 2.1 und 2.2), sind Rechenstörungen in ein Geflecht äußerer Einflussfaktoren eingebettet, die eine Brückenfunktion einnehmen und sich lernförderlich oder -hinderlich auf die Symptomatik der Rechenstörung auswirken können (vgl. Kapitel 2.2). Dies bedeutet, dass die Ausprägung der Symptomatik auch durch den schulischen Unterricht, das familiäre Umfeld und andere externe Faktoren beeinflusst werden kann (z. B. Jordan & Levine, 2009; Schuchardt et al., 2014). So gibt es beispielsweise Hinweise darauf, dass die häusliche Lernumgebung die vorschulische Entwicklung mathematischer Kompetenzen von Kindern beeinflussen kann (Susperreguy et al., 2020; Zhao & Gibson, 2023). Durch spezifische frühzeitige Interventionen im Vorschulbereich können anfängliche Defizite bis zum Schuleintritt verringert werden (z. B. Moraske et al., 2018).

Daraus darf jedoch nicht der Schluss gezogen werden, dass der Fokus in Zukunft ausschließlich auf vorschulischen Früherkennungsprozessen liegen sollte. Denn Eltern

gelingt es weniger gut, Schwierigkeiten in der mathematischen Entwicklung ihrer Kinder zuverlässig zu erkennen, und auch das frühpädagogische Fachpersonal muss diesbezüglich gezielter geschult werden (Dollinger, 2013; Tabeling et al., 2022).

Zudem können Kinder mit Anzeichen einer Rechenstörung leicht übersehen werden. Dies hat mehrere Gründe. Einerseits gelingt es Kindern mit durchschnittlichen kognitiven Fähigkeiten gut, Defizite in den basisnumerischen Fertigkeiten zu kompensieren (Fritz & Ricken, 2005; Tröster, 2009). Andererseits können Kinder ohne Risiko für eine Rechenstörung aufgrund der Prävalenz besser identifiziert werden als Kinder mit einem Risiko (Gallit et al., 2017). Grund dafür ist die Prävalenz. Denn während die Prävalenz der Rechenstörung zwischen 2 und 8 % liegt (vgl. Kapitel 2.1), ist der Anteil nicht betroffener Kinder mit 92 bis 98 % um ein Vielfaches höher, was die Identifikation nicht beeinträchtigter Kinder deutlich einfacher macht (Gallit et al., 2017).

Darüber hinaus nimmt die prognostische Validität zunehmend ab. Beispielsweise zeigen die Ergebnisse von Krajewski (2008), dass die Trefferquote bei der Risikovorhersage von Rechenschwierigkeiten mit der Zeit abnimmt. Bei einem Screening kurz vor der Einschulung wurden 61 % (11 von 18) der Erstklässler*innen korrekt als rechenschwach identifiziert, in der zweiten Klasse waren es 47 % und in der vierten Klasse nur noch 15 % (Krajewski, 2005). Aus diagnostischer Sicht unterstreicht dies die Notwendigkeit regelmäßiger Screenings zu verschiedenen Zeitpunkten vor und während der Grundschulzeit. Es bleibt abzuwarten, ob Grundschullehrkräfte dazu zukünftig den FERMAT-Screeningfragebogen nutzen. Da die Akzeptanz diagnostischer Verfahren im schulischen Setting insbesondere von der ökonomischen Effizienz abhängt (Marx & Lenhard, 2011), sollte der FERMAT digitalisiert werden, sodass die Testergebnisse automatisch generiert werden. Dies könnte eine ressourcenschonende Implementation in der Schul- und Unterrichtspraxis begünstigen (Kuhl et al., 2024).

7 Literatur

Akin, A. (2022). Is reading comprehension associated with mathematics skills: A meta-analysis research. *International Online Journal of Primary Education*, *11*(1), 47–61. https://doi.org/10.55020/iojpe.1052559

American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders. Fifth Edition (5th ed., DSM-5)*. American Psychiatric Association.

American Psychiatric Association (APA). (2022). *Diagnostic and statistical manual of mental disorders. Fifth Edition Text Revision (5th ed., text rev., DSM-5-TR)*. American Psychiatric Association.

An, W., & Bai, H. (2015). *Propensity score analysis: Fundamentals and development*. Guilford Press.

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 57*(3), 175–193. https://doi.org/10.2378/peu2010.art13d

Andersson, U., & Östergren, R. (2012). Number magnitude processing and basic cognitive functions in children with mathematical learning disabilities. *Learning and Individual Differences, 22*(6), 701–714. https://doi.org/10.1016/j.lindif.2012.05.004

Aro, T., Eklund, K., Eloranta, A.-K., Ahonen, T., & Rescorla, L. (2021). Learning Disabilities Elevate Children's Risk for Behavioral-Emotional Problems: Differences Between LD Types, Genders, and Contexts. *Journal of Learning Disabilities*, *55*(6), 465–481. https://doi.org/10.1177/00222194211056297

Artelt, C., & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*(34), 157–160. https://doi.org/10.1024/1010-0652.23.34.157

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-83412-6_4

von Aster, M. (2013). Wie kommen Zahlen in den Kopf und was kann sie daran hindern? Ein Modell der normalen und abweichenden Entwicklung zahlenverarbeitender Hirnfunktionen. In M. von Aster & J. H. Lorenz (Hrsg.), *Rechenstörungen bei Kindern* (S. 15–38). Vandenhoeck & Ruprecht. https://doi.org/10.13109/9783666462580.15

von Aster, M., Schweiter, M., & Weinhold-Zulauf, M. (2007). Rechenstörungen bei Kindern: Vorläufer, Prävalenz und psychische Symptome. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39*, 85–96.

von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology, 49*, 868–873. https://doi.org/10.1111/j.1469-8749.2007.00868.x

von Aster, M., Weinhold-Zulauf, M., & Horn, R. (2013). *Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern (ZAREKI-R)*. Pearson.

Auerbach, J. G., Gross-Tsur, V., Manor, O., & Shalev, R. S. (2008). Emotional and Behavioral Characteristics Over a Six-Year Period in Youths With Persistent and Nonpersistent Dyscalculia. *Journal of Learning Disabilities, 41*(3), 263–273. https://doi.org/10.1177/0022219408315637

Aunio, P., & Niemivirta, M. (2011). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences, 20(5),* 427–435. https://doi.org/10.1016/j.lindif.2010.06.003

Baddeley, A. D. (1986). *Working memory.* Oxford University Press.

Balke-Melcher, C., Schuchardt, K., Wolpers, J., & Mähler, C. (2016). Modellprojekt zur Lernförderung bei Schriftsprachschwierigkeiten in der Grundschule. *Lernen und Lernstörungen, 5*(1), 17–31. https://doi.org/10.1024/2235-0977/a000122

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., & Jacobsen, S. J. (2005). Math learning disorder: incidence in a population-based birth cohort, 1976-82, Rochester, Minn. *Ambulatory Pediatrics*, *5*(5), 281–289. https://doi.org/10.1367/A04-209R.1

Barroso, C., Ganley, C. M., McGraw, A. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. (2021). A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin, 147*(2), 134–168. https://doi.org/10.1037/bul0000307

Bartelet, D., Ansari, D., Vaessen, A., & Blomert, L. (2014). Cognitive subtypes of mathematics learning difficulties in primary education. *Research in Developmental Disabilities, 35*(3), 657–670. https://doi.org/10.1016/j.ridd.2013.12.010

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2

Beauducel, A., & Leue, A. (2014). *Psychologische Diagnostik*. Hogrefe.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 43–55. https://doi.org/10.1037/1045-3830.23.1.43

Bender, F., Brandelik, K., Jeske, K., Lipka, M., Löffler, C., Mannhaupt, G., Naumann, C. L., Nolte, M., Ricken, G., Rosin, H., Scheerer-Neumann, G., von Aster, M., & von Orloff, M. (2017). Die integrative Lerntherapie: Therapieform zur Behandlung von Lernstörungen. *Lernen und Lernstörungen*. *6*(2), 65–73. https://doi.org/10.1024/2235-0977/a000167

Bender, L., Hotz, L., & Renkl, A. (2024). Lehrkräftewissen zur Rechenstörung in der Grundschule (und darüber hinaus). *SEMINAR, 30*(4), 54–69. wbv Publikation. https://doi.org/10.3278/SEM2404W005

Benz, C., Peter-Koop, A., & Grüßing, M. (2015). *Frühe mathematische Bildung*. Springer Spektrum. https://doi.org/10.1007/978-3-8274-2633-8

Betz, D., & Breuninger, H. (1982). *Teufelskreis Lernstörungen: Theoretische Grundlegung und Standardprogramm*. Beltz.

Blumenthal, Y., Kuhlmann, K., & Hartke, B. (2014). Diagnostik und Prävention von Lernschwierigkeiten im Aptitude Treatment Interaction- (ATI-) und Response to Intervention- (RTI-) Ansatz. In M. Hasselhorn, W. Schneider, & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik: Tests und Trends Band 12* (S. 61–81). Hogrefe.

de Boer, H., Bosker, R. J., & Van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*(1), 168–179. https://doi.org/10.1037/a0017289

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Bradley, L., & Bryant, P. E. (1978). Difficulties in auditory organisation as a possible cause of reading backwardness. *Nature, 271*, 746–747. https://doi.org/10.1038/271746a0

Braeuning, D., Hornung, C., Hoffmann, D., Lambert, K., Ugen, S., Fischbach, A., Schiltz, C., Hübner, N., Nagengast, B., & Moeller, K. (2021). Long-term relevance and interrelation of symbolic and non-symbolic abilities in mathematical-numerical development: Evidence from large-scale assessment data. *Cognitive Development, 58*, 101008. https://doi.org/10.1016/j.cogdev.2021.101008

Brambrink, J., Kunina-Habenicht, O., & Kuhn, J.-T. (2024). *Conflict quantity processing: A diffusion-model analysis in adults with mathematical learning disorder* [Posterpräsentation]. Vortrag auf dem 53. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Wien. https://doi.org/10.13140/RG.2.2.11734.87366

Brankaer, C., Ghesquière, P., & De Smedt, B. (2011). Numerical magnitude processing in children with mild intellectual disabilities. *Research in Developmental Disabilities, 32*(6), 2853–2859. https://doi.org/10.1016/j.ridd.2011.05.020

Brankaer, C., Ghesquière, P., & De Smedt, B. (2013). The development of numerical magnitude processing and its association with working memory in children with mild intellectual disabilities. *Research in Developmental Disabilities, 34*(10), 3361–3371. https://doi.org/10.1016/j.ridd.2013.07.001

Breitenbach, E. (2020). *Diagnostik: Eine Einführung* (1. Aufl.). Springer Fachmedien https://doi.org/10.1007/978-3-658-25150-5

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S.

Krauss, & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Waxmann.

Buchholtz, N. (2021). Voraussetzungen und Qualitätskriterien von Mixed-Methods-Studien in der mathematikdidaktischen Forschung. *Journal für Mathematik-Didaktik, 42*, 219–242. https://doi.org/10.1007/s13138-020-00173-0

Bulthé, J., Prinsen, J., Vanderauwera, J., Duyck, S., Daniels, N., Gillebert, C. R., Mantini, D., De Beeck, H. P. O., De Smedt, B. (2019). Multi-method brain imaging reveals impaired representations of number as well as altered connectivity in adults with dyscalculia. *NeuroImage, 190,* 289–302. https://doi.org/10.1016/j.neuroimage.2018.06.012

Bundesministerium der Justiz (o. J.). *Sozialgesetzbuch (SGB) - Achtes Buch (VIII) - Kinder- und Jugendhilfe - (Artikel 1 des Gesetzes v. 26. Juni 1990, BGBl. I S. 1163). § 35a Eingliederungshilfe für Kinder und Jugendliche mit seelischer Behinderung oder drohender seelischer Behinderung*. https://www.gesetze-im-internet.de/sgb_8/__35a.html

Busch, J., Schmidt, C., Studte, S., & Grube, D. (2018). Kognitive Merkmale rechenschwacher Kinder in Abhängigkeit vom Cut-off Kriterium. *Lernen und Lernstörungen, 8*(3), 167–178. https://doi.org/10.1024/2235-0977/a000258

Butterworth, B., & Kovas, Y. (2013). Understanding neurocognitive developmental disorders can improve education for all. *Science, 340*(6130), 300–305. https://doi.org/10.1126/science.1231022

Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: from brain to education. *Science, 332,* 1049–1053. https://doi.org/10.1126/science.1201536

Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language,*

Speech, and Hearing Services in Schools, 40(2), 161–173. https://doi.org/10.1044/0161-1461(2009/07-0099)

Cai, D., Zhang, L., Li, Y., Wei, W., & Georgiou, G. K. (2018). The Role of Approximate Number System in Different Mathematics Skills Across Grades. *Frontiers in Psychology, 9*, 1733. https://doi.org/10.3389/fpsyg.2018.01733

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica, 148*, 163–172. https://doi.org/10.1016/j.actpsy.2014.01.016

Chideridou-Mandari, A., Padeliadu, S., Karamatsouki, A., Sandravelis, A., & Karagiannidis, C. (2016). Secondary mathematics teachers: What they know and don't know about dyscalculia. *International Journal of Learning, Teaching and Educational Research, 15*(9), 84–98.

Chinn, S. (2023). *Checklist for Dyscalculia.* https://www.stevechinn.co.uk/dyscalculia/the-dyscalculia-checklist

Chodura, S., Kuhn, J.-T., & Holling, H. (2015). Interventions for children with mathematical difficulties. *Zeitschrift für Psychologie, 223(2),* 129–144. https://doi.org/10.1027/2151-2604/a000211

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have Gender Gaps in Math Closed? Achievement, Teacher Perceptions, and Learning Behaviors Across Two ECLS-K Cohorts. *AERA Open*, *2*(4). https://doi.org/10.1177/2332858416673617

Cirino, P. T. (2011). The interrelationships of mathematical precursors in kindergarten. *Journal of Experimental Child Psychology, 108*(4), 713–733. https://doi.org/10.1016/j.jecp.2010.11.004

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record*, *115*(6), 1–29.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(1)*, 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*(2), 141–146. https://doi.org/10.1037/0022-0663.78.2.141

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Decarli, G., Paris, E., Tencati, C., Nardelli, C., Vescovi, M., Surian, L., et al. (2020). Impaired large numerosity estimation and intact subitizing in developmental dyscalculia. *PloS One, 15*, e0244578. https://doi.org/10.1371/journal.pone.0244578

Decarli, G., Sella, F., Lanfranchi, S., Gerotto, G., Gerola, S., Cossu, G., & Zorzi, M. (2023). Severe developmental dyscalculia is characterized by core deficits in both symbolic and nonsymbolic number sense. *Psychological Science, 34*(1), 8–21. https://doi.org/10.1177/09567976221097947

Dehaene S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1–2), 1–42. https://doi.org/10.1016/0010-0277(92)90049-n

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition, 1*(1), 83–120.

Demaray, M. K., & Elliot, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8–24. https://doi.org/10.1037/h0088969

Desoete, A., Roeyers, H., & De Clercq, A. (2004). Children with Mathematics Learning Disabilities in Belgium. *Journal of Learning Disabilities, 37*(1), 50–61. https://doi.org/10.1177/00222194040370010601

Devine, A., Soltész, F., Nobes, A., Goswami, U., & Szűcs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction, 27,* 31–39. https://doi.org/10.1016/j.learninstruc.2013.02.004

Dögnitz, S. (2022). *Diagnostik von besonderen Rechenschwierigkeiten in der Sekundarstufe I.* Springer Spektrum. https://doi.org/10.1007/978-3-658-40071-2

Dollinger, S. (2013). *Diagnosegenauigkeit von ErzieherInnen und LehrerInnen: Einschätzung schulrelevanter Kompetenzen in der Übergangsphase.* Springer VS. https://doi.org/10.1007/978-3-658-01660-9

Dolores de Hevia, M., Girelli, L., & Vallar, G. (2006). Numbers and space: A cognitive illusion? *Experimental Brain Research, 168*, 254–264. https://doi.org/10.1007/s00221-005-0084-0

Dornheim, D. (2008). *Prädiktion von Rechenleistung und Rechenschwäche: Der Beitrag von Zahlen-Vorwissen und allgemein-kognitiven Fähigkeiten*. Logos.

Dowker, A. (2024). Developmental Dyscalculia in Relation to Individual Differences in Mathematical Abilities. *Children*, *11*(6), 623. https://doi.org/10.3390/children11060623

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Erlbaum.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*(3), 247–265. https://doi.org/10.1002/pits.20147

Education Scotland (o. J.). *Possible Indicators of Dyscalculia Teacher Checklist*. https://education.gov.scot/media/uehfve1v/nih325-possible-indicators-of-dyscalculia-teacher-checklist.pdf

Ehm, J.-H., Duzy, D., & Hasselhorn, M. (2011). Das akademische Selbstkonzept bei Schulanfängern: Spielen Geschlecht und Migrationshintergrund eine Rolle?. *Frühe Bildung, 0(0)*, 37–45. https://doi.org/10.1026/2191-9186/a000008

Ehm, J.-H., Hasselhorn, M., & Schmiedek, F. (2021). Der wechselseitige Einfluss von Selbstkonzept und Leistung bei Grundschulkindern im Lichte verschiedener längsschnittlicher Analysemethoden. *Zeitschrift für pädagogische Psychologie, 36*(4), 279–288. https://doi.org/10.25656/01:25235

Endlich, D., Lenhard, W., Marx, P., & Richter, T. (2024). Differential Switch Costs in Typically Achieving Children and Children With Mathematical Difficulties. *Journal of Learning Disabilities, 57*(4), 255–271. https://doi.org/10.1177/00222194231204619

Esser, G. (1992). Der langfristige Verlauf von Teilleistungsschwächen. In H.-C. Steinhausen (Hrsg.), *Hirnfunktionsstörungen und Teilleistungsschwächen* (S. 187–211). Springer. https://doi.org/10.1007/978-3-642-77072-2_12

Faix, A.-C., Peter-Koop, A., & Wild, E. (2023). Diagnostik, Förderung und Beratung bei Rechenschwäche: Wie können Selbstwirksamkeitsüberzeugungen angehender Lehrkräfte gesteigert werden?. *HLZ – Herausforderung Lehrer*innenbildung*, *6*(1), 130–145. https://doi.org/10.11576/hlz-6027

Falkai, P., & Wittchen, H. U. (2015). *Diagnostische Kriterien DSM-5: Deutsche Ausgabe. American Psychiatric Association*. Hogrefe.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102*, 453–462. https://doi.org/10.3200/JOER.102.6.453-462

Fischbach, A., Schuchardt, K., Brandenburg, J., Klesczewski, J., Balke-Melcher, C., Schmidt, C., Büttner, G., Grube, D., Mähler, C., & Hasselhorn, M. (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

Fischbach, A., Schuchardt, K., Mähler, C., & Hasselhorn, M. (2010). Zeigen Kinder mit schulischen Minderleistungen sozio-emotionale Auffälligkeiten? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*(4), 201–210. https://doi.org/10.1026/0049-8637/a000025

Fischer, U., Moeller, K., Cress, U., & Nuerk, H.-C. (2013). Interventions supporting children's mathematics school success: A meta-analytic review. *European Psychologist, 18*, 89–113. https://doi.org/10.1027/1016-9040/a000141

Fischer, U., Roesch, S., & Moeller, K. (2017). Diagnostik und Förderung bei Rechenschwäche: Messen wir, was wir fördern wollen? *Lernen und Lernstörungen, 6*(1), 25–38. https://doi.org/10.1024/2235-0977/a000160

Fischer, U., Rösch, S., Nuerk, H.-C., & Moeller, K. (2015). Erkennen von Rechenschwäche durch Lehrkräfte und Testungen im Klassenverband. *Lernen und Lernstörungen, 4*(4), 269–282. https://doi.org/10.1024/2235-0977/a000116

Förster, N., & Souvignier, E. (2017). Förderung diagnostischer Kompetenz durch Bereitstellung formativer Diagnostik. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 231–238). Waxmann.

Fritz, A., Ehlert, A., & Leutner, D. (2018). Arithmetische Konzepte aus kognitiv-entwicklungspsychologischer Sicht. *Journal für Mathematik-Didaktik, 39*(1), 7–41. https://doi.org/10.1007/s13138-018-0131-6

Fritz, A., & Ricken, G. (2005). Früherkennung von Kindern mit Schwierigkeiten beim Erwerb von Rechenfertigkeiten. In M. Hasselhorn, H. Marx, & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen. Trends und Tests Band 4* (S. 5–28). Hogrefe.

Fuchs, L. S., Fuchs, D., Seethaler, P. M., & Zhu, N. (2018). Three Frameworks for Assessing Responsiveness to Instruction as a Means of Identifying Mathematical Learning Disabilities. In A. Fritz-Stratmann, V. G. Haase, & P. Räsänen (Hrsg.), *International Handbook of Math Learning Difficulties: From the Lab to the Classroom* (S. 669–682). Springer. https://doi.org/10.1007/978-3-319-97148-3_39

Gabriel, T., Griepenburg, C., & Schuchardt, K. (2021). Grundschulkindern Lernstörungen erklären: Evaluation einer psychoedukativen Lehreinheit zur Aufklärung über Lernstörungen. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 70*, 316–332. https://doi.org/10.13109/prkk.2021.70.4.316

Gallit, F., Wyschkon, A., & Esser, G. (2017). Prädiktion von Rechenschwäche und Rechenstörung. *Frühförderung interdisziplinär, 36*(2), 63–81. https://doi.org/10.2378/fi2017.art06d

Gallit, F., Wyschkon, A., Poltz, N., Moraske, S., Kucian, K., von Aster, M., & Esser, G. (2018). Henne oder Ei: Reziprozität mathematischer Vorläufer und Vorhersage des Rechnens. *Lernen und Lernstörungen, 7*(2), 81–92. https://doi.org/10.1024/2235-0977/a000205

Gaupp, N., Zoelch, C., & Schumann-Hengsteler, R. (2004). Defizite numerischer Basiskompetenzen bei rechenschwachen Kindern der 3. und 4. Klassenstufe. *Zeitschrift für Pädagogische Psychologie, 18*(1), 31–42.

Geary D. C. (2013). Early Foundations for Mathematics Learning and Their Relations to Learning Disabilities. *Current Directions in Psychological Science*, *22*(1), 23–27. https://doi.org/10.1177/0963721412469398

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology, 77*, 236–263. https://doi.org/10.1006/jecp.2000.2561

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability. *Journal of Learning Disabilities, 45*(4), 291–307. https://doi.org/10.1177/0022219410392046.

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development, 78*, 1343–1359. https://doi.org/10.1111/j.1467-8624.2007.01069.x

Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology, 33*(3), 277–299.

Gebhardt, M., Kuhl, J., Wittich, C., & Wember, F. B. (2018). Inklusives Modell in der Lehramtsausbildung nach den Anforderungen der UN-BRK. In S. Hußmann & B. Welzel (Hrsg.), *DoProfiL – Dortmunder Profil für inklusionsorientierte Lehrerinnen- und Lehrerbildung* (S. 279–292). Waxmann. https://doi.org/10.25656/01:16573

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods, 43*(4), 981–986. https://doi.org/10.3758/s13428-011-0097-5

Gerardi, K., Goette, L., & Meier, S. (2013). Numerical ability predicts mortgage default. *Proceedings of the National Academy of Sciences, 110*(28), 11267–11271. https://doi.org/10.1073/pnas.1220568110

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early Identification and Interventions for Students With Mathematics Difficulties. *Journal of learning disabilities*, *38*(4), 293–304. https://doi.org/10.1177/00222194050380040301

Gliksman, Y., & Henik, A. (2019). Enumeration and alertness in developmental dyscalculia. *Journal of Cognition, 2*(1), 5. https://doi.org/10.5334/joc.55

Gloor, N. (2023). Erfassung von numerischen Kompetenzen im Kindergarten. *Lernen und Lernstörungen, 12*(4), 175–187. https://doi.org/10.1024/2235-0977/a000409

Gnas, J., Mack, E., & Preckel, F. (2022). When classmates influence teacher judgment accuracy of students' cognitive ability: Studying frame-of-reference effects in primary school. *Contemporary Educational Psychology, 69*, Article 102070. https://doi.org/10.1016/j.cedpsych.2022.102070

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. doi:10.1111/2041-210X.12504

Habermann, S., Donlan, C., Göbel, S. M., & Hulme, C. (2020). The critical role of Arabic numeral knowledge as a longitudinal predictor of arithmetic development. *Journal of experimental child psychology*, *193*, 104794. https://doi.org/10.1016/j.jecp.2019.10479

Haberstroh, S., & Schulte-Körne, G. (2019). Diagnostik und Behandlung der Rechenstörung. *Deutsches Ärzteblatt*, *116*(7), 107–114 .https://doi.org/10.3238/arztebl.2019.0107

Haberstroh, S., & Schulte-Körne, G. (2022). The Cognitive Profile of Math Difficulties: A Meta-Analysis Based on Clinical Criteria. *Frontiers in Psychology, 13*, 842391. https://doi.org/10.3389/fpsyg.2022.842391

Haffner, J., Baro, K., Parzer, P., & Resch, F. (2005). *Heidelberger Rechentest (HRT 1–4). Erfassung mathematischer Basiskompetenzen im Grundschulalter*. Hogrefe.

Hakkarainen, A., Holopainen, L., & Savolainen, H. (2013). Mathematical and reading difficulties as predictors of school achievement and transition to secondary education. *Scandinavian Journal of Educational Research, 57*(5), 488–506. https://doi.org/10.1080/00313831.2012.696207

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in nonverbal number acuity correlate with maths achievement. *Nature, 455*(7213), 665–668. https://doi.org/10.1038/nature07246

Hartung, N., Schurig, M., Vossen, A., & Gebhardt, M. (2021). Pädagogische Diagnostik im Rahmen des RTI-Modells. In J. Kuhl, A. Vossen, N. Hartung, & C. Wittich (Hrsg.), *Evidenzbasierte Förderung bei Lernschwierigkeiten in der Grundschule* (S. 28–39). Ernst Reinhardt.

Hasselhorn, M., & Schuchardt, K. (2006). Lernstörungen: Eine kritische Skizze zur Epidemiologie. *Kindheit und Entwicklung, 15*(4), 208–215. https://doi.org/10.1026/0942-5403.15.4.208

Helmke, A. (2017). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (7. Aufl.). Klett-Kallmeyer.

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition, 10*, 389–395. https://doi.org/10.3758/bf03202431

Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3., vollst. überarb. u. erw. Aufl.). utb.

Hilkenmeier, J., Ricken, G., Nolte, M., & Ehlers, A. (2020). Lerntherapie in Schule. *Lernen und Lernstörungen, 9*(4), 213–224. https://doi.org/10.1024/2235-0977/a000315

Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research*, *59*(3), 297–313. https://doi.org/10.3102/00346543059003297

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social psychology of education*, *20*, 471-490. https://doi.org/10.1007/s11218-017-9384-z

Holzer, N., Lenart, F., & Schaupp, H. (2017). *Eggenberger Rechentest für Jugendliche und Erwachsene (ERT JE)*. Hogrefe.

Holzer, N., Schaupp, H., & Lenart, F. (2010). *Eggenberger Rechentest 3+ (ERT 3+): Diagnostikum für Dyskalkulie für das Ende der 3. Schulstufe bis zur Mitte der 4. Schulstufe*. Hogrefe.

Hornung, C., Schiltz, C., Brunner, M., & Martin, R. (2014). Predicting first-grade mathematics achievement: The contributions of domain-general cognitive

abilities, nonverbal number sense, and early number competence. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00272

Hosenfeld, I., Helmke, A., & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (Zeitschrift für Pädagogik, Beiheft 45, S. 65–82). Beltz. https://doi.org/10.25656/01:3939

Huguet, P., & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology, 99*(3), 545–560. https://doi.org/10.1037/0022-0663.99.3.545

van Hulst, B. M., de Zeeuw, P., Vlaskamp, C., Rijks, Y., Zandbelt, B. B., & Durston, S. (2018). Children with ADHD symptoms show deficits in reactive but not proactive inhibition, irrespective of their formal diagnosis. *Psychological Medicine*. https://doi.org/10.1017/S0033291718000107

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik* (6. Aufl.). Beltz.

Ise, E., Dolle, K., Pixner, S., & Schulte-Körne, G. (2012). Effektive Förderung rechenschwacher Kinder. *Kindheit und Entwicklung, 21*(3), 181–192. https://doi.org/10.1026/0942-5403/a000083

Jacobs, C., & Petermann, F. (2003). Dyskalkulie - Forschungsstand und Perspektiven. *Kindheit und Entwicklung*, *12*(4), 197–211. https://doi.org/10.1026//0942-5403.12.4.197

Jacobs, C., & Petermann, F. (2012). *Diagnostik von Rechenstörungen*. Hogrefe.

Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (1999). *Bielefelder Screening zur*

Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC). Hogrefe.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development, 74*(3), 834–850. https://doi.org/10.1111/1467-8624.00571

Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews, 15*(1), 60–68. https://doi.org/10.1002/ddrr.46

Jordan, N. C., Rinne, L., & Hansen, N. (2018). Mathematical Learning and Its Difficulties in the United States: Current Issues in Screening and Intervention. In A. Fritz-Stratmann, V. G. Haase & P. Räsänen (Hrsg.), *International Handbook of Mathematical Learning Difficulties* (S. 183–199). Springer. https://doi.org/10.1007/978-3-319-97148-3_12

Jussim, L., & Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of students' achievement. *Journal of Personality and Social Psychology, 63*(6), 947–961. https://doi.org/10.1037/0022-3514.63.6.947

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5

Karing, C., & Artelt, C. (2014). Urteilsgenauigkeit von Lehrer(inne)n im emotionalmotivationalen Bereich und im Leistungsbereich. In M. Mudiappa & C.

Artelt (Hrsg.) *BiKS – Ergebnisse aus den Längsschnittstudien. Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich* (S. 111–118). University of Bamberg Press.

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern.* Waxmann.

Karst, K., & Bonefeld, M. (2020). Stereotype, Urteile und Urteilsakkuratheit von Lehrkräften: Eine Zusammenschau im Rahmen des Heterogenitätsdiskurses. In S. Glock & H. Kleen (Hrsg.), *Stereotype in der Schule* (S. 281–308). Springer VS. https://doi.org/10.1007/978-3-658-27275-3_9

Karst, K., Hartig, J., Kaiser, J., & Lipowsky, F. (2017). Mehrebenenmodelle als Werkzeuge zur Analyse diagnostischer Kompetenz von Lehrkräften – ein lineares Mischmodell (LMM). In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 153–174). Waxmann

Karst, K., Schoreit, E., & Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern. *Zeitschrift für pädagogische Psychologie, 28*, 237–248. https://doi.org/10.1024/1010-0652/a000133

Kaufmann, E. (2020). How accurately do teachers judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Kaufmann, L., & von Aster, M. (2012). Diagnostik und Intervention bei Rechenstörung. *Deutsches Ärzteblatt, 109*(45), 767–778.

Kaufmann, S., & Wessolowski, S. (2021). *Rechenstörungen – Diagnose und Förderbausteine* (8. Aufl.). Klett-Kallmeyer.

Kenny, D. T., & Chekaluk, E. (1993). Early Reading Performance: A Comparison of Teacher-Based and Test-Based Assessments. *Journal of Learning Disabilities, 26*(4), 227–236. https://doi.org/10.1177/002221949302600403

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment, 30*(2), 148–159. https://doi.org/10.1177/0734282911412722

Kißler, C., Schwenk, C., & Kuhn, J.-T. (2020). Zur Additivität kognitiver Defizitprofile bei komorbiden Lernstörungen. *Lernen und Lernstörungen, 10*(2), 89–101. https://doi.org/10.1024/2235-0977/a000310

Kißler, C., Schwenk, C., & Kuhn, J.-T. (2021). Two Dyscalculia Subtypes With Similar, Low Comorbidity Profiles: A Mixture Model Analysis. *Frontiers in Psychology, 12*, 1828. https://doi.org/10.3389/fpsyg.2021.589506

Klapproth, F., & von der Lippe, H. (2024). A Gender Bias in Curriculum-Based Measurement across Content Domains: Insights from a German Study. *Education Sciences*, *14*(1), 76. https://doi.org/10.3390/educsci14010076

Kohn, J., Wyschkon, A., & Esser, G. (2013). Psychische Auffälligkeiten bei Umschriebenen Entwicklungsstörungen: Gibt es Unterschiede zwischen Lese-Rechtschreib- und Rechenstörungen? *Lernen und Lernstörungen, 1*, 7–20. https://doi.org/10.1024/2235-0977/a000027

Korhonen, J., Linnanmäki, K., & Aunio, P. (2014). Learning difficulties, academic well-being and educational dropout: A person-centred approach. *Learning and Individual Differences, 31,* 1–10. https://doi.org/10.1016/j.lindif.2013.12.011

Koriakin, T. A., McCurdy, M. D., Pritchard, A. E., Zabel, T. A., & Jacobson, L. A. (2019). Screening for learning difficulty using teacher ratings on the Colorado Learning Difficulties Questionnaire. *Learning Disabilities: A Multidisciplinary Journal*, *24*(1), 55–63. https://doi.org/10.18666/LDMJ-2019-V24-I1-9355

Kosel, C., Bauer, E., & Seidel, T. (2024). Where experience makes a difference: teachers' judgment accuracy and diagnostic reasoning regarding student learning characteristics. *Frontiers in Psychology*, *15*, 1278472. https://doi.org/10.3389/fpsyg.2024.1278472

Krämer, S., & Zimmermann, F. (2020). Zum Einfluss von störendem Schülerverhalten im Unterricht auf Leistungsbeurteilungen: Explizite Einschätzungen und experimentelle Befunde. *Zeitschrift für Pädagogische Psychologie, 34*(2), 99–115. https://doi.org/10.1024/1010-0652/a000250

Krajewski, K. (2005). Vorschulische Mengenbewusstheit von Zahlen und ihre Bedeutung für die Früherkennung von Rechenschwäche. In M. Hasselhorn, H. Marx, & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen. Trends und Tests Band 4* (S. 49–70). Hogrefe.

Krajewski, K. (2008). *Vorhersage von Rechenschwäche in der Grundschule* (2. Korr. Aufl.). Verlag Dr. Kovač

Krajewski, K., & Schneider, W. (2006). Mathematische Vorläuferfertigkeiten im Vorschulalter und ihre Vorhersagekraft für die Mathematikleistungen bis zum Ende der Grundschulzeit. *Zeitschrift für Psychologie in Erziehung und Unterricht, 53*(4), 246–262.

Kuckartz, U. (2014). *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren.* Springer VS. https://doi.org/10.1007/978-3-531-93267-5

Kuhl, J., Hußmann, A., & Schulze, S. (2024). Fördern und Unterstützen – Der Einsatz von digitalen Medien im inklusiven Unterricht. In V. Heitplatz & L. Wilkens (Hrsg.), *Die Rehabilitationstechnologie im Wandel: Eine Mensch-Technik-Umwelt-Betrachtung – Eine Festschrift anlässlich des 20-jährigen Bestehens des Fachgebiets Rehabilitationstechnologie und der Verabschiedung von Prof. Dr.-Ing. Christian Bühler* (S. 375–390). Eldorado. https://d-nb.info/1330532457/34

Kuhn, J.-T. (2017). Rechenschwäche – eine interdisziplinäre Einführung. In A. Fritz, S. Schmidt, & G. Ricken (Hrsg.), *Handbuch Rechenschwäche: Lernwege, Schwierigkeiten und Hilfen bei Dyskalkulie* (3. Aufl., S. 14–29). Beltz.

Kuhn, J.-T., Ise, E., Raddatz, J., Schwenk, C., & Dobel, C. (2016). Basic numerical processing, calculation, and working memory in children with dyscalculia and/or ADHD symptoms. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 44*(5), 365–375. https://doi.org/10.1024/1422-4917/a000450

Kuhn, J.-T., Raddatz, J., Holling, H., & Dobel, C. (2013). Dyskalkulie vs. Rechenschwäche: Basisnumerische Verarbeitung in der Grundschule. *Lernen und Lernstörungen*, *2*(4), 229–247. https://doi.org/10.1024/2235-0977/a000044

Kuhn, J-T., & Schwenk, C. (2018). Onlinebasierte Diagnostik mathematischer Kompetenzen: Möglichkeiten und Grenzen. *Lernen und Lernstörungen, 7*(4), 231–235 https://doi.org/10.1024/2235-0977/a000232

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

Kuhn, J.-T., Schwenk, C., Souvignier, E., & Holling, H. (2019). Arithmetische Kompetenz und Rechenschwäche am Ende der Grundschulzeit: Die Rolle statusdiagnostischer und lernverlaufsbezogener Prädiktoren. *Empirische Sonderpädagogik, 11*(2), 95–117. https://doi.org/10.25656/01:17773

Kultusministerkonferenz (KMK) (2004). *Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 07.10.2022).* https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Kultusministerkonferenz (KMK). (2007). *Grundsätze zur Förderung von Schülerinnen und Schülern mit besonderen Schwierigkeiten im Lesen und Rechtschreiben oder*

im Rechnen. Beschluss der Kultusministerkonferenz vom 04.12.2003 i. d. F. vom 15.11.2007. https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/Beschluesse_Veroeffentlichungen/allg_Schulwesen/304_Legasthenie.pdf

Lamb, S., Anderson, S., Schulze, S., Zimmermann, J. S., Schröter, A., Hußmann, A., Kuhn, J.-T., & Steinmayr, R. (2023a). Diversitätswissen von Lehramtsstudierenden: Ergebnisse einer Fragebogenerhebung an der Technischen Universität Dortmund. In S. Hußmann & B. Welzel (Hrsg.), *DoProfiL 2.0 – Das Dortmunder Profil für inklusionsorientierte Lehrerinnen- und Lehrerbildung* (S. 263–277). Waxmann. https://doi.org/10.25656/01:28145

Lamb, S., Schulze, S., Anderson, S., Zimmermann, J. S., Schröter, A., Hußmann, A., Kuhn, J.-T., & Steinmayr, R. (2023b). Prädiktive Faktoren des Diversitätswissens von Lehramtsstudierenden. In S. Hußmann & B. Welzel (Hrsg.), *DoProfiL 2.0 – Das Dortmunder Profil für inklusionsorientierte Lehrerinnen- und Lehrerbildung* (S. 278–290). Waxmann. https://doi.org/10.25656/01:28145

Lamb, S., Krieger, F., & Kuhn, J.-T. (2024a). Delayed development of basic numerical skills in children with developmental dyscalculia. *Frontiers in Psychology, 14*, 1187785. https://doi.org/10.3389/fpsyg.2023.1187785

Lamb, S., Schulz, A.-K., & Kuhn, J.-T. (2024b). Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule. *Lernen und Lernstörungen, 13*(4), 165–177. https://doi.org/10.1024/2235-0977/a000456

Lamb, S., Schulz, A.-K., & Kuhn, J.-T. (2025). Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule. *Empirische Sonderpädagogik, 17*(1), 35–49. https://doi.org/10.25656/01:34364

Landerl, K. (2013). Development of numerical processing in children with typical and dyscalculic arithmetic skills-a longitudinal study. *Frontiers in Psychology*, *4*, 459. https://doi.org/10.3389/fpsyg.2013.00459

Landerl, K. (2019). Neurocognitive Perspective on Numerical Development. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International Handbook of Mathematical Learning Difficulties* (pp. 9–24). Springer. https://doi.org/10.1007/978-3-319-97148-3_2

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition, 93*(2), 99–125. https://doi.org/10.1016/j.cognition.2003.11.004

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology, 103*(3), 309–324. https://doi.org/10.1016/j.jecp.2009.03.006

Landerl, K., & Kölle, C. (2009). Typical and atypical development of basic numerical skills in elementary school. *Journal of Experimental Child Psychology, 103*(4), 546–565. https://doi.org/10.1016/j.jecp.2008.12.006

Landerl, K., & Moll, K. (2010). Comorbidity of learning disorders: Prevalence and familial transmission. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 51*(3), 287–294. https://doi.org/10.1111/j.1469-7610.2009.02164.x

Landerl, K., Vogel, S., & Kaufmann, L. (2022). *Dyskalkulie: Modelle, Diagnostik, Intervention*. Ernst Reinhardt.

Landesregierung Nordrhein-Westfalen. (2020). *Lese-Rechtschreib-Schwäche und Dyskalkulie an den Schulen in NRW. Antwort der Landesregierung auf die Kleine Anfrage 4529 vom 7. Oktober 2020*.

https://www.landtag.nrw.de/portal/WWW/dokumentenarchiv/Dokument/MMD17-11816.pdf

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363–374. https://doi.org/10.2307/2529786

Laursen, B., Richmond, A., Kiuru, N., Lerkkanen, M. K., & Poikkeus, A. M. (2021). Off on the wrong foot: Task avoidance at the outset of primary school anticipates academic difficulties and declining peer acceptance. *European Journal of Developmental Psychology*, *19*(4), 601–615. https://doi.org/10.1080/17405629.2021.1936491

Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF). (2024). Dyskalkulie, Rechenstörung, Rechenschwäche: Erkennen, Testen, Nachteilsausgleich. *Deutscher Bildungsserver*. https://www.bildungsserver.de/dyskalkulie-rechenstoerung-rechenschwaeche-794-de.html

Lenart, F., Schaupp, H., & Holzer, N. (2013). *Eggenberger Rechentest 0+ (ERT 0+): Diagnostikum für Dyskalkulie-Disposition für das Ende des Kindergartenalters bis Mitte der 1. Schulstufe.* Huber.

Lerkkanen, M. K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J. E. (2005). Mathematical performance predicts progress in reading comprehension among 7-year olds. *European Journal of Psychology of Education, 20*, 121–137. https://doi.org/10.1007/BF03173503

Lindberg, S., Linkersdörfer, J., Ehm, J. H., Hasselhorn, M., & Lonnemann, J. (2013). Gender differences in children's math self-concept in the first years of elementary school. *Journal of Education and Learning, 2*(3). https://doi.org/10.5539/jel.v2n3p1

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for Explaining Teachers' Diagnostic Judgements by Cognitive Modeling (DiaCoM). *Teaching and Teacher Education, 91,* 103059. https://doi.org/10.1016/j.tate.2020.103059

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen*. University of Bamberg Press. https://fis.uni-bamberg.de/server/api/core/bitstreams/07e40c0c-aa6e-4d53-942e-9696614f6267/content

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie, 23*(3–4), 211–222. https://doi.org/10.1024/1010-0652.23.34.211

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science, 17*(5), 714–726. https://doi.org/10.1111/desc.12152

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mack, E., Gnas, J., Vock, M., & Preckel, F. (2023). The domain-specificity of elementary school teachers' judgment accuracy. *Contemporary Educational Psychology, 72*, 102142. https://doi.org/10.1016/j.cedpsych.2022.102142

Mähler, C. (2021). Diagnostik von Lernstörungen: Zeit zum Umdenken. *Zeitschrift für Pädagogische Psychologie, 35*(4), 217–227. https://doi.org/10.1024/1010-0652/a000291

Mähler, C., & Schuchardt, K. (2016). Working memory in children with specific learning disorders and/or attention deficits. *Learning and Individual Differences, 49*, 341–347. https://doi.org/10.1016/j.lindif.2016.05.007

Mammarella, I. C., Toffalini, E., Caviola, S., Colling, L., & Szűcs, D. (2021). No evidence for a core deficit in developmental dyscalculia or mathematical learning disabilities. *Journal of Child Psychology and Psychiatry, 62(6)*, 704–714. https://doi.org/10.1111/jcpp.13397

Marx, P., & Lenhard, W. (2011). Diagnostische Merkmale von Screeningverfahren. In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen* (S. 68–84). Hogrefe.

Mathematisches Institut zur Behandlung der Rechenschwäche/Dyskalkulie (o. J.). *Symptomfragebogen zum Erkennen von elementaren Lernschwierigkeiten im Grundlagenbereich der Mathematik (Rechenschwäche/Dyskalkulie) für Lehrer, Kinder- und Jugendärzte, Psychologen, Kliniken und Beratungsstellen (Grund-, Förder- und weiterführende Schulen bis Klasse 5).* https://www.rechenschwaeche.de/Arbeitsweise/Symptome_Fragebogen_interaktiv.pdf

Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 633–648). Springer VS. https://doi.org/10.1007/978-3-658-21308-4_42

Mazzocco, M. M. M. (2007). Defining and differentiating mathematical learning disabilities and difficulties. In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 29–47). Paul H. Brookes Publishing Co

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82(4)*, 1224–1237. https://doi.org/10.1111/j.1467-8624.2011.01608.x

McCaskey, U., von Aster, M. G., Maurer, U., Martin, E., O'Gorman Tuura, R., & Kucian, K. (2017). Longitudinal Brain Development of Numerical Skills in Typically Developing Children and Children with Developmental Dyscalculia. *Frontiers in Human Neuroscience, 11*, 629. https://doi.org/10.3389/fnhum.2017.00629

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223

Meier, P., McCaskey, U., & Kucian, K. (2021). Typical errors made by children and adolescents with developmental dyscalculia. *Lernen und Lernstörungen, 10*(3), 135–150. https://doi.org/10.1024/2235-0977/a000348

Menon, V. (2016). Working memory in children's math learning and its disruption in dyscalculia. *Current Opinion in Behavioral Sciences, 10,* 125–132. https://doi.org/10.1016/j.cobeha.2016.05.014

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., & Schulte Körne, G. (2014). Specific Learning Disorder: Prevalence and Gender Differences. *PLOS ONE, 9*(7), e103537. https://doi.org/10.1371/journal.pone.0103537

Moosbrugger, H., & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien"). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 13–38). Springer. https://doi.org/10.1007/978-3-662-61532-4_2

Moraske, S., Wyschkon, A., Poltz, N., Kohn, J., Kucian, K., von Aster, M., & Esser, G. (2018). Indizierte Prävention von Rechenschwächen im Vorschulalter: Effekte bis Klasse 3. *Lernen und Lernstörungen, 7*(4), 175–186. https://doi.org/10.1024/2235-0977/a000224

Moser Opitz, E. M., & Bern, P. H. (2008). *Rechenschwäche erfassen: Screening für die Schuljahre 4–8*. Universitätsbibliothek Dortmund. https://eldorado.tu-dortmund.de/bitstream/2003/31691/1/141.pdf

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology, 103*(4), 490–502. https://doi.org/10.1016/j.jecp.2009.02.003

Muntoni, F., Dunekacke, S., Heinze, A., & Retelsdorf, J. (2020). Geschlechtsspezifische Erwartungseffekte in Mathematik. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *51*(2), 84–96. https://doi.org/10.1026/0049-8637/a000212

Murphy, M. M., Mazzocco, M. M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive Characteristics of Children With Mathematics Learning Disability (MLD) Vary as a Function of the Cutoff Criterion Used to Define MLD. *Journal of learning disabilities*, *40*(5), 458–478. https://doi.org/10.1177/00222194070400050901

Mussolin, C., Mejias, S., & Noël, M. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition, 115*, 10–25. https://doi.org/10.1016/j.cognition.2009.10.006

Mutlu, Y., Çalışkan, E. F., & Yasul, A. F. (2022). We asked teachers: Do you know what dyscalculia is? *International Online Journal of Primary Education, 11*(2), 361–378. https://doi.org/10.55020/iojpe.1067560

Noël, M.-P., & Rousselle, L. (2011). Developmental changes in the profiles of dyscalculia: an explanation based on a double exact-and-approximate number

representation model. *Frontiers in Human Neuroscience*, *5*, 165. https://doi.org/10.3389/fnhum.2011.00165

Olczyk, M., Gentrup, S., Schneider, T., Volodina, A., Casoni, V. P., Washbrook, E., & Waldfogel, J. (2023). Teacher judgements and gender achievement gaps in primary education in England, Germany, and the US. *Social Science Research, 116*, 102938. https://doi.org/10.1016/j.ssresearch.2023.102938

Olsson, L., Östergren, R., & Träff, U. (2016). Developmental dyscalculia: A deficit in the approximate number system or an access deficit? *Cognitive Development, 39*(2), 154–167. https://doi.org/10.1016/j.cogdev.2016.04.006

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* National Research and Development Centre for Adult Literacy and Numeracy https://core.ac.uk/download/pdf/111025087.pdf

Peng, P., Cuicui, W., & Jessica, N. (2018). Understanding the Cognition Related to Mathematics Difficulties: A Meta-Analysis on the Cognitive Deficit Profiles and the Bottleneck Theory. *Review of Educational Research, 88*(3), 434–476. https://doi.org/10.3102/0034654317753350

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

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin, 146*(7), 595–634. https://doi.org/10.1037/bul0000231

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

socioeconomic status. *Psychological Bulletin, 145*(2), 189–236. https://doi.org/10.1037/bul0000182

Peter-Koop, A., Wollring, B., Spindeler, B., & Grüßing, M. (2007). *Elementarmathematisches Basisinterview (EMBI)*. Mildenberger.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., et al. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition, 116*(1), 33–41. https://doi.org/10.1016/j.cognition.2010.03.012

Pielsticker, F., Pielsticker, C., & Witzke, I. (2024). Darstellung neurowissenschaftlicher Ergebnisse zu besonderen Schwierigkeiten beim Mathematiklernen – eine theoriegeleitete Diskussion. In F. Dilling, K. Holten, & I. Witzke (Hrsg.), *Interdisziplinäres Forschen und Lehren in den MINT-Didaktiken. MINTUS – Beiträge zur mathematisch-naturwissenschaftlichen Bildung* (S. 215–247). Springer Spektrum. https://doi.org/10.1007/978-3-658-43873-9_10

Praetorius, A.-K., Karst, K., & Lipowsky, F. (2012). Diagnostische Kompetenz von Lehrkräften: Aktueller Forschungsstand, unterrichtspraktische Umsetzbarkeit und Bedeutung für den Unterricht. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht. Implikationen für Theorie und Praxis* (S. 115–146). Klinkhardt.

Praetorius, A.-K., & Südkamp, A. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), Diagnostische Kompetenz von Lehrkräften (S. 13–17). Waxmann.

Price, G. R., Holloway, I., Räsänen, P., Vesterinen, M., & Ansari, D. (2007). Impaired parietal magnitude processing in developmental dyscalculia. *Current Biology, 17*(24), R1041–R1042. https://doi.org/10.1016/j.cub.2007.10.013

Raddatz, J., Kuhn, J.-T., Holling, H., Moll, K., & Dobel, C. (2017). Comorbidity of Arithmetic and Reading Disorder: Basic Number Processing and Calculation in Children With Learning Impairments. *Journal of Learning Disabilities, 50*(3), 298–308. https://doi.org/10.1177/0022219415620899

Reid, E. E., Diperna, J. C., Missall, K., & Volpe, R. J. (2014). Reliability and structural validity of the teacher rating scales of early academic competence. *Psychology in the Schools, 51*(6), 535–553. https://doi.org/10.1002/pits.21769

Reigosa-Crespo, V., Valdés-Sosa, M., Butterworth, B., Estévez, N., Rodríguez, M., Santos, E., Torres, P., Suárez, R., & Lage, A. (2012). Basic numerical capacities and prevalence of developmental dyscalculia: The Havana Survey. *Developmental Psychology, 48*(1), 123–135. https://doi.org/10.1037/a0025356

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

Ricken, G. (2014). Lerntherapie geht in die Schule: Überlegungen zu einer Ressourcenverknüpfung. *Lernen und Lernstörungen, 3*(3), 179–184. https://doi.org/10.1024/2235-0977/a000074

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, *77*. https://doi.org/10.1186/1471-2105-12-77

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*(4), 1262–1281. https://doi.org/10.1037/a0035073

Rogalla, M., & Vogt, F. (2008). Förderung adaptiver Lehrkompetenz: Eine Interventionsstudie. *Unterrichtswissenschaft. Zeitschrift für Lernforschung, 36*(1), 17–36.

Roick, T., Gölitz, D., & Hasselhorn, M. (2011). *Kettenrechner für dritte und vierte* Klassen (KR 3–4). Hogrefe.

Roick, T., & Hasselhorn, M. (2005). Der Kettenrechner 3–4. Zusätzliche Differenzierung durch komplexe arithmetische Faktenaufgaben. In M. Hasselhorn, W. Schneider, & U. Trautwein (Hrsg.), *Diagnostik von Mathematikleistungen. Tests und Trends Band 4* (S. 233–250). Hogrefe.

Rousselle, L., & Noël, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs. non-symbolic number magnitude processing. *Cognition, 102*(3), 361–395. https://doi.org/10.1016/j.cognition.2006.01.005

Rubinsten, O., & Tannock, R. (2010). Mathematics anxiety in children with developmental dyscalculia. *Behavioral and Brain Functions, 6*, 46. https://doi.org/10.1186/1744-9081-6-46

Saga, M., Rkhaila, A., Oubaha, D., & Ounine, K. (2022). The impact of anxiety and life quality on the mathematical performance of dyscalculic middle school children. *Applied Neuropsychology: Child*, *12*(4), 318–326. https://doi.org/10.1080/21622965.2022.2105146

Salvador, L., Moura, R., Wood, G., & Haase, V. G. (2019). Cognitive heterogeneity of math difficulties: A bottom-up classification approach. *Journal of Numerical Cognition, 5*(1), 55–85. https://doi.org/10.5964/jnc.v5i1.60

Schabmann, A., & Schmidt, B. M. (2009). Sind Lehrer gute Lese-Rechtschreibdiagnostiker? Der Einfluss von problematischem Schülerverhalten

auf die Einschätzungen der Lesekompetenz durch Lehrkräfte. *Heilpädagogische Forschung, 35*(3), 133–145.

Schleifer, P., & Landerl, K. (2011). Subitizing and counting in typical and atypical development. *Developmental Science, 14*(2), 280–291. https://doi.org/10.1111/j.1467-7687.2010.00976.x

Schmitterer, A. M., & Brod, G. (2021). Which Data Do Elementary School Teachers Use to Determine Reading Difficulties in Their Students? *Journal of Learning Disabilities, 54*(5), 349–364. https://doi.org/10.1177/0022219420981990

Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science, 20*(3), e12372. https://doi.org/10.1111/desc.12372

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Peter Lang.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerinnen- und Lehrerbildung, 31*(2), 154–165. https://doi.org/10.25656/01:13843

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik, 1*(1), 27–52.

Schrader, F.-W., Helmke, A., Hosenfeld, I., Halt, A. C., & Hochweber, J. (2006). Komponenten der Diagnosegenauigkeit von Lehrkräften: Ergebnisse aus Vergleichsarbeiten in der Grundschule. In F. Eder, A. Gastager & F. Hoffmann (Hrsg.), *Qualität durch Standards*? *Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 265–278). Waxmann.

Schrader, F.-W., & Praetorius, A.-K. (2018). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost, J. R. Sparfeldt & S. R. Buch (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (5., überarb. u. erw. Aufl., S. 92–98). Beltz.

Schuchardt, K., Griepenburg, C., & Mähler, C. (2021). Umgang mit Lernstörungen in der Klasse. Gegen Stigmatisierung und Ausgrenzung. In C. Bätge, P. Cloos, F. Gerstenberg, & K. Riechers (Hrsg.), *Inklusive Bildungsforschung der frühen Kindheit: Empirische Perspektiven und multidisziplinäre Zugänge* (S. 302–319). Beltz Juventa.

Schuchardt, K., Maehler, C., & Hasselhorn, M. (2008). Working Memory Deficits in Children With Specific Learning Disorders. *Journal of Learning Disabilities, 41*(6), 514–523. https://doi.org/10.1177/0022219408317856

Schuchardt, K., Piekny, J., Grube, D., & Mähler, C. (2014). Einfluss kognitiver Merkmale und häuslicher Umgebung auf die Entwicklung numerischer Kompetenzen im Vorschulalter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*(1), 24–34. https://doi.org/10.1026/0049-8637/a000099

Schütky, R. (2022). Der Einfluss nicht-kognitiver Merkmale (Selbstkonzept, Stereotypen und Stereotype Threat) auf Mathematikleistungen im Bereich der Größen im Verlauf der Grundschulzeit. *Mathematik im Unterricht, 13*, 31–46.

Schulte-Körne, G. (2021). Verpasste Chancen: Die neuen diagnostischen Leitlinien zur Lese-, Rechtschreib- und Rechenstörung der ICD-11. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 49*(6), 463–467. https://doi.org/10.1024/1422-4917/a000791

Schulte-Körne, G., & Haberstroh, S. (2018). *S3-Leitlinie: Diagnostik und Behandlung der Rechenstörung*, https://register.awmf.org/assets/guidelines/028-046l_S3_Rechenst%C3%B6rung-2018-03_1-abgelaufen.pdf

Schulte-Körne, G., & Hasselhorn, M. (2025). *LONDI-Hilfssystem. Unterstützung bei Diagnostik und Förderung von Lernstörungen bei Kindern*. https://hilfssystem.londi.de/

Schulz, F., Wyschkon, A., Gallit, F., Poltz, N., Moraske, S., Kucian, K., & Esser, G. (2018). Rechenprobleme bei Grundschulkindern: Persistenz und Schulerfolg nach fünf Jahren. *Lernen und Lernstörungen, 7*(2), 67–80. https://doi.org/10.1024/2235-0977/a000206

Schwenk, C., Sasanguie, D., Kuhn, J.-T., Kempe, S., Doebler, P., & Holling, H. (2017). (Non-)symbolic magnitude processing in children with mathematical difficulties: A meta-analysis. *Research in Developmental Disabilities, 64*, 152–167. https://doi.org/10.1016/j.ridd.2017.03.003

Selter, C., Walter, D., Heinze, A., Brandt, J., & Jentsch, A. (2020). Mathematische Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky, & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 25–56). Waxmann. https://doi.org/10.31244/9783830993193

Semeraro, C., Coppola, G., Taurino, A., & Cassibba, R. (2020). Understanding the impact of diagnosis: Emotional well-being, peers and teachers. In D. Lucangeli (Hrsg.), *Understanding Dyscalculia* (S. 94–119). Routledge. https://doi.org/10.4324/9780429423581-6

Shalev, R. S., Auerbach, J., Manor, O., & Gross-Tsur, V. (2000). Developmental dyscalculia: Prevalence and prognosis. *European Child & Adolescent Psychiatry, 9*(Suppl 2), 58–64. https://doi.org/10.1007/s007870070009

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237–243. https://doi.org/10.1111/1467-9280.02438

Sikora, S., & Voß, S. (2018). *Mathematikunterricht in der inklusiven Grundschule* (1. Aufl.). Kohlhammer

Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly, 32*(4), 435–448. https://doi.org/10.1037/spq0000197

Skagerlund, K., & Träff, U. (2014). Development of magnitude processing in children with developmental dyscalculia: Space, time, and number. *Frontiers in Psychology, 5*, 675. https://doi.org/10.3389/fpsyg.2014.00675

Skagerlund, K., & Träff, U. (2016). Number Processing and Heterogeneity of Developmental Dyscalculia: Subtypes With Different Cognitive Profiles and Deficits. *Journal of Learning Disabilities, 49*(1), 36–50. https://doi.org/10.1177/0022219414522707

De Smedt, B., Noël, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in neuroscience and education, 2*(2), 48-55. https://doi.org/10.1016/j.tine.2013.06.001

Sousa, P., Dias, P. C., & Cadime, I. (2017). Predictors of primary school teachers' knowledge about developmental dyscalculia. *European Journal of Special Needs Education, 32*(2), 204–220. https://doi.org/10.1080/08856257.2016.1216635

Special Educational Needs and Disability Independent Support Service (SENDISS) (o. J.). *Dyscalculia:Checklist.*

https://sendiss.co.uk/downloads/dyscalculia/3_DYSCALCULIA_CHECKLIST.pdf

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19*(1/2), 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Statistisches Bundesamt. (2022). *Bevölkerung nach Geschlecht.* https://www.zensus2022.de/DE/Ergebnisse-des-Zensus/Sonderauswertung_Bevoelkerung_nach_Geschlecht.html

Steinmayr, R., Heyder, A., & Tometten, L. (2022). DiWi. Test zum Wissen über verschiedene Diversitätsbereiche von (angehenden) Lehrkräften [Verfahrensdokumentation, Testbogen Lang- und Kurzversion, Auswertungssyntax und Datenmaske]. In Leibniz-Institut für Psychologie (ZPID) (Hrsg.), *Open Test Archive*. ZPID. https://doi.org/10.23668/psycharchives.5136

Steinmayr, R., Weidinger, A. F., Heyder, A., & Bergold, S. (2019). Warum schätzen Mädchen ihre mathematischen Kompetenzen geringer ein als Jungen? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 51*(2), 71–83. https://doi.org/10.1026/0049-8637/a000213

Stern, E. (2003). Früh übt sich: Neuere Ergebnisse aus der LOGIK-Studie zum Lösen mathematischer Textaufgaben in der Grundschule. In A. Fritz, G. Ricken, & S. Schmidt (Hrsg.), *Handbuch Rechenschwäche – Lernwege, Schwierigkeiten und Hilfen* (S. 116–130). Beltz.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Susperreguy, M. I., Di Lonardo Burr, S., Xu, C., Douglas, H., & LeFevre, J. (2020). Children's Home Numeracy Environment Predicts Growth of their Early

Mathematical Skills in Kindergarten. *Child Development, 91*(5), 1663–1680. https://doi.org/10.1111/cdev.13353

Tabeling, L., Gasteiger, H., Aumann, L., & Puca, R. M. (2022). Elterliche Einschätzung früher mathematischer Kompetenzen. *Frühe Bildung, 11*(1), 20–28. https://doi.org/10.1026/2191-9186/a000558

Tennant, L. J., & Tennant, R. F. (2010). Dyscalculia: More than mathematics phobia. *Middle East Educator, 14*, 46–49.

Thomas, K., Schulte Körne, G., & Hasselhorn, M. (2015). Entwicklungsstörungen schulischer Fertigkeiten. *Zeitschrift für Erziehungswissenschaft, 18*(3), 431–451. https://doi.org/10.25656/01:12669

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology, 19*(2), 331–351. https://doi.org/10.1037/0096-1523.19.2.331

Tröster, H. (2009). Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen. Hogrefe.

Tröster, H. (2018). *Diagnostik in schulischen Handlungsfeldern: Methoden, Konzepte, praktische Ansätze*. Kohlhammer. https://doi.org/10.17433/978-3-17-025149-6

Vigna, G., Ghidoni, E., Burgio, F., Danesin, L., Angelini, D., Benavides-Varela, S., & Semenza, C. (2022). Dyscalculia in early adulthood: Implications for numerical activities of daily living. *Brain Sciences*, *12*(3), 373. https://doi.org/10.3390/brainsci12030373

De Visscher, A., Noël, M. P., Pesenti, M., & Dormal, V. (2018). Developmental Dyscalculia in Adults: Beyond Numerical Magnitude Impairment. *Journal of Learning Disabilities, 51*(6), 600–611. https://doi.org/10.1177/0022219417732338

Visser, L., Linkersdörfer, J., Rothe, J., Görgen, R., Hasselhorn, M., & Schulte-Körne, G. (2020). Comorbidities Between Specific Learning Disorders and Psychopathology in Elementary School Children in Germany. *Frontiers in Psychiatry*, *11*, 292. https://doi.org/10.3389/fpsyt.2020.00292

Vogel, S. E., & De Smedt, B. (2021). Developmental brain dynamics of numerical and arithmetic abilities. *NPJ Science of Learning*, *6*, 22. https://doi.org/10.1038/s41539-021-00099-3

Voß, S. (2017). Datenbasierte Förderentscheidungen. In B. Hartke (Hrsg.), *Handlungsmöglichkeiten Schulische Inklusion: Das Rügener Modell kompakt* (S. 33–56). Kohlhammer.

Vossen, A., & Krizan, A. (2021). Response-to-Intervention als Rahmenmodell schulischer Lernförderung. In J. Kuhl, A. Vossen, N. Hartung, & C. Wittich (Hrsg.), *Evidenzbasierte Förderung bei Lernschwierigkeiten in der Grundschule* (S. 18–27). Ernst Reinhardt.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J. et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1)*,* 58 – 76. https://doi.org/10.3758/s13423-017-1323-7

Wagner, L. (2024). Einflussfaktoren auf die Diagnosekompetenz (angehender) Lehrkräfte – ein systematisches Literaturreview. *Unterrichtswissenschaft, 53*, 99–128. https://doi.org/10.1007/s42010-024-00215-3

Wagner, L., & Ehlert, A. (2016). Diagnostische Kompetenzen von Lehramtsstudierenden auf dem Prüfstand [Posterpräsentation]. *Tagung der Arbeitsgruppe Empirische Sonderpädagogische Forschung (AESF)*, Dortmund. https://doi.org/10.13140/RG.2.2.19950.41287

Wagner, L., & Ehlert, A. (2017). Kompetenzen angehender Lehrkräfte auf dem Prüfstand – Diagnostizieren und Interpretieren. In Institut für Mathematik der Universität Potsdam (Hrsg.), *Beiträge zum Mathematikunterricht 2017* (S. 1171–1174). WTM.

Wagner, L., & Ehlert, A. (2019). Diagnostic Competence of Math Teacher Students: An Important Skill in Inclusive Settings. In D. Kollosche, R. Marcone, M. Knigge, M. Godoy Penteado & O. Skovsmose (Hrsg.), *Inclusive Mathematics Education: State-of-the-Art Research from Brazil and Germany* (S. 561–579). Springer. https://doi.org/10.1007/978-3-030-11518-0_32

Walter, J. (2020). Ein Screening-Verfahren zur Prognose von Rechenschwierigkeiten in der Grundschule. *Zeitschrift für Heilpädagogik, 71*(5), 238–253.

Wehrmann, M. (2003). *Qualitative Diagnostik von Rechenschwierigkeiten im Grundlagenbereich Arithmetik*. Verlag Dr. Köster.

Weißhaupt, S., Peucker, S., & Wirtz, M. (2006). Diagnose mathematischen Vorwissens im Vorschulalter und Vorhersage von Rechenleistungen und Rechenschwierigkeiten in der Grundschule. *Psychologie in Erziehung und Unterricht, 53*(4), 236–245.

Weltgesundheitsorganisation Regionalbüro für Europa. (2020). *Vorsorgeuntersuchung und Screening: ein kurzer Leitfaden. Wirksamkeit erhöhen, Nutzen maximieren und Schaden minimieren.* https://iris.who.int/handle/10665/330853

White, R., & Gunstone, R. (1992). *Probing understanding*. The Falmer Press.

Wieneke, R. (o. J.). *Symptomorientierter Kriterienkatalog – Rechenschwäche.* Zentrum zur Therapie der Rechenschwäche. https://www.ztr-rechenschwaeche.de/wp-content/uploads/2021/09/Symptomkatalog-Rechenschw%C3%A4che-f%C3%BCr-Lehrer.pdf

Wilkey, E. D., Pollack, C., & Price, G. R. (2018). Dyscalculia and Typical Math Achievement Are Associated With Individual Differences in Number-Specific Executive Function. *Child Development, 91*(2), 596–619. https://doi.org/10.1111/cdev.13194

Willcutt, E. G., Boada, R., Riddle, M. W., Chhabildas, N., DeFries, J. C., & Pennington, B. F. (2011). Colorado Learning Difficulties Questionnaire: Validation of a parent-report screening measure. *Psychological Assessment, 23*(3), 778–791. https://doi.org/10.1037/a0023290

Wilson, A. J., & Dehaene, S. (2007). Number sense and developmental dyscalculia. In D. Coch, G. Dawson, & K. W. Fischer (Eds.), *Human Behavior, Learning, and the Developing Brain: Atypical Development,* (pp. 212–238). Guilford Press.

Wittich, C., & Kuhl, J. (2021). Grundlagen der evidenzbasierten Förderung bei Lernschwierigkeiten in der inklusiven Schulpraxis. In J. Kuhl, A. Vossen, N. Hartung, & C. Wittich (Hrsg.), *Evidenzbasierte Förderung bei Lernschwierigkeiten in der Grundschule* (S. 7–17). Ernst Reinhardt.

World Health Organization (WHO). (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (Vol. 1). World Health Organization.

World Health Organization (WHO). (1993). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research* (Vol. 2). World Health Organization.

World Health Organization (WHO). (2020). *International classification of diseases for mortality and morbidity statistics* (11th revision). https://icd.who.int/browse/2025-01/mms/en

World Health Organization (WHO). (2025). *ICD-11 für Mortalitäts- und Morbiditätsstatistiken (MMS) – ICD-11 in Deutsch – Entwurfsfassung*.

https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-11/uebersetzung/_node.html

Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A Meta-analysis of the Longitudinal Relationship Between Academic Self-Concept and Academic Achievement. *Educational Psychology Review, 33*(4), 1749–1778. https://doi.org/10.1007/s10648-021-09600-1

Wyschkon, A., Kohn, J., Ballaschk, K., & Esser, G. (2009). Sind Rechenstörungen genauso häufig wie Lese-Rechtschreibstörungen? *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 37*(6), 499–512. https://doi.org/10.1024/1422-4917.37.6.499

Zentrum für Rechentherapie (o. J.) *Symptomfragebogen für Eltern und Lehrkräfte.* www.rechentherapie.net/pdf/symptome.pdf

Zhao, Y. V., & Gibson, J. L. (2023). Early home learning support and home mathematics environment as predictors of children's mathematical skills between age 4 and 6: A longitudinal analysis using video observations and survey data. *Child Development, 94*(6), e377–e392. https://doi.org/10.1111/cdev.13971

Zuber, J., Pixner, S., Moeller, K., & Nuerk, H.-C. (2009). On the language specificity of basic number processing: Transcoding in a language with inversion and its relation to working memory capacity. *Journal of Experimental Child Psychology, 102*(1), 60–77. https://doi.org/10.1016/j.jecp.2008.04.003

Anhang A: Studie I

Lamb, S., Krieger, F., & Kuhn, J.-T. (2024a). Delayed development of basic numerical skills in children with developmental dyscalculia. *Frontiers in Psychology, 14*, 1187785. https://doi.org/10.3389/fpsyg.2023.1187785

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/fpsyg.2023.1187785

OPEN ACCESS

EDITED BY
Ann Dowker,
University of Oxford, United Kingdom

REVIEWED BY
Gisella Decarli,
University of Trento, Italy
Karin Landerl,
University of Graz, Austria
Stephen Barlow,
University of Oxford, United Kingdom

*CORRESPONDENCE
Sarah Lamb
✉ sarah.lamb@tu-dortmund.de

RECEIVED 16 March 2023
ACCEPTED 04 December 2023
PUBLISHED 11 January 2024

CITATION
Lamb S, Krieger F and Kuhn J-T (2024) Delayed development of basic numerical skills in children with developmental dyscalculia.
Front. Psychol. 14:1187785.
doi: 10.3389/fpsyg.2023.1187785

Delayed development of basic numerical skills in children with developmental dyscalculia

Sarah Lamb *, Florian Krieger and Jörg-Tobias Kuhn

Department of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany

Research suggests that children with developmental dyscalculia (DD) have deficits in basic numerical skills. However, there is conflicting evidence on whether basic numerical skills in children with DD are qualitatively different from those in typically developing children (TD) or whether basic numerical skills development in children with DD is simply delayed. In addition, there are also competing hypotheses about deficits in basic numerical skills, assuming (1) a general deficit in representing numerosities (Approximate Number System, ANS), (2) specific deficits in an object-based attentional system (Object Tracking System, OTS), or (3) deficits in accessing numerosities from symbols (Access Deficit, AD). Hence, the purpose of this study was to investigate whether deficits in basic numerical skills in children with DD are more indicative of a developmental delay or a dyscalculia-specific qualitative deviation and whether these deficits result from (selective) impairment of core cognitive systems involved in numerical processing. To address this, we tested 480 children (68 DD and 412 TD) in the 2nd, 3rd, and 4th grades with different paradigms for basic numerical skills (subitizing, counting, magnitude comparison tasks, number sets, and number line estimation tasks). The results revealed that DD children's impairments did not indicate qualitatively different basic numerical skills but instead pointed to a specific developmental delay, with the exception of dot enumeration. This result was corroborated when comparing mathematical profiles of DD children in 4th grade and TD children in 2nd grade, suggesting that DD children were developmentally delayed and not qualitatively different. In addition, specific deficits in core markers of numeracy in children with DD supported the ANS deficit rather than the AD and OTS deficit hypothesis.

KEYWORDS

developmental dyscalculia, basic numerical skills, domain-specific deficits, dot enumeration, magnitude comparison, number sets, number line

1 Introduction

According to DSM-5 (American Psychiatric Association, 2013) and ICD-11 (World Health Organization, 2019), developmental dyscalculia (DD) is classified as a specific learning disorder characterized by several impairments in acquiring mathematical competency compared to typically developing children (TD) with a prevalence of 3–7% depending on diagnostic criteria (e.g., Moll et al., 2014). Common definitions agree that low mathematical achievement (i.e., more than one standard deviation below average) must occur despite adequate education and normal intelligence (i.e., not more than two standard deviations below average; e.g., Szardenings et al., 2018). While a cutoff at the

25th percentile in a mathematical achievement test identifies children with learning difficulties with a broad etiological basis, a stringent cutoff (e.g., ≤ 10th percentile) more likely comprises children whose underachievement is associated with neurobiological deficits, for example, children with DD (see Mazzocco et al., 2011).

The key causes of DD continue to be debated (e.g., Decarli et al., 2023). While some authors assume that domain-general deficits play a key role (e.g., Geary, 2004), others suppose that deficits in domain-specific skills, especially in basic numerical skills, lie at the core of DD (e.g., Butterworth, 2010). In the current study, we will focus on deficits in basic numerical skills as key causes of DD.

Although current research mostly shows that deficits in basic numerical skills characterize DD (Butterworth et al., 2011), there is still conflicting evidence on whether these deficits indicate qualitative differences in basic numerical skills between DD and TD or whether DD children's skills are not qualitatively different but rather developmentally delayed. Most findings point to lower efficiency (e.g., Schwenk et al., 2017) and accuracy (e.g., Geary et al., 2008) in numerical processing, suggesting a developmentally delayed basic numerical profile in children with DD. Other findings show that the basic numerical skills of children with DD are persistently qualitatively different (e.g., Landerl, 2013), indicating that children with DD have a disproportionate impairment in basic numerical skills, resulting in abnormal numerical cognition (Landerl, 2013).

The situation is comparable concerning evidence at the neuronal level for children with DD. Studies show that children with DD display persistent structural and functional brain anomalies (Rotzer et al., 2008; Kaufmann et al., 2011; McCaskey et al., 2020). However, it remains unclear whether these abnormalities are due to developmental delay or specific markers of DD (McCaskey et al., 2020).

Based on the ability-level-match approach proposed by Bradley and Bryant (1978), the current study addresses whether DD children's impairments in basic numerical skills are more likely to indicate a developmental delay or a dyscalculia-specific qualitative deviation. This approach suggests that children's development is delayed if children with low abilities differ from their TD peers but are similar to younger children with the same ability. Alternatively, if the performance of low-achieving children differs from their TD peers and younger children with the same ability, this indicates a qualitatively different development pattern. The idea of an ability-level-match design, well known from reading-related research (e.g., Cain et al., 2000), has been applied in other fields as well. For example, Poloczek et al. (2012) investigated working memory in children with intellectual disabilities compared to matched TD children of the same mental age. In a study focusing on mathematical strategy development, Torbeyns et al. (2004) found differences in strategy use among 2nd grade children with strong and low mathematical abilities but not between 2nd and 3rd grade children matched on mathematical ability. This suggests that the mathematical strategies of children with low mathematical abilities reflect an immature level of numerical ability characterized by a delay (see Li et al., 2020).

In DD research, this approach has been used less frequently. Skagerlund and Träff (2014) examined the basic numerical skills of DD children in 4th grade, compared to an age-matched control group of TD children in 4th grade and a math ability-matched control group of TD children in 2nd grade. Based on reaction times (RTs), no significant differences were found between the DD and TD groups. Although descriptive, DD children showed lower RTs than TD children in 2nd grade but longer RTs than TD children in 4th grade, suggesting that the abnormalities are due to developmental delay. In the most recent studies, qualitative deviations were inferred based on statistical interactions or varying slopes. However, it is important to take age into account to ensure that qualitatively different deviations are not moderated by age-related differences. There are substantial changes in children's mathematical development in the first years of schooling (e.g., Landerl, 2013). Especially in primary school, differences in mathematical skills between TD and DD children can, therefore, be influenced by children's numerical age development. Thus, the question of whether DD is related to a qualitative difference or a delay can ultimately only be answered if we take children's numerical age development (grade level) into account. Statistical evidence of dyscalculia-specific *qualitative* impairments refers to an (over-additive) statistical interaction effect between group (TD/DD) and basic numerical skills, as well as the absence of interactions moderated by grade level. In the case of *delayed* (additive) impairment, the performance of children with DD is generally slower or less accurate than that of children without DD, but there are no qualitative differences between the groups. This is reflected in the lack of statistical interaction effects.

Strongly intertwined with the characterization of basic numerical deficits (qualitative difference in DD or TD or developmental delay of DD children) is the question about the causes of basic numerical deficits. In fact, there are conflicting hypotheses and heterogeneous results pertaining to the causes of basic numerical deficits. These hypotheses consider deficits in core systems for numerical processing and derive predictions on how deficits in one of these systems affect basic numerical skills in general.

Due to the heterogeneous evidence on characterization and causes of impaired basic numerical skills, the goals of this study were to comprehensively investigate (1) whether DD children's impairments in basic numerical skills are more likely to indicate a developmental delay or a dyscalculia-specific qualitative deviation and (2) whether these deficits derive from a (selective) impairment in cognitive core systems of numerical processing. Answering these research questions is essential to design and improve interventions for children with DD. We selected groups of children who matched in chronological school age (2nd, 3rd, and 4th grades) but who differed in mathematical ability to compare their performance in basic numerical skills in a cross-sectional design. Inspired by the ability-level-match design approach and similar to Skagerlund and Träff (2014), we compared the basic numerical profiles of 4th grade DD children with the performance of 2nd grade TD children to examine whether DD is a developmental delay or results in a qualitatively different numerical processing profile. In the following, we will describe core systems of numerical processing.

1.1 Core systems of numerical processing

Feigenson et al. (2004) distinguish two independent core systems of numerical processing. One system is used for *approximate* recognition and estimation of numerosities (approximate number system, ANS). In contrast, the other core system (object tracking system, OTS) is utilized for *accurate* discrimination and recognition of three to four objects without counting them (*subitizing*). These innate core systems are the foundation for developing the symbolic

number system (Feigenson et al., 2004). According to the Triple Code Model (TCM) by Dehaene (1992), building on the ANS that processes large (> 4) representations of non-symbolic numerosities (e.g., •••), children develop two additional modules of numerical processing during preschool and school (von Aster and Shalev, 2007): the verbal-phonological module is responsible for processing written and spoken number words (e.g., /three/), whereas the visual-Arabic module processes written Arabic numerals (e.g., 3). Several theories propose that the cause of DD could be a deficit in one of the core systems of numerical processing (see Figure 1).

According to the *OTS deficit hypothesis*, some authors see the cause for DD in a defective OTS (e.g., Butterworth, 2010; see Figure 1A), and according to the *ANS deficit hypothesis,* some authors see the cause in a defective ANS but non-impaired OTS (e.g., Piazza et al., 2010; see Figure 1B). In addition, the *access deficit hypothesis* (AD) (e.g., Rousselle and Noël, 2007; see Figures 1C,D) claims that the ANS is not impaired *per se*, but *accessing* the visual Arabic module is impaired, which indicates an access deficit. Thus, DD children should be (disproportionately) impaired in symbolic numerical processing compared to non-symbolic numerical processing (see Noël and Rousselle, 2011; Schwenk et al., 2017).

Because the development of numerical processing is based on the core systems (e.g., Wilson and Dehaene, 2007), all three hypotheses predict problems in processing representations from symbols. However, the findings pertaining to non-symbolic numerical processing are less clear. While some results support the ANS hypothesis (e.g., Decarli et al., 2020), others are more in line with the AD hypothesis (e.g., Schwenk et al., 2017). In addition to a lack of consistent criteria and methods to critically assess empirical results with respect to these hypotheses (see Olsson et al., 2016), age could explain the conflicting results. According to Rousselle and Noël (2007), deficits in processing non-symbolic numerosities could also

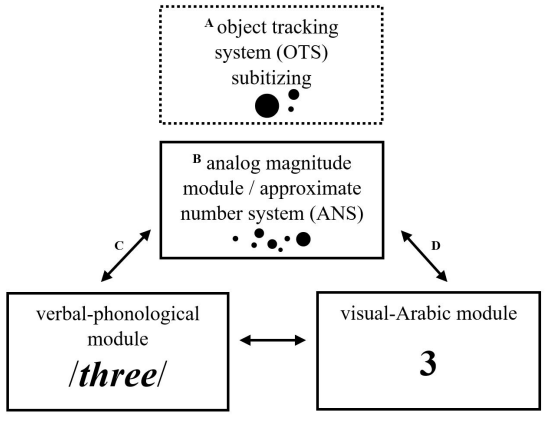


FIGURE 1
Core systems of numerical processing. Object tracking system (OTS) and approximate number system (ANS) are seen as innate, independent core systems of numerical processing (Feigenson et al., 2004). ANS/analog magnitude, verbal-phonological, and visual-Arabic modules are regarded as neurocognitive numerical processing components of the Triple Code Model by Dehaene (1992). The letters pertain to the hypotheses about the causes of developmental dyscalculia. **(A)** OTS deficit = specific deficit in subitizing; **(B)** ANS deficit = general deficit in representing numerosities; **(C,D)** Access deficit (AD) = deficits in accessing numerosities from symbols (based on Landerl et al., 2021).

be interpreted as a consequence of (poor) mathematical development. A review by de Smedt et al. (2013) revealed that 9-year-old or older children with DD showed more inefficient non-symbolic numerical processing than TD children; younger DD children, however, did not. This result points to different assumptions about the causes of DD depending on a child's age. Nevertheless, especially in primary school, age could be an influencing factor, as early math learning represents a sensitive developmental phase for mathematical competencies (see Raddatz et al., 2017). Children are increasingly detached from visual materials and the number space expands, and they are taught the basics of arithmetic (see Landerl, 2013). This is reflected in developmental leaps in basic numerical skills in the first years of schooling (e.g., Moore and Ashcraft, 2015) and in increasing efficiency and accuracy in performing basic numerical tasks from grade to grade (see Landerl and Kölle, 2009; Landerl, 2013; Moore and Ashcraft, 2015). In the 2nd grade, children still stand at the very beginning of their mathematical education. In the 4th grade, in contrast, they have already reached an adult-like level in some basic numerical tasks (Moore and Ashcraft, 2015). Thus, there are particularly large developmental leaps from 2nd grade to 4th grade.

Importantly, previous studies (e.g., Andersson and Östergren, 2012; Decarli et al., 2023) that contrasted hypotheses referring to the causes of DD often did not consider age-related differences in primary school children. In turn, many studies focusing on DD children's numerical age development failed to systematically compare the hypotheses about the causes of DD (e.g., Landerl, 2013) or investigated only a small spectrum of basic numerical skills (e.g., Landerl and Kölle, 2009; Schleifer and Landerl, 2011). However, these studies did not explicitly investigate whether DD children's basic numerical skills are qualitatively different or, instead, developmentally delayed. Other studies either only considered DD children in 4th grade (Skagerlund and Träff, 2014) or did not compare basic numerical profiles of 4th grade DD children and 2nd grade TD children (e.g., Landerl, 2013).

To summarize, the question of whether numerical processing deficits in children with DD stem from an ANS and/or OTS deficit or whether they can be explained by the AD hypothesis has not yet been conclusively answered. Furthermore, we cannot conclusively say whether DD children's deficits are more indicative of a developmental delay or a qualitatively different numerical deficit profile. Addressing this gap, we presented different basic numerical tasks to children with and without DD, allowing a detailed view of numerical processing. The following subsection introduces the state of research on basic numerical deficits in DD along the different numerical processing task paradigms (see Landerl, 2019) investigated in the current study.

1.2 Subitizing vs. counting

To investigate whether DD children are impaired in numerical judgments of small countable objects, enumeration tasks are typically used. These tasks test the ability to count a limited number of objects (e.g., dots) as rapidly as possible (e.g., Landerl, 2019). About three to four dots can be processed in a preattentive way by tapping the OTS (subitizing), while more than three to four dots require serial counting (Trick and Pylyshyn, 1993) by tapping the verbal-phonological module. The latter is related to the ANS (von Aster and Shalev, 2007) in that lower efficiency in counting also suggests a defective ANS.

Several studies report subitizing and/or counting problems for DD children (Schleifer and Landerl, 2011; Andersson and Östergren, 2012; Landerl, 2013; Decarli et al., 2023). Some results (e.g., Landerl et al., 2004; Raddatz et al., 2017) suggest that DD children have a qualitatively different approach to counting, while other results (e.g., Schleifer and Landerl, 2011; Landerl, 2013) indicate a dyscalculia-specific impairment in subitizing, suggesting that DD children need to count even three or fewer dots serially (Kuhn et al., 2013). Generally, both subitizing and counting skills steadily increase across children's development (Reeve et al., 2012; Moore and Ashcraft, 2015). However, while TD children become systematically more efficient at subitizing, there is evidence that this is not the case for DD children (Landerl, 2013). Previous study results showed that children with DD display more difficulties and larger slopes in the subitizing range than TD children (e.g., Schleifer and Landerl, 2011; Landerl, 2013), suggesting that the RTs of the DD children were not only slower but disproportionately slower, indicating qualitatively different numerical processing. Consistent throughout elementary school, this difference is an indication of a different development in DD and a dyscalculia-specific subitizing problem (Landerl, 2013). Although DD children were also substantially more inefficient in the counting range than TD children, the developmental trajectory did not appear to be qualitatively different (Schleifer and Landerl, 2011; Landerl, 2013).

1.3 Non-symbolic, symbolic, and mixed magnitude comparison

The most common method of generating information about the cognitive representation of numbers and the precision or efficiency of the ANS is to use magnitude comparison tasks (e.g., Halberda et al., 2008). This task type involves selecting the numerically larger of two (non-)symbolic numerosities (e.g., dots or numbers) as quickly as possible without counting (e.g., Moyer and Landauer, 1967; Halberda et al., 2008; Landerl, 2019). The estimation of numerosities is based on the ANS (Feigenson et al., 2004). The mental representation of numerosities in children with DD is thought to be less precise than in children with TD (e.g., Piazza et al., 2010), resulting in less correct differentiation of close numbers (also known as *distance effect*) by DD than by TD children. As a result, children with DD need longer RTs and/or show lower accuracy. In line with the ANS deficit, DD children should display impaired performance in non-symbolic magnitude comparison tasks (see Wilson and Dehaene, 2007). In contrast to the comparison of non-symbolic quantities, before comparing symbolic quantities (i.e., numbers), the visual-Arabic input must be converted to the analog quantity (Henik and Tzelgov, 1982). Thus, there is an interaction between the ANS and the visual-Arabic module, which, according to the AD hypothesis, is disturbed in DD. Some studies support the ANS deficit hypothesis (e.g., Feigenson et al., 2004; Piazza et al., 2010; Decarli et al., 2020), while other studies show that children with DD are only impaired in symbolic comparisons or at least have difficulty in non-symbolic comparisons to a much lesser extent (e.g., Olsson et al., 2016; Schwenk et al., 2017), thus supporting the AD hypothesis.

Age seems to be an important moderator of the results. Children with and without DD generally become more efficient with increasing age (Landerl, 2013). However, the development of symbolic numerical processing in DD children proceeds less systematically than in TD children (Landerl, 2013). In line with the ANS deficit hypothesis, Skagerlund and Träff (2014) revealed that the DD children showed significantly noisier ANS representations than TD children in the 4th and 2nd grades (Skagerlund and Träff, 2014). Piazza et al. (2010) observed that 10-year-old DD children's number acuity was comparable to that of TD children 5 years younger, suggesting a delayed mathematical development. Several studies reported differences in non-symbolic comparisons only between older TD and DD children, but not in younger children with and without DD (for a review, see de Smedt et al., 2013), supporting the AD hypothesis by Rousselle and Noël (2007), which interprets that non-symbolic deficits occur as a consequence of (poor) mathematical development. In summary, age seems to influence DD and TD children's performance in non-symbolic processing, but the evidence remains unclear (e.g., de Smedt et al., 2013; Skagerlund and Träff, 2014).

1.3.1 Comparison distance effect

In addition, in the context of the magnitude comparison task, it has reliably been reported that the larger the distance between the two quantities (e.g., dots or numbers) being compared, the better the observed performance (e.g., Moyer and Landauer, 1967; Henik and Tzelgov, 1982). This finding is referred to as the comparison distance effect (Moyer and Landauer, 1967; Dehaene, 1992; Holloway and Ansari, 2008). In line with the most common interpretation, the distance effect is based on the assumption that cognitive magnitude representations are ordered along a mental number line/or mapped on the ANS (see also mental number line theory; Dehaene, 2011). It follows that the distance effect is also an indicator of the ANS acuity (e.g., Peters et al., 2008). Due to the ANS deficit, DD children should have a mental number line that is less mature, making it more challenging to represent and distinguish (non)-symbolic magnitudes (e.g., Mussolin et al., 2010; Piazza et al., 2010; Andersson and Östergren, 2012). Consequently, DD children should also display a greater distance effect than TD children. However, research results concerning the distance effect are not consistent. In line with the ANS deficit, a larger distance effect has been observed in children with DD in some studies (e.g., Mussolin et al., 2010). In contrast, a recent meta-analysis (Schwenk et al., 2017) reported that no qualitatively different distance effect in symbolic or non-symbolic comparison paradigms could be found in children with DD. Most results suggested a relatively more inefficient than a qualitatively different numerical processing pattern in DD (Landerl and Kölle, 2009; Kuhn et al., 2013; Landerl, 2013). Generally, the distance effect was observed for children with and without DD in symbolic and non-symbolic magnitude comparisons from 2nd grade onward (Landerl and Kölle, 2009; Landerl, 2013). Some studies reported that the distance effect was stable across development (Reeve et al., 2012; Landerl, 2013), while others reported a steadily decreasing distance effect with age (Holloway and Ansari, 2008; Moore and Ashcraft, 2015). DD children were generally more inefficient than TD children across the elementary school (Holloway and Ansari, 2009). The results did not suggest a qualitatively different processing pattern in DD; instead, they suggested a developmental delay. A systematic analysis investigating small and large distances for symbolic and non-symbolic magnitude comparisons between children with and without DD in the 2nd, 3rd, and 4th grades is still lacking.

For the less frequently used mixed comparison task, tapping several core systems of numerical processing (OTS, ANS, and

visual-Arabic modules, respectively; Raddatz et al., 2017), there is currently a lack of knowledge. The task addresses children's mapping skills in addition to their numerical processing skills, as in solving the mixed comparison task, two quantities in different modalities (e.g., point sets and Arabic numerals) are compared (Kuhn et al., 2013). To solve the mixed comparison task efficiently, children need to know how the non-symbolic and written Arabic numerals are related to each other (Kuhn et al., 2013). Previous results have shown that DD children were less efficient compared to TD children but did not seem to be disproportionately impaired (Kuhn et al., 2013). Research examining the influence of the age of children with and without DD in the context of this task is still lacking.

1.4 Number sets

The number sets task developed by Geary et al. (2009) primarily taps the visual-Arabic module (Raddatz et al., 2017). In this task, children compare an Arabic number (target) with a number set displayed below (Kuhn et al., 2013). The number set consists of two Arabic numbers, two different numerosities of dots, or combining an Arabic number and one numerosity of dots (Raddatz et al., 2017). Children decide whether the target matches the number of the total number set (e.g., target = 9, set 4 (Arabic number) + five dots) (Raddatz et al., 2017). Non-symbolic numerosities in this task can be in the subitizing or counting range, requiring ANS processing. In addition, numerosities from number symbols tap the visual-Arabic module. Although this task does not differentiate well between single hypotheses pertaining to the causes of DD, it significantly distinguishes between children with and without DD (see Geary et al., 2009). Previous studies showed that children with DD show impaired performance in this task compared to TD children (e.g., Kuhn et al., 2013; Raddatz et al., 2017; von Wirth et al., 2021). Brankaer et al. (2014) investigated similar mapping tasks and observed that mapping abilities continue to develop through primary school. Although some authors examined the performance of elementary school students with DD in this task (e.g., Kuhn et al., 2013), it is still unclear whether performance in children with DD develops qualitatively different from children without DD.

1.5 Number line estimation

The number line estimation task entails transcoding a numerical value into a spatial position on a visual line bounded by two numbers (e.g., 0 and 100) (Siegler et al., 2009). Number line tasks require an understanding of ordinality and estimation skills (see von Aster and Shalev, 2007), thus tapping the ANS (Feigenson et al., 2004). A less mature mental number line results in difficulties representing and distinguishing between numerosities. For example, younger children overrepresent small numbers on the number line (e.g., they locate 300 at about 450 on a number line from 0 to 1,000; Booth and Siegler, 2006). This logarithmic rather than linear conception of the mental number line leads to less accurate estimates (Booth and Siegler, 2006). In line with the defective ANS and AD hypotheses, children with DD should display problems locating numbers on the number line (see Wilson and Dehaene, 2007). Several empirical results support this assumption (Geary et al., 2008; Decarli et al., 2023). DD children's estimations deviated more strongly from the target number and corresponded more closely to a logarithmic than a linear pattern. Generally, mental number line precision improved with age (Booth and Siegler, 2006; Landerl, 2013). However, Landerl's (2013) findings indicated that DD children's estimates were less accurate than the TD children's estimates throughout elementary school. Nevertheless, only at the first measurement point were numbers represented in a logarithmically compressed way. The low accuracy of DD children's performance supports the assumption of a persistent general inaccuracy, which, however, does not differ qualitatively from TD children. Whether DD children's mental number line precision develops with a delay or whether it remains unspecified has remained unclear.

1.6 Current study

Most findings indicate delayed rather than qualitatively different numerical processing in dyscalculia. In line with extant research, we expected that DD children display deficits in basic numerical skills, suggesting a lower efficiency and accuracy of numerical processing rather than a qualitative difference (*disproportionate* impairment). Generally, an impairment is present when DD children perform significantly worse than TD children. In detail, we addressed the following two overarching research questions (RQs): (1) Are DD children qualitatively different in basic numerical skills compared to TD children, or is there, instead, a developmental delay for DD children? (2) Are deficits in the basic numerical skills most compatible with the ANS-, OTS-, and/or AD-deficit hypothesis?

We examined various basic numerical tasks in children with and without DD in a cross-sectional design. Based on the numerical development in the first school years, it would be conceivable that differences between children with and without DD are due to numerical age development. To control for this effect of age development, we contrasted children's performance for three grades (2nd–4th grades). Concerning age-related mathematical development, we expected that children become increasingly efficient. The increasing efficiency in numerical processing should, however, vary depending on the task.

With regard to RQ1, we investigated whether there is evidence for qualitatively different numerical processing in children with DD that is (not) moderated by grade level, indicating qualitatively different numerical processing (*over-additive impairment*) in children with DD. If the deficits are moderated by grade level, this would suggest that the development of the DD children is delayed in view of the ability-level-match design.

Furthermore, inspired by the ability-level-match design, we compared the basic numerical profiles of 4th grade DD children with the performance of 2nd grade TD children to examine whether DD is a developmental delay or whether DD results in a qualitatively different numerical processing profile. The interval of two school years (similar to Skagerlund and Träff, 2014) was chosen as previous work showed that the developmental delay varies between 1 and 5 years depending on the task (e.g., Piazza et al., 2010). Assuming that children with DD are developmentally delayed, the following patterns of results were expected: (a) There is a developmental delay of more than 2 school years in children with DD. In this case, TD children in 2nd grade will be substantially more efficient in solving basic numerical tasks than the DD children in 4th

grade. (b) The developmental delay of the DD children is less than 2 school years; the DD children in 4th grade will be significantly better than the TD children in 2nd grade. (c) DD children's numerical profile indicates qualitatively different numerical processing, suggesting a disproportionately large impairment, indicated by an interaction of the factors group and tasks.

With regard to RQ2, we compared different basic numerical paradigms for children with and without DD. Given that AD holds, we expected that accessing numerosities from symbols would be disproportionately impaired compared to non-symbolic numerosities. For example, the pattern of the following results would be consistent with the AD hypothesis: in a magnitude comparison task, children with DD are substantially slower than TD children in comparing symbolic magnitudes (i.e., numbers), whereas the difference between DD and TD children is much smaller when comparing non-symbolic numbers (Rousselle and Noël, 2007). Based on the literature, we anticipated that children with DD would show longer RTs for magnitude comparisons but no qualitatively different distance effect. Thus, children with DD are expected to be more inefficient and less accurate in magnitude comparisons (indicated by an effect of the group), but their mental number line representation should not seem qualitatively abnormal. Thus, we expected no disproportionate impairment, as indicated by the absence of an interaction between group and task conditions (small vs. large magnitudes). Pertaining to the three hypotheses about the causes of DD (OTS-, ANS-, and AD-deficit hypothesis), we expected patterns of impairment in the following tasks (see Table 1).

2 Materials and methods

2.1 Participants and procedure

A total of $N = 480$ children (2nd–4th grade) composed of 68 DD (mathematical abilities, see instruments: T-score ≤ 38; IQ > 70) and 412 TD (mathematical abilities: T-score > 38; IQ > 70) were recruited from 46 classes in 8 primary schools in Germany. The local ethics committee approved the study protocol, and parental consent was obtained before testing. DD and TD samples included 47 and 78 children, respectively, with comorbid reading disorders (reading fluency, PR ≤ 16). We did not exclude these children because it has been shown that children with reading disorders were only selectively impaired in verbal number tasks (e.g., counting; Raddatz et al., 2017).

Gender was evenly distributed between the grades, χ^2 (2) $= 0.916$, $p = 0.633$ and groups, χ^2 (1) $= 3.392$, $p = 0.655$. The detailed demographics of all subjects are summarized in Table 2.

2.2 Instruments

Tasks measuring reading fluency, non-verbal intelligence, and mathematical abilities were administered in class, while tasks assessing basic numerical skills were computer-administered in smaller groups of 13 pupils.

2.2.1 Arithmetic

Children's mathematical abilities were examined by the Arithmetic Operations of the *Heidelberger Rechentest* (HRT 1–4; Engl.: Heidelberger Numeracy Test 1–4; Haffner et al., 2005). The paper-based speed test consists of two scales, which comprise a total of 11 subtests. The first scale Arithmetic Operations (test–retest reliability $r_{tt} = 0.93$; Haffner et al., 2005) includes the following subtests: addition (e.g., 3 + 5), subtraction (e.g., 5–3), multiplication (e.g., 2 · 3), division (e.g., 9 ÷ 3), complement task (e.g., __ + 3 = 10), and greater (>)/less (<)/equal (=) number comparisons (e.g., 3 __ 71); and the second scale Numerical-logical and Visual-spatial skills ($r_{tt} = 0.87$; Haffner et al., 2005) comprises: numerical sequences (e.g., 12 11 10 9 8 7_ _ _), length estimation, dice counting), counting objects/figures, and connecting numbers. Each subtest consists of at least 10 (e.g., connecting numbers) and at most 40 (e.g., addition) tasks arranged in order of increasing difficulty. The children had to solve as many tasks as possible within 2 min for each subtest. To calculate a total test score, both subscales were combined.

2.2.2 Intelligence

The non-verbal intelligence of 2nd and 3rd graders was examined by three subtests of the *Grundintelligenztest Skala 1-Revision* (CFT 1-R; Engl.: Culture Fair Intelligence Test 1-R Scale 1; Weiß and Osterland, 2013): series completion, classification, and matrices ($r_{tt} = 0.95$). The 4th graders were assessed using four subtests of the *Grundintelligenztest Skala 2-Revision* (CFT 20-R; Engl.: Culture Fair Intelligence Test 20-R Scale 2; Weiß, 2006): series completion, classification, matrices, and topologies ($r_{tt} = 0.80$). The test score (IQ) was calculated from the subscales used in each case.

2.2.3 Reading fluency

Reading fluency was measured using the paper-based *Salzburger Lese-Screening für die Klassenstufen 1–4* (SLS 1–4; Engl.: Salzburg reading screening test for grades 1–4; parallel-forms reliability = 0.90; Mayringer and Wimmer, 2003). The children were presented with a list of 70 simple sentences containing correct and incorrect statements (e.g., "bananas can talk"). Within 3 min, as many sentences as possible had to be read and judged with regard to their correctness. The test score depends on the total number of sentences judged correctly.

All tasks assessing basic numerical skills given below consisted of the *CODY-M 2–4 battery* (Kuhn et al., 2017), except the Panamath task.

2.2.4 Dot enumeration (DE)

Children were presented with several black dots (1–9) and asked to count the dots as quickly as possible. We chose a data-driven approach to divide the number of points to be counted into the subitizing and counting range; based on the model fit of the piecewise regressions, we set the subitizing range for each child individually between 1 and 3 or 1 and 4. More than half of the children ($n = 266$) could subitize three items. Depending on the number of dots, the subitizing (1–3/1–4) or counting (4–9/5–9) skills are captured. Each magnitude 1 to 9 was presented twice (a total of 18 items; time limit: 2 min). Based on the total area of the points, no clear conclusions can be drawn about the quantity to be counted. The median of the children's RTs, averaged over all correct answers and separately for the subitizing and the counting range, was used as the test score (Raddatz et al., 2017).

TABLE 1 Expected pattern of impairments in line with the ANS deficit hypothesis, OTS deficit hypothesis, and AD hypothesis.

Tasks	ANS	OTS	AD
Dot enumeration subitizing range (1–3/1–4)	Not impaired	Impaired	Not impaired
Dot enumeration counting range (4–9/5–9)[2]	Impaired	Not impaired	Not impaired
Number comparison[1,2]	Impaired	Not impaired	Impaired
Mixed comparison[1,2]	Impaired	(Not impaired)*	Impaired
Dot magnitude comparison/Panamath[1]	Impaired	Not impaired	Not impaired
Number sets[3]	Impaired	(Impaired)*	Impaired
Number line estimation[1]	Impaired	Not impaired	Impaired

The study's tasks substantially tap the analog magnitude,[1] verbal-phonological,[2] and visual-Arabic[3] module of the Triple Code Model by Dehaene (1992). *Tasks cannot differentiate sharply between hypotheses because tasks substantially tap the ANS and visual-Arabic module, but points (tapping the ANS) can be in the subitizing (tapping OTS) or counting range (tapping verbal-phonological module). ANS deficit = general deficit in representing numerosities; OTS deficit = specific deficits in an object-based attentional system; AD = deficits in accessing numerosities from symbols (based on Andersson and Östergren, 2012).

TABLE 2 Demographic characteristics, scores on mathematical abilities, intelligence, and reading fluency by grade and mathematical ability.

	2nd grade		3rd grade		4th grade	
Details	TD	DD	TD	DD	TD	DD
n (boys)	177 (85)	17 (8)	180 (84)	34 (10)	55 (29)	17 (6)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Age, in months	98.56_{a} (6.88)	98.62_{a} (10.93)	110.27_{b} (6.77)	111.12_{b} (6.45)	120.90_{c} (6.23)	120.93_{c} (9.03)
HRT 1–4	104.88_{a} (11.40)	77.38_{b} (3.66)	103.32_{a} (12.62)	75.44_{b} (5.43)	102.65_{a} (13.69)	77.13_{b} (5.07)
CFT[1]	103.20_{a} (13.04)	$95.68_{a,c}$ (12.31)	104.29_{a} (12.03)	94.50_{c} (13.23)	$96.44_{b,c}$ (13.59)	88.03_{c} (9.46)
SLS 1–4	97.51_{a} (15.94)	79.35_{b} (13.53)	99.72_{a} (14.56)	78.00_{b} (16.58)	101.00_{a} (18.71)	82.47_{b} (14.61)

1, NA for gender and 39 NA for age; HRT 1–4, Heidelberger Rechentest (mathematical abilities); CFT, Culture Fair Intelligence Test (intelligence, intelligence quotient); SLS 1–4, Salzburger Lese-Screening für die Klassenstufen 1–4 (reading fluency, reading quotient). Scaling for all measures (*M*: 100/*SD*: 15). *Post-hoc* comparisons between groups were based on Holm's method ($p < 0.05$). Mean scores with the same indices (per row) do not differ significantly. [1] Culture Fair Intelligence Test 1–Revision (CFT 1-R) was used for 2nd and 3rd graders, and Culture Fair Intelligence Test 2–Revision (CFT 20-R) was used for 4th graders.

2.2.5 Number comparison (NC)

Pairs of single-digit Arabic numbers were presented on a screen. The task was to compare the Arabic numbers and select the larger Arabic number as quickly as possible (27 items; lime limit: 1.5 min). Numerical distances between both stimuli (small: 1–3; large: 4–6) varied systematically between one and six. Each difference appeared four times in a random order, but the same for all participants (Raddatz et al., 2017). The median of the children's RTs for correct answers for small (1–3) and large (4–6) distances were calculated (e.g., Holloway and Ansari, 2009; Kuhn et al., 2013).

2.2.6 Mixed comparison (MC)

This task works similarly to number comparison, differing only in one quantity being represented as a cloud of one to nine dots, instead of a number. When creating the dot stimuli, care was taken to ensure that the overall area of the dots does not allow unambiguous conclusions to be drawn about the quantity to be counted (Kuhn et al., 2017). The test score was calculated in the same way as in the number comparison task. Together, these three tasks represent the basic numerical processing scale, which has good test–retest reliability, $r_{tt} = 0.72$ (Kuhn et al., 2017).

2.2.7 Dot magnitude comparison (Panamath)

Children were presented with 48 items consisting of yellow and blue dots (ranging from 5 to 21) on the screen and had to decide as quickly as possible which point cloud contained more dots without counting them (e.g., Halberda et al., 2008). In line with Raddatz et al. (2017), the items were presented with four ratios between the two sets: 1.2, 1.4, 1.6, and 2.6 (12 items each). The total score of correct answers and the median of the children's RTs for correct answers were calculated.

2.2.8 Number sets test (NS)

The speed and accuracy in identifying and processing quantities in different representations were measured using a task based on Geary et al. (2009). Participants had to compare a number set (consisting of numerosities of dots and/or an Arabic number) shown at the bottom of the screen with a target (Arabic number) at the top of the screen, then they had to decide as quickly as possible whether the sum of the number set shown at the bottom was equal to the target shown above (Raddatz et al., 2017). The time limit for each task type (target number 5 or 9) was 1.5 min. To calculate the test score, the incorrect answers (false alarms) were subtracted from the correct answers (hits) (Kuhn et al., 2017).

2.2.9 Number line estimation (NL)

The acuity/precision of the mental number line was measured using a task based on Siegler and Booth (2004). Participants were asked to locate a presented number (1–99) on a number line with endpoints 0 and 100. For the analyses, the mean deviation between the target number and the answer was calculated (Kuhn et al., 2017). The time limit per item (23 in total) was 3.5 min (Raddatz et al., 2017). The number sets test and number line estimation are part of the complex number processing scale, which has good test–retest reliability, $r_{tt} = 0.76$ (Kuhn et al., 2017).

2.3 Statistical analyses

All statistical calculations were performed using the statistical software R (version 4.3.1; R Core Team, 2023).

Our observations are cross-sectional clustered data: classes within schools (level 3), students within classes (level 2), and children's performance in different tasks within students (level 1). We used a multilevel approach to account for the clustering.

First, intercept-only models with schools as fixed effects were specified separately for each task. Intraclass correlation coefficients (ICCs) were calculated. When a school was associated with children's performance on basic numerical tasks, school was retained as a fixed effect in the model specification of linear mixed models (LMMs).

Second, random-intercept-constant-slope (rics) models were specified and tested separately for each task. In the rics models, we included the predictor's group (typical/dyscalculic), grade (2/3/4), and task condition (e.g., dot enumeration: subitizing/counting) and their interactions as fixed effects, while we considered classes (grade: 2/3/4) and students as random intercepts. The covariates IQ and reading fluency were regarded as fixed factors. Only a two-level structure (classes within schools (level 2) and students within classes (level 1)) was used for number sets and the number line task, as no task conditions were clustered within children.

To account for heteroscedasticity, we used a bias-reduced linearization (BRL) generalization, which corrects cluster-robust variance estimation (CRVE) in conjunction with a Satterthwaite approximation for t-tests. The methods are implemented in an R package called clubSandwich (Pustejovsky and Tipton, 2018). Non-normal distributions are less critical for LMMs (Schielzeth et al., 2020).

Third, pairwise contrasts based on fitted models with Satterthwaite approximation for degrees of freedom using Holm's method were calculated for all significant effects to identify the effects more precisely based on mean differences. When the results of the pairwise contrasts differed from the rics model, the results of the rics model were used because they are more robust.

An interaction of group × task (e.g., number comparison: small and large distances) in the rics model would provide evidence for qualitatively different numerical processing (over-additive impairment) in children with DD. A significant effect of group and the absence of the above interaction for a given dependent variable (basic numerical task) would indicate a less efficient but not qualitatively different numerical processing (additive impairment) in children with DD. An interaction of group × grade would suggest a difference in numerical development between children with and without DD. If qualitative differences between children with and without DD are moderated by grade level, this would be shown in an interaction of group × task × grade, indicating that the qualitative differences between groups are not constant across grades. Whether interactions with the grade level indicate a qualitatively different developmental trajectory or rather a developmental delay needs to be tested with pairwise contrasts.

If the basic numerical skills of children with DD develop with a delay, children with DD would have to reach a comparable level of numerical processing as the TD children later. Therefore, we examined whether DD children in 4th grade catch up in basic numerical skills with TD children in 2nd grade or whether there is evidence of persistent inefficiency or qualitative differences in numerical processing. A significant effect of group (TD2/DD4) indicates a developmental delay across basic numerical skills, whereas an interaction of group × task indicates a qualitative difference in numerical processing.

3 Results

3.1 Dot enumeration

The predictors group (typical/dyscalculic), grade (2/3/4), and dot enumeration (subitizing/counting) and their interactions, except dot enumeration × group × grade, were significant (see Table 3). Pairwise contrasts based on the rics model showed that the DD group required significantly longer RTs than the TD group, $p < 0.0001$. RTs decreased systematically across grades (g2 > g3 > g4, $ps < 0.05$). Point sets in the counting range (4–9/5–9) resulted in longer RTs than point sets in the subitizing range (1–3/1–4), $p < 0.0001$. The large mean difference between the DD and TD children for the counting range (see Figure 2) resulted in a significant interaction of dot enumeration × group. The DD group required significantly longer RTs than the TD group in the counting range ($p < 0.0001$); however, not in the subitizing range ($p = 0.235$). The interaction of dot enumeration × grade may have been related to a relatively small decrease in RTs from 3rd to 4th grade for both task conditions, especially for the subitizing range ($ps < 0.01$, except 3rd vs. 4th grade for subitizing). The group × grade interaction

TABLE 3 Results of the linear mixed model using task condition, group, grade, reading fluency, and intelligence to predict children's performance on dot enumeration.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	2829.701	185.953	15.217	28.9	< 0.001
DE[a]	2357.316	61.734	38.185	14.0	< 0.001
Group[b]	316.240	133.730	2.365	8.9	0.043
Grade[c] 3	−227.064	54.752	−4.147	32.7	< 0.001
Grade 4	−301.789	63.290	−4.768	8.7	0.001
Reading fluency	−8.004	1.285	−6.228	26.3	< 0.001
IQ	−2.549	2.074	−1.229	29.6	0.229
DE × group	682.125	185.825	3.671	8.5	0.006
DE × grade 3	−281.861	75.630	−3.727	26.6	0.001
DE × grade 4	−468.362	121.968	−3.840	6.9	0.007
Group × grade 3	−284.101	154.505	−1.839	16.4	0.084
Group × grade 4	−370.736	138.137	−2.684	9.9	0.023
DE × group × grade 3	−74.698	229.762	−0.325	16.2	0.749
DE × group × grade 4	−162.668	293.630	−0.554	10.1	0.592
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	290.73				
$\tau_{00\ IDclass}$	154.83				
δ^2	453.10				
ICC $_{IDstudent:\ IDclass}$	0.297				
ICC $_{IDclass}$	0.141				

The model summary based on *p*-values for fixed effects was calculated using Satterthwaite approximations.
DE = Dot enumeration.
[a]Subitizing is the reference.
[b]Typical developed is the reference.
[c]Grade 2 is the reference.

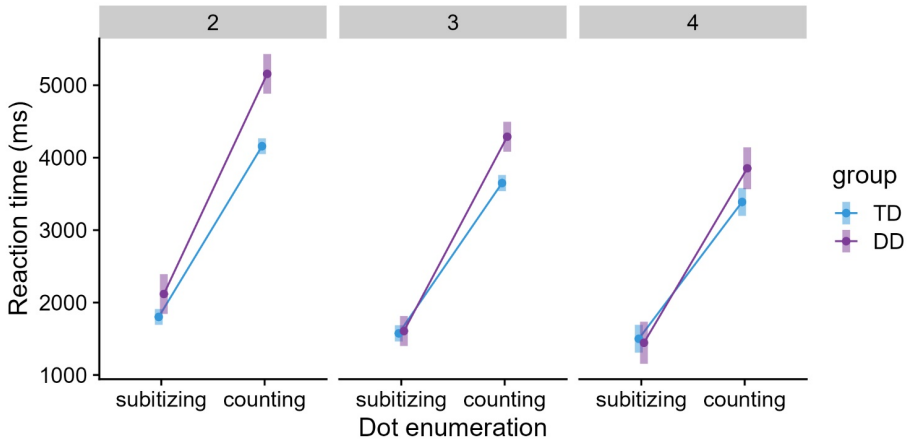


FIGURE 2
Means and 95% confidence intervals (CIs) for dot enumeration grouped by group and grade.

showed that children with and without DD differed significantly in 2nd and 3rd grade ($p < 0.01$) but not in 4th grade ($p = 0.393$). Grade level trends within groups indicated that DD and TD children reached a stable level in 3rd grade (TD2 > TD3 = TD4; DD2 > DD3 = DD4, $ps < 0.0001$). In addition, higher levels of reading fluency were associated with lower RTs on the dot enumeration task, but the effect of IQ was not significant.

3.2 Number comparison

The predictors group (typical/dyscalculic), grade (2/3/4), and number comparisons (large/small distances) were significant (see Table 4). Subsequent pairwise comparisons based on the rics model showed that DD children required significantly longer RTs than TD children ($p < 0.0001$). RTs decreased across grades (g2 > g3, $p < 0.05$, except g4 vs. g3). As expected, smaller numerical distances led to longer RT ($p < 0.0001$) (see Figure 3A). No interactions were found. Using reading fluency as a fixed effect showed that higher reading fluency was associated with lower RTs on the number comparison task, but the effect of IQ was not significant. A fixed effect of the school indicated that students from school A required longer RTs relative to school H.

3.3 Mixed comparison

The predictors group (typical/dyscalculic), grade (2/3/4), and mixed comparison (small/large distances) were significant. Subsequent pairwise comparisons based on the rics model revealed longer RTs for DD than TD children ($p < 0.01$), a systematic decrease in RTs across grades (g2 > g3 > g4, $ps < 0.001$), and longer RTs for small than for large distances ($p < 0.0001$). Figure 3B shows that children in all grades needed longer RTs for small distances, especially in 2nd grade. This may have resulted in the mixed comparison × grade interaction. The pairwise comparisons showed that children's RTs decreased across grades when comparing the small distances ($ps <$ .001). In contrast, there was only a significant difference between 2nd and 3rd graders ($p <$ .001) for large distances. We found no evidence of a mixed comparison × group effect

TABLE 4 Results of the linear mixed model using task condition, group, grade, reading fluency, and intelligence to predict children's performance on number comparison.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	1253.521	82.264	15.238	28.7	< 0.001
NC[a]	109.712	11.386	9.636	14.0	< 0.001
Group[b]	270.653	79.434	3.407	9.0	0.008
Grade[c] 3	−142.202	33.080	−4.299	33.0	< 0.001
Grade 4	−225.639	35.220	−6.407	8.5	< 0.001
Reading fluency	−1.894	0.457	−4.144	26.3	< 0.001
IQ	−0.380	0.787	−0.484	29.7	0.632
School A [d]	235.377	33.430	7.041	4.9	0.001
NC × group	−17.888	60.545	−0.295	8.5	0.775
NC × grade 3	−21.079	14.617	−1.442	26.6	0.161
NC × grade 4	−41.612	19.944	−2.086	6.9	0.076
Group × grade 3	−180.461	88.724	−2.034	16.4	0.058
Group × grade 4	−164.044	83.246	−1.971	9.5	0.079
NC × group × grade 3	84.623	68.657	1.233	16.2	0.235
NC × group × grade 4	66.230	74.782	0.886	10.1	0.396
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	186.59				
$\tau_{00\ IDclass}$	63.26				
δ^2	90.74				
ICC $_{IDstudent:\ IDclass}$	0.678				
ICC $_{IDclass}$	0.180				

The model summary based on *p*-values for fixed effects was calculated using Satterthwaite approximations. NC = number comparison.
[a]Large distance is the reference.
[b]Typical developed is the reference.
[c]Grade 2 is the reference.
[d]School H is the reference.

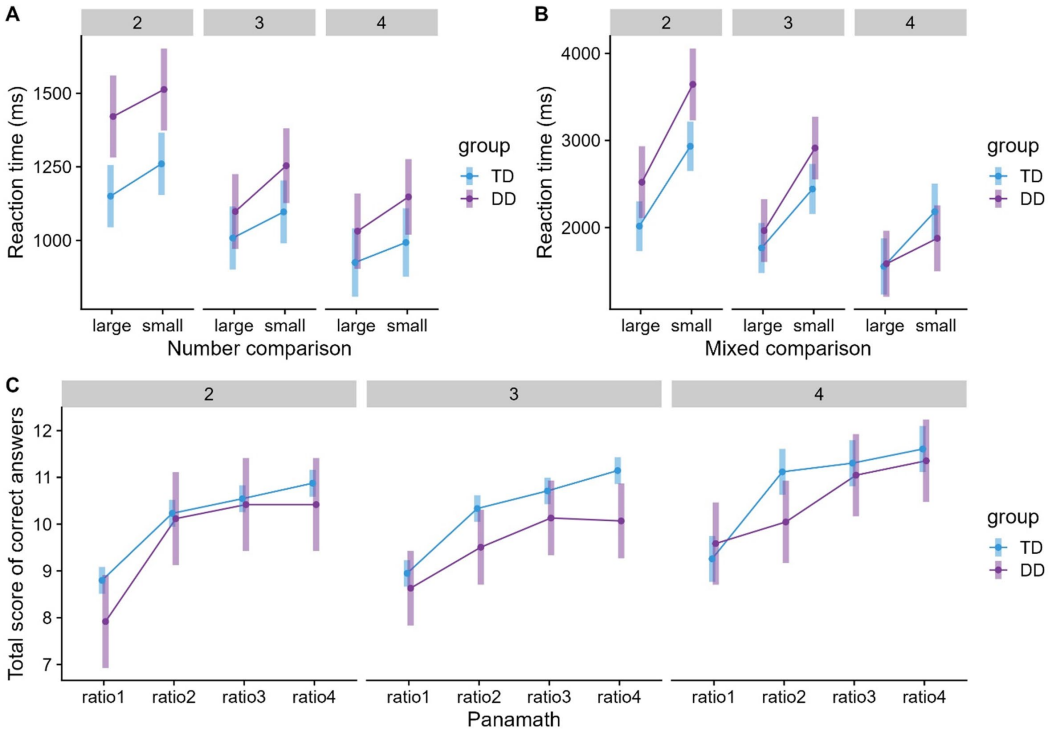


FIGURE 3
Means and 95% confidence intervals (CIs) for **(A)** number comparison, **(B)** mixed comparison and **(C)** Panamath grouped by group and grade. Ratio 1 = 1.2; ratio 2 = 1.4; ratio 3 = 1.6; and ratio 4 = 2.6 (12 items each).

(see Table 5). However, the large difference between children with and without DD for small distances in 2nd grade (see Figure 3B) may have resulted in a significant interaction of mixed comparison × group × grade. Subsequent comparisons showed that in 2nd and 3rd grades, DD children needed longer RTs in small distances compared to TD children ($ps < 0.05$). At large distances, children with and without DD did not differ significantly ($ps > 0.05$). Grade level trends within groups showed that RTs of DD children for large distances decreased from 2nd to 3rd grade (DD2 > DD3 = DD4, $p < 0.01$), but there was no significant development for TD children (TD2 = TD3 = TD4, $ps > 0.05$). In small distances, TD children reached a stable level as early as 3rd grade (TD2 > TD3 = TD4, $p < 0.001$). In contrast, DD children continued to develop until 4th grade (DD2 > DD3 > DD4, $ps < 0.05$). Using reading fluency and IQ as fixed effects, higher IQ was associated with lower RTs in the mixed comparison task. The effect of reading fluency was not significant. In addition, it was found that belonging to school A was associated with poorer performance compared to school H (see Table 5).

3.4 Dot magnitude comparison (Panamath)

Not all children completed this task due to a computer technical problem. Data for the Panamath task were only available for $n = 322$ children ($n = 39$ with DD). TD children's total score was not significantly higher than DD children's score (see Table 6). The predictor Panamath (ratios: 1.2/1.4/1.6/2.4) was significant. Subsequent pairwise comparisons based on the rics model showed that the total test score of correct answers decreased systematically over the ratios (r4 > r3 > r2 > r1, $ps < 0.01$, except r3 vs. r4). Furthermore, an interaction of Panamath × group × grade was found. Pairwise comparisons focused on group differences within a grade level and condition (ratio) revealed no significant differences between TD and DD children. However, children's total test scores of correct answers increased from ratio 1 to ratio 2 (also see Figure 3C). While this was evident for TD children in all grades, for DD children, the effect was only found in 2nd grade ($ps < 0.01$). Reading fluency and IQ were not associated with the Panamath task (see Table 6).

3.5 Dot magnitude (non-symbolic) vs. number (symbolic) comparison

The predictors grade (2/3/4), distance (small/large), and task (non-symbolic/symbolic) were significant. Subsequent pairwise comparisons based on the rics model indicated that RTs decreased from 2nd to 4th grade ($p < 0.0001$) but not systematically across grades. There was no significant difference between 2nd and 3rd grade (see Table 7). Small distances resulted in longer RTs than large distances ($p < 0.0001$) and symbolic tasks resulted in longer Rts than non-symbolic tasks ($p < 0.01$). The interaction of task × distance showed that children required longer RTs for small distances than for large distances in both task conditions ($ps < 0.01$). Additionally, large distances in symbolic comparisons resulted in higher RTs than large distances in non-symbolic comparisons ($p < 0.01$). In contrast, the mean difference between small distances in symbolic comparisons and small distances in non-symbolic comparisons was not significant ($p = 0.631$). The task × group interaction showed that DD children required substantially longer RTs to compare symbolic and non-symbolic tasks than the TD children ($ps < 0.05$). Furthermore, within-group effects showed that in contrast to TD children, DD children did not need longer RTs for

TABLE 5 Results of the linear mixed model using task condition, group, grade, reading fluency, and intelligence to predict children's performance on mixed comparison.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	2400.484	192.238	12.487	28.4	< 0.001
MC[a]	916.887	72.139	12.710	14.0	< 0.001
Group[b]	504.641	239.509	2.107	9.0	0.065
Grade[c] 3	−250.340	58.506	−4.279	30.6	< 0.001
Grade 4	−462.388	74.686	−6.191	8.1	< 0.001
Reading fluency	−1.325	2.141	−0.619	26.3	0.541
IQ	−5.489	1.963	−2.797	30.2	0.009
School A[d]	599.333	76.148	7.871	4.6	0.001
MC × group	206.172	193.653	1.065	8.5	0.316
MC × grade 3	−240.143	85.361	−2.813	26.6	0.009
MC × grade 4	−288.414	85.098	−3.389	6.9	0.012
Group × grade 3	−305.458	276.649	−1.104	16.3	0.286
Group × grade 4	−474.660	267.062	−1.777	9.6	0.107
MC × group × grade 3	65.113	232.200	0.280	16.2	0.783
MC × group × grade 4	−541.909	230.088	−2.355	10.1	0.040
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	490.44				
$\tau_{00\ IDclass}$	138.33				
δ^2	471.95				
ICC $_{IDstudent:\ IDclass}$	0.483				
ICC $_{IDclass}$	0.100				

The model summary based on p-values for fixed effects was calculated using Satterthwaite approximations. MC = mixed comparison.
[a]Large distance is the reference.
[b]Typical developed is the reference.
[c]Grade 2 is the reference.
[d]School H is the reference.

TABLE 6 Results of the linear mixed model using task condition, group, grade, reading fluency, and intelligence to predict children's performance on Panamath.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	7.323	0.587	12.471	16.8	< 0.001
Ratio 2[a]	1.432	0.162	8.838	7.4	< 0.001
Ratio 3	1.746	0.144	12.114	7.4	< 0.001
Ratio 4	2.076	0.142	14.671	7.4	< 0.001
Group[b]	−0.881	0.386	−2.281	6.8	0.058
Grade[c] 3	0.150	0.230	0.652	15.4	0.524
Grade 4	0.459	0.301	1.523	5.9	0.180
Reading fluency	0.003	0.003	0.810	15.8	0.430
IQ	0.012	0.006	2.076	19.3	0.052
Ratio 2 × group	0.768	0.428	1.795	6.5	0.119
Ratio 3 × group	0.754	0.465	1.623	6.5	0.152
Ratio 4 × group	0.424	0.787	0.539	6.5	0.608
Ratio 2 × grade 3	−0.047	0.254	−0.185	15.5	0.856
Ratio 3 × grade 3	0.017	0.204	0.081	15.5	0.937
Ratio 4 × grade 3	0.120	0.222	0.542	15.5	0.596
Ratio 2 × grade 4	0.428	0.335	1.277	5.1	0.257
Ratio 3 × grade 4	0.301	0.280	1.074	5.1	0.331
Ratio 4 × grade 4	0.273	0.284	0.961	5.1	0.380
Group × grade 3	0.563	0.760	0.741	12.3	0.472
Group × grade 4	1.208	0.679	1.779	7.8	0.114
Ratio 2 × group × grade 3	−1.278	0.591	−2.162	12.3	0.051
Ratio 3 × group × grade 3	−1.017	0.510	−1.993	12.3	0.069
Ratio 4 × group × grade 3	−1.183	0.974	−1.215	12.3	0.247
Ratio 2 × group × grade 4	−2.167	0.888	−2.441	7.8	0.041
Ratio 3 × group × grade 4	−1.339	0.773	−1.732	7.8	0.123
Ratio 4 × group × grade 4	−1.003	0.948	−1.059	7.8	0.321
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	1.15				
$\tau_{00\ IDclass}$	0.00				
δ^2	1.09				
ICC $_{IDstudent:\ IDclass}$	0.534				
ICC $_{IDclass}$	0.006				

The model summary based on *p*-values for fixed effects was calculated using Satterthwaite approximations.
[a]Ratio 1 is the reference.
[b]Typical developed is the reference.
[c]Grade 2 is the reference.

symbolic than non-symbolic tasks. Subsequent pairwise comparisons of the task × group × grade interaction, focusing on group differences within a grade level and a task condition, revealed that children with and without DD did not differ in any grade level. Grade-level trends within groups showed that there was no significant development in non-symbolic and symbolic tasks for DD and TD children. With one exception: TD children's RTs for symbolic tasks decreased from 2nd to 3rd grade ($ps < 0.01$). Furthermore, DD and TD children substantially needed longer RTs to compare symbolic than non-symbolic tasks in 2nd grade ($ps < 0.01$), but there were no significant differences in 3rd or 4th grade ($p = 1.000$). When rerunning analyses without the 2nd graders, the task × group effect vanished. No significant effects were found for the covariates, but children from school A showed poorer performance in comparison to children from school H (see Table 7).

3.6 Number sets

The predictors group (typical/dyscalculic) and grade (2/3/4) were significant. Subsequent pairwise comparisons based on the rics model indicated a better test score for the TD than DD children ($p < 0.0001$), and an improvement across grades (g2 > g3 > g4, $ps < 0.01$, also see Figure 4A). Using reading fluency and IQ as fixed effects, both were

TABLE 7 Results of the linear mixed model using task conditions, group, grade, reading fluency, and intelligence to predict children's performance on dot magnitude and number comparisons.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	1137.349	138.610	8.205	17.0	< 0.001
Task[a]	155.792	41.326	3.770	7.4	0.006
Distance[b]	214.475	23.899	8.974	7.4	< 0.001
Group[c]	−4.594	74.967	−0.061	6.7	0.953
Grade 3[d]	−93.850	52.314	−1.794	16.0	0.092
Grade 4	−203.681	60.157	−3.386	5.8	0.016
Reading fluency	−0.551	0.812	−0.678	16.0	0.507
IQ	−1.411	1.315	−1.073	19.2	0.296
School A[e]	176.186	46.803	3.764	3.4	0.026
Task × distance	−111.449	29.202	−3.816	7.4	0.006
Task × group	254.008	85.669	2.965	6.5	0.023
Distance × group	43.225	144.160	0.300	6.5	0.774
Task × grade 3	−82.182	56.917	−1.444	15.5	0.169
Task × grade 4	−107.374	42.616	−2.520	5.1	0.053
Distance × grade 3	−60.671	29.796	−2.036	15.5	0.059
Distance × grade 4	−59.475	25.645	−2.319	5.1	0.068
Group × grade 3	279.338	280.175	0.997	12.2	0.338
Group × grade 4	110.484	123.294	0.896	7.6	0.398
Task × distance × group	−16.901	168.704	−0.100	6.5	0.923
Task × distance × grade 3	54.146	35.177	1.539	15.5	0.144
Task × distance × grade 4	26.821	33.322	0.805	5.1	0.457
Task × group × grade 3	−449.712	268.516	−1.675	12.3	0.119
Task × group × grade 4	−228.849	92.204	−2.482	7.8	0.039
Distance × group × grade 3	−129.560	165.147	−0.785	12.3	0.448
Distance × group × grade 4	146.159	151.149	0.967	7.8	0.363
Task × distance × group × grade 3	145.173	187.632	0.774	12.3	0.454
Task × distance × group × grade 4	−129.125	196.356	−0.658	7.8	0.530
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	228.17				
$\tau_{00\ IDclass}$	30.48				
δ^2	271.79				
ICC $_{IDstudent:\ IDclass}$	0.396				
ICC $_{IDclass}$	0.057				

The model summary based on p-values for fixed effects was calculated using Satterthwaite approximations.
[a]Non-symbolic is the reference.
[b]Large distance is the reference.
[c]Typical developed is the reference.
[d]Grade 2 is the reference.
[e]School H is the reference.

significant. A higher level of IQ and reading fluency was associated with a better test score (see Table 8).

3.7 Number line estimation

The predictors group (typical/dyscalculic) and grade (2/3/4) were significant. Subsequent pairwise comparisons based on the rics model indicated a larger average deviation in the DD group than for the TD group ($p < 0.0001$), and a decrease in average deviation across grades (g2 > g3, g4, except g3 vs. g4, $ps < 0.001$). Figure 4B shows that especially DD children in 2nd grade were less accurate. A relatively small decrease in the mean deviations from 3rd to 4th grade resulted in a group × grade interaction. Pairwise comparisons of the interaction showed that the precision of arranging the digit on the number line increased within groups, but both groups reached a stable level as early as 3rd grade. Furthermore, the group × grade interaction revealed that only DD children in 2nd grade were less accurate than TD children ($p < 0.001$). There were no differences between the two groups in the 3rd and 4th grades ($ps > 0.05$). Further fixed effects of reading fluency and IQ were significant. A higher level of IQ and reading fluency was associated with a better test score (see Table 9).

3.8 DD in 4th grade vs. TD in 2nd grade

For the profile analyses, the variables were z-standardized. Accordingly, scale differences between the tasks largely disappeared (Table 10). The predictor group (typical/dyscalculic) was significant. Subsequent pairwise comparisons based on the rics model showed that on average (across tasks), the TD children from the 2nd grade performed substantially worse than DD children in the 4th grade. No interaction between subgroup and grade was found, indicating a delayed basic numerical profile in DD children.

4 Discussion

We investigated whether DD children's basic numerical skills are qualitatively different compared to TD or, instead, if there is a developmental delay for DD children (RQ 1). To answer this question, we investigated a wide range of basic numerical paradigms and compared basic numerical profiles of DD children in 4th grade with TD children in 2nd grade. Finally, competing hypotheses on the causes of DD were compared and it was examined whether these assumptions are stable across different grades (RQ 2). Because substantial mathematical development can be observed in the first years of schooling, we considered the children's grade level (2–4) as a possible influencing factor. Thereby, we ruled out that differences between children with and without DD were due to mathematical age development.

Regarding the first RQ, DD children consistently displayed deficits in core markers of numeracy, with the exception of subitizing and dot magnitude comparison. DD children's difficulties were manifested in lower efficiency and/or accuracy. Apart from counting, we found no evidence that DD children's basic numerical skills were qualitatively different from those of TD children. Disproportionate impairments in processing numerosities from symbols and the qualitatively different

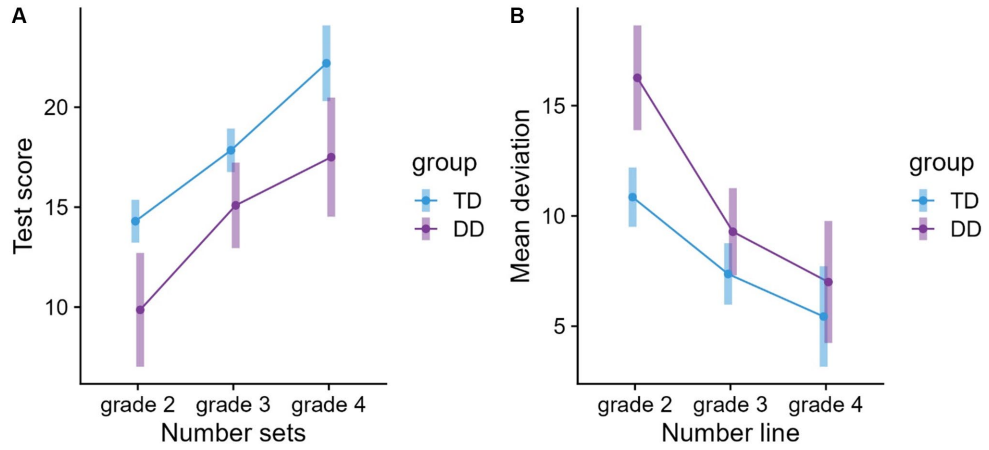


FIGURE 4
Means and 95% confidence intervals (CIs) for **(A)** number sets and **(B)** number line estimation grouped by grade and group.

TABLE 8 Results of the linear mixed model using group, grade, reading fluency, and intelligence to predict children's performance on number sets.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	−8.391	2.844	−2.951	28.4	0.006
Group[a]	−4.433	1.614	−2.746	9.1	0.022
Grade 3[b]	3.547	0.746	4.757	31.4	< 0.001
Grade 4	7.900	0.776	10.177	8.5	< 0.001
Reading fluency	0.104	0.018	5.876	26.3	< 0.001
IQ	0.125	0.022	5.633	30.2	< 0.001
Group × grade 3	1.680	1.976	0.850	16.4	0.407
Group × grade 4	−0.264	2.857	−0.092	9.9	0.928
Random effects					
$\tau_{00\ IDclass}$	1.31				
δ^2	5.62				
$ICC_{IDclass}$	0.179				

The model summary based on *p*-values for fixed effects was calculated using Satterthwaite approximations.
[a]Typical developed is the reference.
[b]Grade 2 is the reference.

TABLE 9 Results of the linear mixed model using group, grade, reading fluency, and intelligence to predict children's performance on number line.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	24.266	3.040	7.981	29.6	< 0.001
Group[a]	5.412	1.279	4.232	9.0	0.002
Grade 3[b]	−3.484	1.012	−3.443	36.4	0.001
Grade 4	−5.413	0.885	−6.114	9.7	< 0.001
Reading fluency	−0.048	0.016	−3.056	25.9	0.005
IQ	−0.087	0.018	−4.798	28.4	< 0.001
Group × grade 3	−3.496	1.440	−2.428	16.5	0.027
Group × grade 4	−3.847	1.722	−2.234	9.6	0.051
Random effects					
$\tau_{00\ IDclass}$	2.58				
δ^2	4.11				
$ICC_{IDclass}$	0.380				

The model summary based on p-values for fixed effects was calculated using Satterthwaite approximations.
[a]Typical developed is the reference.
[b]Grade 2 is the reference.

distance effect in the mixed comparison task were moderated by grade level. The results suggest that children with DD exhibit developmentally delayed rather than qualitatively different numerical processing. Additionally, DD children showed a different developmental trajectory for mixed comparison, dot magnitude vs. number comparison, and number line, but in line with the profile analysis, this can be interpreted as a developmental delay. Overall, the results provided evidence for delayed numerical processing in DD.

The results regarding the question of the causes of DD supported the defective ANS rather than the AD hypothesis (see Table 11). We found no evidence for the OTS deficit hypothesis. These results appeared consistent across grades. In the following, we address the results pertaining to each task in detail.

4.1 Dot enumeration

The results confirmed prior studies (e.g., Andersson and Östergren, 2012; Decarli et al., 2020, 2023): Larger point sets led to longer RTs, supporting a qualitative difference in the perception and representation of large vs. small quantities. In contrast to previous studies (e.g., Kuhn et al., 2013; Decarli et al., 2023) and the OTS defective hypothesis, we found no evidence of a deficit in the subitizing range (similar to Skagerlund and Träff, 2014). The present results suggest, in line with the defective ANS hypothesis, that DD children's counting skills were disproportionately impaired. This is unusual because most studies showed that children with DD were additively or over-additively impaired in subitizing (e.g., Landerl, 2013).

TABLE 10 Results of the linear mixed model using task condition, group, grade, reading fluency, and intelligence to predict children's performance on all basic numerical tasks.

Predictors	*B*	Robust S.E.	*t*	*df*	*p*
(Intercept)	0.233	0.401	0.580	10.6	0.574
DE counting[a]	0.375	0.233	1.609	3.6	0.190
NC	0.482	0.183	2.636	3.6	0.064
MC	−0.020	0.191	−0.105	3.6	0.922
NS	0.339	0.394	0.861	3.2	0.448
NL	−0.019	0.515	−0.038	3.6	0.972
Panamath	0.280	0.429	0.651	3.6	0.554
Group[b]	0.737	0.271	2.721	4.7	0.045
Reading fluency	−0.004	0.002	−2.102	10.7	0.060
IQ	−0.003	0.003	−0.937	12.3	0.367
DE counting[a] × group	−0.470	0.240	−1.960	4.3	0.116
NC × group	−0.501	0.193	−2.600	4.3	0.056
MC × group	−0.067	0.222	−0.300	4.3	0.778
NS × group	−0.854	0.419	−2.036	3.9	0.113
NL × group	−0.591	0.536	−1.104	4.3	0.327
Panamath × group	−0.326	0.461	−0.708	4.3	0.515
Random effects					
$\tau_{00\ IDstudent:\ IDclass}$	0.14				
$\tau_{00\ IDclass}$	0.23				
δ^2	0.92				
ICC $_{IDstudent:\ IDclass}$	0.025				
ICC $_{IDclass}$	0.043				

The model summary based on *p*-values for fixed effects was calculated using Satterthwaite approximations. NC = number comparison; MC = mixed comparison; NS = number sets; NL = number line; Panamath, median of the children's RTs for correct answers.
[a]Subitizing is the reference.
[b]Dyscalculic grade 4 is the reference.

TABLE 11 Results of impairments in line with the ANS deficit hypothesis, OTS deficit hypothesis, and AD hypothesis.

Tasks	ANS	OTS	AD	Results
Dot enumeration subitizing range (1–3 / 1–4)	Not impaired	Impaired	Not impaired	Not impaired
Dot enumeration counting range (4–9 / 5–9)[2]	Impaired	Not impaired	Not impaired	Qualitative impaired
Number comparison[1,2]	Impaired	Not impaired	Impaired	Delayed impaired
Mixed comparison[1,2]	Impaired	(Not impaired)*	Impaired	Delayed impaired
Dot magnitude comparison/ Panamath[1]	Impaired	Not impaired	Not impaired	Not impaired
Number sets[3]	Impaired	(Impaired)*	Impaired	Delayed impaired
Number line estimation[1]	Impaired	Not impaired	Impaired	Delayed impaired

The study's tasks substantially tap the analog magnitude,[1] verbal-phonological,[2] and visual-Arabic[3] module of the Triple Code Model by Dehaene (1992).
*Tasks cannot differentiate sharply between hypotheses because tasks substantially tap the ANS and visual-Arabic module, but points (tapping the ANS) can be in the subitizing (tapping OTS) or counting range (tapping verbal-phonological module). ANS deficit, general deficit in representing numerosities; OTS deficit, specific deficits in an object-based attentional system; AD, deficits in accessing numerosities from symbols (based on Andersson and Östergren, 2012).

However, we implemented a data-driven approach to determine the subitizing range. Other authors (e.g., Kuhn et al., 2013) set the subitizing range at three dots. Due to the variability in the subitizing range, it is possible that some studies have mixed the counting and subitizing skills, resulting in lower RTs in the counting range. We found no evidence that DD children's over-additive impairment in the counting range was moderated by age development (grade level). Nevertheless, the disproportionately impaired counting skills in DD must be interpreted with some reservations because prior studies that found no evidence of a qualitatively different approach to counting did not include children with comorbid reading disorders (e.g., Schleifer and Landerl, 2011; Kuhn et al., 2013). In our study, a higher level of reading fluency was associated with lower RTs. Thus, verbal problems could cause a deficit in counting efficiency (see Schleifer and Landerl, 2011). Averaged across all items 1–9 (not differentiating between the counting and subitizing range), the differences between children with and without DD were moderated by grade level. Children with DD performed poorer than TD children in the 2nd and 3rd grade. However, in the 4th grade, there were no longer substantial group differences. Grade level trends within both groups reached a stable level in the 3rd grade, suggesting that DD children catch up and are merely developmentally delayed. Our results showed that counting skills and subitizing developed differently (similar to Schleifer and Landerl, 2011). Complementary to the findings of Schleifer and Landerl (2011), a stable competence level in subitizing was achieved in the 3rd grade, thus earlier than in counting. All in all, the results showed disproportionately impaired counting skills in DD. Future studies should determine the subitizing and counting range data-driven and consider the influence of reading-related difficulties.

4.2 Number comparison

The current results revealed that smaller numerical distances resulted in longer RTs, and DD children performed poorer than TD children but showed no greater distance effect than TD children (similar to Holloway and Ansari, 2008; Decarli et al., 2020, 2023). In line with Landerl and Kölle (2009), children with and without DD became more efficient across grades, but there was no evidence that the distance effect was moderated by grade level (similar to Holloway and Ansari, 2009; Landerl and Kölle, 2009; Reeve et al., 2012), or that the developmental trajectory of children with DD differed from that of TD children. As deficits in symbolic numerical processing are associated with an impaired ANS and AD, number comparisons cannot clearly discriminate between the hypotheses. However, it is certain that, for DD children tapping the visual-Arabic module causes

problems. Although we found a fixed effect for reading fluency, we do not assume that our result was significantly confounded by this covariate because studies (e.g., Decarli et al., 2023) showed that children with DD and combined impaired children did not differ.

4.3 Mixed comparison

As expected, smaller distances resulted in longer RTs. This distance effect was evident in all grades. A stable level of competence was reached earlier in large distances than in small distances. Similar to the results of Kuhn et al. (2013), DD children performed worse than TD children. Children with and without DD did not differ in large distances, but DD children required longer RTs in small distances in the 2nd and 3rd grades. At first glance, a larger distance effect was observed in children with DD, in line with the ANS deficit (similar to Mussolin et al., 2010). However, there was no significant difference between DD and TD in small distances in the 4th grade. Furthermore, no interaction between mixed comparison and group was found. Thus, the result implies that children's grade level moderates the distance effect, indicating that DD children's performance is developmentally delayed. The result contradicts other studies that found no over-additive distance or age effects (see Schwenk et al., 2017). However, the finding may be related to the specificity of the task condition. Most studies compared symbolic or non-symbolic tasks, but not a mixture of both conditions. The over-additive deficit in comparing small distances is associated with 2nd and 3rd grade children's ability to compare two quantities in different modalities (e.g., dot sets and Arabic numerals). Based on the results, we cannot draw firm conclusions about the hypotheses regarding the causes of DD (see Table 1). We found that the higher the IQ, the lower the RT. IQ should be further investigated in future studies. It cannot be excluded that IQ plays a central role in solving this task. Nevertheless, the results extend the current state of research because mixed comparisons were rarely used, and when they were, the grade level of the DD and TD children was not systematically investigated.

4.4 Dot magnitude comparison (Panamath)

In line with previous findings (e.g., Rousselle and Noël, 2007; Landerl and Kölle, 2009; Decarli et al., 2023), children with DD showed no significant problems with non-symbolic magnitude comparisons, contradicting the ANS deficit hypothesis.

The causes of heterogeneous evidence regarding non-symbolic comparisons continue to be debated. Some authors discussed whether difficulties in non-symbolic comparisons are associated with severe mathematical difficulties (Wong et al., 2017). The discussion stemmed from the fact that studies found no impairments (e.g., Rousselle and Noël, 2007) using less stringent cutoff criteria to classify DD ($PR = 15$). In contrast, studies that have demonstrated differences (e.g., Piazza et al., 2010) used more stringent criteria (2 standard deviations below average, similar to the current study). We found no evidence of a deficit in DD despite the strict criterion, contradicting the assumption. Additionally, some authors argued that deficits in the non-symbolic comparison tasks result from poor mathematical development (see de Smedt et al., 2013), supporting the AD hypothesis. We investigated whether children's grade level affects differences between children with and without DD. The present study could not confirm this assumption (similar to Skagerlund and Träff, 2014). However, Skagerlund and Träff (2014) investigated not only correctly answered trials (as the current study did), arguing that Weber fractions would be a more sensitive measure of performance on this task. Their findings showed that in line with the ANS hypothesis, DD children in 4th grade showed noisier ANS representations than TD children in 4th and 2nd grade. In other words, different test scoring methods also have an impact on results. Furthermore, methodological aspects of the task itself were discussed. Studies that used small ratios (e.g., Wong et al., 2017) found differences between children with and without DD compared to studies that used rather large ratios (e.g., Iuculano et al., 2008). We used small and large ratios and showed that DD and TD children did not differ in small ratios. However, the results showed that it is more challenging to compare small distances. In the TD group, it is clear that the children performed better at ratio 2 than at ratio 1. This distance effect is observed in all grades. In the DD group, this effect is less pronounced. DD children only performed better in 2nd grade. The developmental trajectories seem to differ, at least for ratios 1 and 2.

4.5 Dot magnitude (Panamath) vs. number comparison

To test whether DD children were disproportionately more impaired in symbolic comparisons, we contrasted children's performance in symbolic vs. non-symbolic comparisons and distance effects for both tasks within one analysis. This methodical approach has rarely been used, and when it has been used, it has yet to be focused on children with DD (e.g., Holloway and Ansari, 2008). Similar to Kuhn et al. (2013), Landerl (2013), and Landerl and Kölle (2009), we found no interaction between group and distances. The result argues against a qualitative distance effect for children with DD. Similar to Schwenk et al. (2017), DD children required substantially longer RTs than TD children, particularly on symbolic comparisons. At first glance, this finding suggests a disproportionate impairment in symbolic tasks, consistent with the AD hypothesis. Pairwise comparisons of the interaction among task, group, and grade showed that TD and DD children did not differ at any grade level, suggesting that the disproportionate impairment could not be seen at grade level due to small group sizes. However, for both groups, we found that in the 2nd grade, symbolic magnitude comparisons resulted in significantly longer RTs than non-symbolic tasks. Thus, we reran the analysis without the 2nd graders and showed that the interaction between task and group effect disappeared. Against this background, the AD hypothesis seems to be tenable only in the 2nd grade.

4.6 Number sets

Consistent with other studies (e.g., Kuhn et al., 2013; von Wirth et al., 2021), DD children displayed difficulties in the number set task. DD children's impairments were stable in different grades. There was

no evidence suggesting a different developmental trajectory for children with DD. Consistent with previous research (e.g., Brankaer et al., 2014), children's mapping skills developed across grade levels. Whether the deficits are more likely to result from deficient ANS/OTS or are indicative of AD remains unanswered, as both hypotheses predict deficits in this task.

4.7 Number line estimation

The results of this study are in line with previous research (e.g., Booth and Siegler, 2006; Decarli et al., 2023) and showed that children with DD displayed problems locating numbers on the number line. Children with DD were less accurate than TD children in 2nd grade; by the 3rd and 4th grades, there were no longer substantial group differences, suggesting DD children catch up and are merely developmentally delayed. Whether DD children in 3rd grade really catch up or if the task is not sensitive enough due to the small number range (0–100) remains unanswered, as Landerl (2013) examined number line estimation with a range of 0–1,000 and found that DD children became more accurate by 4th grade. In general, our results are in line with the defective ANS and AD hypotheses (see Wilson and Dehaene, 2007). In our study, children were only tested on locating written Arabic numerals on the number line. However, future studies should expand the task to rule out an ANS deficit, as Lafay et al. (2017) showed that DD children had no difficulty placing non-symbolic numerosities on the number line, suggesting that the mental number line or ANS is not damaged per se.

4.8 DD in 4th grade vs. TD in 2nd grade

Finally, we investigated whether the basic numerical profile of DD children in 4th grade is qualitatively different from that of TD children in 2nd grade or whether there is a developmental delay in DD children. The results revealed that DD children in 4th grade performed better than TD children in 2nd grade and caught up with TD children. Adapted from the ability-level-match design (Bradley and Bryant, 1978), this finding suggests that children with DD are developmentally delayed by less than 2 school years. As we found no interaction between tasks and group, the results do not indicate a task-specific DD profile, nor that DD children are disproportionately impaired. The present results confirm and complement the findings of previous studies. Skagerlund and Träff (2014) found no differences in RTs of basic numerical skills between DD children in 4th grade compared to a math ability-matched control group of TD children in 2nd grade, suggesting that the abnormalities are due to developmental delay. However, Skagerlund and Träff (2014) additionally used the Weber fraction and found that DD children in 4th grade had noisier ANS representations than TD children in 4th and 2nd grades. Thus, it may be useful in future to include the Weber fraction, which may be more sensitive to performance measures. Future studies should investigate whether this finding is robust and thus replicable in other samples. To extend the analyses and to be able to make statements about whether this indication of developmental delay persists in the long term, longitudinal studies comparing the basic numerical profiles of children with and without DD in higher grades would be interesting (e.g., DD children in 6th grade compared to TD children in 4th grade).

5 Limitations and suggestions for future research

Our results are more consistent with the ANS than with the AD (deficit) hypotheses and point against a deficit of the OTS. However, recent evidence suggests that there is no core cognitive deficit in DD. Individuals with DD may have deficits in basic numerical processing and domain-general cognitive abilities, but neither is necessarily present (Mammarella et al., 2021). Mammarella et al. (2021) argue that it is more fruitful to locate children with mathematical difficulties in a multidimensional space that reflects the severity of their difficulties and their relative position with respect to various domain-specific and cross-domain influences on mathematical performance. Subsequent studies should take this approach as a starting point for their analyses rather than focusing solely on domain-specific deficits.

Reading ability should also be considered because although recent evidence suggests that impairments in basic numerical skills are clearly associated with DD but not with a reading disorder (Raddatz et al., 2017; Decarli et al., 2023), we found that reading fluency affected basic numerical skills (e.g., dot enumeration). The same was true for IQ (e.g., mixed comparison).

To improve our understanding of the impairments of children with DD, it may be methodologically helpful to focus on a Bayesian statistical approach and to compare different competing models and theories.

Furthermore, our study compared independent groups in a cross-sectional design. Future longitudinal studies and research using the ability-level-match design are needed to definitively answer the intriguing question of whether DD children's basic numerical abilities are qualitatively different from TD or whether DD children are developmentally delayed.

6 Conclusion

Consistent with the ANS deficit rather than the AD hypothesis, DD children consistently showed deficits in basic numerical skills. We found no evidence of deficits in subitizing, contradicting a disrupted OTS. The disproportionate impairment in processing numerosities from symbols and the qualitatively different distance effect in the mixed comparison task were moderated by grade level. Both grade level effects indicate that children with DD have a developmentally delayed rather than qualitatively different numerical processing. We found significant improvements in children's performance with increasing grade levels on all tasks except Panamath. The results suggest developmental leaps between 2nd and 3rd graders. For mixed comparison, dot magnitude vs. number comparison, and number line, children with DD had a different developmental trajectory than TD children. However, the results indicate that children with DD have a developmentally delayed rather than a qualitatively different basic numerical profile. The only disproportionate DD impairment that was not moderated by grade level relates to the counting range of the dot enumeration task. This

result emphasizes the potential of (pre)-school identification, prevention, and intervention initiatives. Verbal counting indicates risks related to math abilities and predicts math achievement in the long term (Koponen et al., 2019). This learned skill (see Krajewski and Schneider, 2009) builds on innate core systems (von Aster and Shalev, 2007). Therefore, interventions should start before further basic numerical skills develop poorly due to defective innate core systems. Moraske et al. (2019) showed that early intervention of basic numerical skills in kindergarten leads to improved later math performance in DD at-risk children and reduces the likelihood of the onset of dyscalculia.

Most of the evidence points to a developmentally delayed basic numerical profile in DD. If teachers have evidence-based knowledge about the causes (von Aster and Shalev, 2007) and deficits (e.g., Butterworth, 2010; Kuhn et al., 2013) in DD, they could locate the deficits of the DD child in the developmental stage (e.g., Krajewski and Schneider, 2009) and initiate adaptive interventions according to the response-to-intervention approach (Voß et al., 2014). This approach has great potential, as appropriate interventions can improve children's performance significantly (Chodura et al., 2015). However, the first step to appropriate intervention is to identify children with DD who are at risk. The fact that even basic numerical skill tasks discriminate between children with and without DD in primary school underlines the importance of teachers including basic numerical skills when identifying or supporting children with DD. To support teachers in identifying children with DD, future studies should focus on the development of simple screening tools.

Data availability statement

The datasets presented in this article are not readily available because the data will be used for further analyses and publications. Requests to access the datasets should be directed to tobias.kuhn@tu-dortmund.de.

Ethics statement

The studies involving involving human participants were reviewed and approved by the Ethikkommission des Fachbereichs 7, Psychologie und Sportwissenschaft, Westfälische Wilhelms-Universität Münster (ethics committee of Faculty 7 - Psychology and Sports Sciences, University of Münster). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

J-TK: contributed the idea for writing this manuscript, funding acquisition, and data collection. J-TK and SL: statistical analysis. SL: created the basic structure of the manuscript (lead). J-TK and FK: provided the writing oversight and revision of the draft. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01GJ1302.

Acknowledgments

The authors would like to thank all the children who participated in our study. In addition, they would like to thank all the employees who supported us in the data collection. The authors acknowledge the financial support by the Deutsche Forschungsgemeinschaft and Technische Universität Dortmund/TU Dortmund University within the funding program open access costs.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1187785/full#supplementary-material

References

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders. 5th Edn. Washington, DC: American psychiatric association.

Andersson, U., and Östergren, R. (2012). Number magnitude processing and basic cognitive functions in children with mathematical learning disabilities. *Learn. Individ. Differ.* 22, 701–714. doi: 10.1016/j.lindif.2012.05.004

Booth, J. L., and Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Dev. Psychol.* 42, 189–201. doi: 10.1037/0012-1649.41.6.189

Bradley, L., and Bryant, P. E. (1978). Difficulties in auditory organisation as a possible cause of reading backwardness. *Nature*, 271, 746–747. doi: 10.1038/271746a0

Brankaer, C., Ghesquière, P., and de Smedt, B. (2014). Children's mapping between non-symbolic and symbolic numerical magnitudes and its association with timed and untimed tests of mathematics achievement. *PLoS One* 9:e93565. doi: 10.1371/journal.pone.0093565

Butterworth, B. (2010). Foundational numerical capacities and the origins of dyscalculia. *Trends Cogn. Sci.* 14, 534–541. doi: 10.1016/j.tics.2010.09.007

Butterworth, B., Varma, S., and Laurillard, D. (2011). Dyscalculia: from brain to education. *Science* 332, 1049–1053. doi: 10.1126/science.1201536

Cain, K., Oakhill, J., and Bryant, P. (2000). Investigating the causes of reading comprehension failure: the comprehension-age match design. *Read. Writ.* 12, 31–40. doi: 10.1023/A:1008016920825

Chodura, S., Kuhn, J.-T., and Holling, H. (2015). Interventions for children with mathematical difficulties: a meta-analysis. *Z. Psychol.* 223, 129–144. doi: 10.1027/2151-2604/a000211

Decarli, G., Paris, E., Tencati, C., Nardelli, C., Vescovi, M., Surian, L., et al. (2020). Impaired large numerosity estimation and intact subitizing in developmental dyscalculia. *PLoS One* 15:e0244578. doi: 10.1371/journal.pone.0244578

Decarli, G., Sella, F., Lanfranchi, S., Gerotto, G., Gerola, S., Cossu, G., et al. (2023). Severe developmental dyscalculia is characterized by Core deficits in both symbolic and nonsymbolic number sense. *Psychol. Sci.* 34, 8–21. doi: 10.1177/09567976221097947

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition* 44, 1–42. doi: 10.1016/0010-0277(92)90049-N

Dehaene, S. (2011). The number sense: How the mind creates mathematics (Rev. and updated ed.). Oxford University Press. New York.

de Smedt, B., Noël, M. P., Gilmore, C., and Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends Neurosci. Educ.* 2, 48–55. doi: 10.1016/j.tine.2013.06.001

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends Cogn. Sci.* 8, 307–314. doi: 10.1016/j.tics.2004.05.002

Geary, D. C. (2004). Mathematics and learning disabilities. *J. Learn. Disabil.* 37, 4–15. doi: 10.1177/00222194040370010201

Geary, D. C., Bailey, D. H., and Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: the number sets test. *J. Psychoeduc. Assess.* 27, 265–279. doi: 10.1177/0734282908330592

Geary, D. C., Hoard, M. K., Nugent, L., and Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Dev. Neuropsychol.* 33, 277–299. doi: 10.1080/87565640901982361

Haffner, J., Baro, K., Parzer, P., and Resch, F. (2005). Heidelberger Rechentest (HRT 1–4) [Heidelberger Numeracy Test (HRT 1–4)]. Göttingen: Hogrefe.

Halberda, J., Mazzocco, M. M. M., and Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature* 455, 665–668. doi: 10.1038/nature07246

Henik, A., and Tzelgov, J. (1982). Is three greater than five: the relation between physical and semantic size in comparison tasks. *Mem. Cogn.* 10, 389–395. doi: 10.3758/bf03202431

Holloway, I. D., and Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Dev. Sci.* 11, 644–649. doi: 10.1111/j.1467-7687.2008.00712.x

Holloway, I. D., and Ansari, D. (2009). Mapping numerical magnitudes onto symbols: the numerical distance effect and individual differences in children's mathematics achievement. *J. Exp. Child Psychol.* 103, 17–29. doi: 10.1016/j.jecp.2008.04.001

Iuculano, T., Tang, J., Hall, C. W. B., and Butterworth, B. (2008). Core information processing deficits in developmental dyscalculia and low numeracy. *Dev. Sci.* 11, 669–680. doi: 10.1111/j.1467-7687.2008.00716.x

Kaufmann, L., Wood, G., Rubinsten, O., and Henik, A. (2011). Meta-analyses of developmental fMRI studies investigating typical and atypical trajectories of number processing and calculation. *Dev. Neuropsychol.* 36, 763–787. doi: 10.1080/87565641.2010.549884

Koponen, T., Aunola, K., and Nurmi, J.-E. (2019). Verbal counting skill predicts later math performance and difficulties in middle school. *Contemp. Educ. Psychol.* 59:101803. doi: 10.1016/j.cedpsych.2019.101803

Krajewski, K., and Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: findings from a four-year longitudinal study. *Learn. Instr.* 19, 513–526. doi: 10.1016/j.learninstruc.2008.10.002

Kuhn, J.-T., Raddatz, J., Holling, H., and Dobel, C. (2013). Dyskalkulie vs. Rechenschwäche: basisnumerische Verarbeitung in der Grundschule [Dyscalculia vs. Severe Math Difficulties: basic Numerical Capacities in Elementary School]. *Lernen und Lernstörungen* 2, 229–247. doi: 10.1024/2235-0977/a000044

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

Lafay, A., St-Pierre, M. C., and Macoir, J. (2017). The mental number line in dyscalculia: impaired number sense or access from symbolic numbers? *J. Learn. Disabil.* 50, 672–683. doi: 10.1177/0022219416640783

Landerl, K. (2013). Development of numerical processing in children with typical and dyscalculic arithmetic skills – a longitudinal study. *Front. Psychol.* 4:459. doi: 10.3389/fpsyg.2013.00459

Landerl, K. (2019). Neurocognitive Perspective on Numerical Development in International. *Handbook of Mathematical Learning Difficulties*. eds. A. Fritz, V. G. Haase and P. Räsänen, (Cham: Springer), 9–24.

Landerl, K., Bevan, A., and Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: a study of 8-9-year old students. *Cognition* 93, 99–125. doi: 10.1016/j.cognition.2003.11.004

Landerl, K., and Kölle, C. (2009). Typical and atypical development of basic numerical skills in elementary school. *J. Exp. Child Psychol.* 103, 546–565. doi: 10.1016/j.jecp.2008.12.006

Landerl, K., Vogel, S., and Grabner, R. H. (2021). Early neurocognitive development of dyscalculia in Heterogeneous. *Contributions to Numerical Cognition: Learning and Education in Mathematical Cognition*. eds. W. Fias and A. Henik, (London: Academic Press), 359–382.

Li, H., Hua, X., Yang, Y., Huang, B., and Si, J. (2020). How does task switching affect arithmetic strategy use in children with low mathematics achievement? Evidence from computational estimation. *Eur. J. of Psychol. of Educ.* 35, 225–240. doi: 10.1007/s10212-019-00425-9

Mammarella, I. C., Toffalini, E., Caviola, S., Colling, L., and Szűcs, D. (2021). No evidence for a core deficit in developmental dyscalculia or mathematical learning disabilities. *J. Child Psychol. Psychiatry* 62, 704–714. doi: 10.1111/2Fjcpp.13397

Mayringer, H., and Wimmer, H. (2003). Salzburger Lese-Screening für die Klassenstufen 1–4 (SLS 1–4) [Salzburg Reading Screening Test for Grades 1–4]. Bern: Huber.

Mazzocco, M. M., Feigenson, L., and Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Develop.* 82, 1224–1237. doi: 10.1111/j.1467-8624.2011.01608.x

McCaskey, U., Von Aster, M., O'Gorman, R., and Kucian, K. (2020). Persistent differences in brain structure in developmental dyscalculia: a longitudinal morphometry study. *Front. Hum. Neurosci.* 14:272. doi: 10.3389/fnhum.2020.00272

Moore, A. M., and Ashcraft, M. H. (2015). Children's mathematical performance: five cognitive tasks across five grades. *J. Exp. Child Psychol.* 135, 1–24. doi: 10.1016/j.jecp.2015.02.003

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., and Schulte-Körne, G. (2014). Specific learning disorder: prevalence and gender differences. *PLoS One* 9, 1–8. doi: 10.1371/journal.pone.0103537

Moraske, S., Wyschkon, A., Poltz, N., Kohn, J., Kucian, K., von Aster, M., et al. (2019). Indizierte Prävention von Rechenschwächen im Vorschulalter: Effekte bis Klasse 3 [prevention of math learning disability: effects of an intervention stimulating numerical competencies in children at risk]. *Lernen und Lernstörungen* 8, 141–153. doi: 10.1024/2235-0977/a000224

Moyer, R. S., and Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature* 215, 1519–1520. doi: 10.1038/2151519a0

Mussolin, C., Mejias, S., and Noël, M.-P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition* 115, 10–25. doi: 10.1016/j.cognition.2009.10.006

Noël, M.-P., and Rousselle, L. (2011). Developmental changes in the profiles of dyscalculia: an explanation based on a double exact-and-approximate number representation model. *Front. Hum. Neurosci.* 5:165. doi: 10.3389/fnhum.2011.00165

Olsson, L., Östergren, R., and Träff, U. (2016). Developmental dyscalculia: a deficit in the approximate number system or an access deficit? *Cogn. Dev.* 39, 154–167. doi: 10.1016/j.cogdev.2016.04.006

Peters, E., Slovic, P., Västfjäll, D., and Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgm. Decis. Mak.* 3, 619–635. doi: 10.1017/S1930297500001571

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., et al. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition* 116, 33–41. doi: 10.1016/j.cognition.2010.03.012

Poloczek, S., Büttner, G., and Hasselhorn, M. (2012). Relationships between working memory and academic skills: are there differences between children with intellectual disabilities and typically developing children? *J. Cogn. Educ. Psychol.* 11, 20–38. doi: 10.1891/1945-8959.11.1.20

Pustejovsky, J. E., and Tipton, E. (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J. Bus. Econ. Stat.* 36, 672–683. doi: 10.1080/07350015.2016.1247004

Raddatz, J., Kuhn, J.-T., Holling, H., Moll, K., and Dobel, C. (2017). Comorbidity of arithmetic and Reading disorder: basic number processing and calculation in children with learning Imparments. *J. Learn. Disabil.* 50, 298–308. doi: 10.1177/0022219415620899

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeve, R., Reynolds, F., Humberstone, J., and Butterworth, B. (2012). Stability and change in markers of core numerical competencies. *J. Exp. Psychol.* 141, 649–666. doi: 10.1037/a0027520

Rotzer, S., Kucian, K., Martin, E., von Aster, M., Klaver, P., and Loenneker, T. (2008). Optimized voxel-based morphometry in children with developmental dyscalculia. *NeuroImage* 39, 417–422. doi: 10.1016/j.neuroimage.2007.08.045

Rousselle, L., and Noël, M.-P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs non-symbolic number magnitude processing. *Cognition*, 102, 361–395. doi: 10.1016/j.cognition.2006.01.005

Schleifer, P., and Landerl, K. (2011). Subitizing and counting in typical and atypical development. *Dev. Sci.* 102, 361–395. doi: 10.1111/j.1467-7687.2010.00976.x

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., et al. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* 11, 1141–1152. doi: 10.1111/2041-210X.13434

Schwenk, C., Sasanguie, D., Kuhn, J., Kempe, S., Doebler, P., and Holling, H. (2017). Non-symbolic magnitude processing in children with mathematical difficulties: a meta-analysis. *Res. Dev. Disabil.* 64, 152–167. doi: 10.1016/j.ridd.2017.03.003

Siegler, R. S., and Booth, J. L. (2004). Development of numerical estimation in young children. *Child Dev.* 75, 428–444. doi: 10.1111/j.1467-8624.2004.00684.x

Siegler, R., Thompson, C., and Opfer, J. E. (2009). The logarithmic-to-linear shift: one learning sequence, many tasks, many time scales. *MBE* 3, 143–150. doi: 10.1111/j.1751-228X.2009.01064.x

Skagerlund, K., and Träff, U. (2014). Development of magnitude processing in children with developmental dyscalculia: space, time, and number. *Front. Psychol.* 5, 1–15. doi: 10.3389/fpsyg.2014.00675

Szardenings, C., Kuhn, J.-T., Ranger, J., and Holling, H. (2018). A diffusion model analysis of magnitude comparison in children with and without dyscalculia: care of response and ability are related to both mathematical achievement and stimuli. *Front. Psychol.* 8:1615. doi: 10.3389/fpsyg.2017.01615

Torbeyns, J., Verschaffel, L., and Ghesquière, P. (2004). Strategy development in children with mathematical disabilities: insights from the choice/no-choice method and the chronological-age/ability-level-match design. *J. Learn. Disabil.* 37, 119–131. doi: 10.1177/00222194040370020301

Trick, L. M., and Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. *J. Exp. Psychol.* 19, 331–351. doi: 10.1037/0096-1523.19.2.331

Voß, S., Blumenthal, Y., Sikora, S., Mahlau, K., Diehl, K., and Hartke, B. (2014). Rügener Inklusionsmodell (RIM)-Effekte eines Beschulungsansatzes nach dem Response to Intervention-Ansatz auf die Rechen- und Leseleistungen von Grundschulkindern [The "Rügener Inklusionsmodell" (RIM) - Effects of a school concept based on the Response to Intervention approach on the mathematics and reading achievement of German elementary school students]. *Empirische Sonderpädagogik* 6, 114–132. doi: 10.25656/01:9248

von Aster, M. G., and Shalev, R. S. (2007). Number development and developmental dyscalculia. *DMCN* 49, 868–873. doi: 10.1111/j.1469-8749.2007.00868.x

von Wirth, E., Kujath, K., Ostrowski, L., Settegast, E., Rosarius, S., Döpfner, M., et al. (2021). The co-occurrence of attention-deficit/hyperactivity disorder and mathematical difficulties: an investigation of the role of basic numerical skills. *Res. Dev. Disabil.* 112:103881. doi: 10.1016/j.ridd.2021.103881

Wilson, A. J., and Dehaene, S. (2007) Number sense and developmental dyscalculia in *Human Behavior, Learning, and The Developing Brain: Atypical Development*. eds. D. Coch, G. Dawson and K. W. Fischer (New York: The Guilford Press), 212–238.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 - Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest - Revision (WS/ZFR). [Culture Fair Intelligence Test 20-R Scale 2 with vocabulary and numerical order test]*. Göttingen: Hogrefe.

Weiß, R. H., and Osterland, J. (2013). *Grundintelligenztest Skala 1 – Revision (CFT 1-R). [Culture Fair Intelligence Test, Scale 1-R Scale 1]*. Göttingen: Hogrefe.

Wong, T. T.-Y., Ho, C. S.-H., and Tang, J. (2017). Defective number sense or impaired access? Differential impairments in different subgroups of children with mathematics difficulties. *J. Learn. Disabil.* 50, 49–61. doi: 10.1177/0022219415588851

World Health Organization (2019). *International statistical classification of diseases and related health problems*. 11th Edn.

Anhang B: Studie II

Lamb, S., Schulz, A.-K., & Kuhn, J.-T. (2024b). Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule. *Lernen und Lernstörungen, 13*(4), 165–177. https://doi.org/10.1024/2235-0977/a000456

Originalarbeit

Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule

Sarah Lamb, Ann-Katrin Schulz und Jörg-Tobias Kuhn

Methoden der empirischen Bildungsforschung, Fakultät Rehabilitationswissenschaften, Technische Universität Dortmund, Deutschland

Zusammenfassung: *Hintergrund:* Bislang fehlen standardisierte Screeninginstrumente zur Identifikation von Rechenstörungen (RS) im Grundschulalter durch Lehrkräfte, die sich effektiv in die Unterrichtspraxis integrieren lassen und zeitgleich allgemein anerkannte Gütekriterien erfüllen. Mit dem Fragebogen zur Erfassung mathematischer Fertigkeiten (FERMAT) wurde ein theoriebasiertes Screeningverfahren für Lehrkräfte entwickelt, das Kinder mit einem erhöhten Risiko für RS ökonomisch, reliabel und valide identifiziert. *Methoden:* Anhand der Daten von *N* = 377 Schüler_innen aus Nordrhein-Westfalen (Klassenstufe zwei bis vier), wurden mittels psychometrischer Analysen (u.a. Receiver-Operating-Characteristics-Analysen) verschiedene Gütekriterien bestimmt. Zur Prüfung der Treffsicherheit wurden verschiedene Cut-Offs verwendet. *Ergebnisse:* Der FERMAT verfügt über gute bis überwiegend zufriedenstellende Screeningeigenschaften. Die Sensitivität und Spezifität variierten je nach Kriterium zwischen 57.6% und 69.8% bzw. 82.5% und 91.4%. Die RATZ-Indizes lagen zwischen .511 und .613. Die Screeningeigenschaften der Teilbereiche Basisnumerik und Rechenfertigkeiten, in die sich der FERMAT gliedert, sind ebenfalls überwiegend zufriedenstellend. *Diskussion:* Der FERMAT identifiziert Kinder mit und ohne erhöhtes Risiko für RS mit angemessener Treffsicherheit. Die praktische Bedeutung des Instruments wird vor dem Hintergrund langfristiger Folgen unbehandelter RS und den Chancen frühzeitiger Risikoidentifikation diskutiert.

Schlüsselwörter: Rechenstörung, Screening, Früherkennung, Grundschule

Development of a Teacher Questionnaire for the Early Identification of Mathematics Disorders in Primary School

Abstract: *Background:* There is a lack of screening instruments for the early identification of mathematics disorders (MD) in primary school that can be effectively integrated into classroom practices while simultaneously meeting widely accepted quality criteria. The questionnaire to assess mathematical skills of children in grades two to four ("Fragebogen zur Erfassung mathematischer Fertigkeiten von Grundschulkindern, FERMAT") has been developed as a theory-based screening instrument to identify children at risk of developing MD in an effective, reliable, and valid manner. *Methods:* Based on data from *N* = 377 students from North Rhine Westphalia in grades two to four, psychometric quality criteria (e.g., Receiver-Operating-Characteristics-analyses) were investigated. Different cut-off criteria were employed to examine the predictive accuracy of the FERMAT. *Results:* The FERMAT exhibits good to mostly satisfactory screening characteristics. For the criteria used, sensitivity ranged from 57.6% to 69.8%, while specificity ranged from 82.5% to 91.4%. The RATZ index ranged from .511 to .613. FERMAT is divided into two subscales (basic numerical skills and numeracy) whose screening characteristics are also predominantly satisfactory. *Discussion:* Thus, the FERMAT can accurately identify children with and without an increased risk for MD. The practical importance of early identification of MD is discussed regarding long-term consequences of untreated MD and the potential benefits of early risk identification.

Keywords: mathematics disorder, screening, early identification, primary school

Einleitung

Basierend auf dem angeborenen Zahlensinn beginnt die Entwicklung mathematischer Vorläuferfertigkeiten bereits im Säuglingsalter (Dehaene, 1992). Im Vorschulalter verfügen die meisten Kinder bereits über ein Repertoire an Fertigkeiten, das von der Mengenunterscheidung bis hin zur Zahlwortreihe und basaler Arithmetik reicht (Ennemoser, Sinner & Krajewski, 2015). Die Entwicklung setzt sich unter formeller Beschulung meistens erfolgreich fort. Allerdings weisen etwa 25% der Kinder am Ende der Grundschulzeit unterdurchschnittliche mathematische Kompetenzen auf (Selter, Walter, Heinze, Brandt & Jentsch, 2020), und etwa 3 bis 5% entwickeln eine klinisch relevante Rechenstörung (RS) (Fischbach et al., 2013). Die Prävalenz ist damit ähnlich hoch wie bei Beeinträchtigungen im schriftsprach-

lichen Bereich (z. B. Fischbach et al., 2013), das zugehörige Forschungsfeld sowie die Anerkennung in Bildungspraxis und -politik jedoch weniger ausgereift (Ennemoser et al., 2015; Lorenz, 2012). Ohne Intervention besteht die Gefahr, dass sich anfängliche Schwierigkeiten zu persistierenden Lernstörungen entwickeln (Lorenz, 2012), die bis in das Erwachsenenalter bestehen bleiben (Butterworth, Varma & Laurillard, 2011), sekundäre emotionalsoziale Komorbiditäten verursachen (Kohn, Wyschkon, Ballaschk, Ihle & Esser, 2013) und in einer benachteiligten Bildungsbiografie münden (Parsons & Bynner, 2005). Diese nachteiligen Folgen unterstreichen die Wichtigkeit der Früherkennung von Kindern mit einem Risiko für RS.

Im Sinne des Response-to-Intervention-Ansatzes ermöglicht die frühzeitige Risikoidentifikation eine weiterführende Diagnostik und gezielte Förderung (z. B. Grosche & Huber, 2012), noch bevor sich anfängliche Schwierigkeiten in einer negativen Abwärtsspirale manifestieren (Vaughn & Fuchs, 2003). Um den Prozess von der Früherkennung über die Diagnostik bis hin zur Intervention (Tröster, 2009) besser zu verzahnen, ist es wichtig, Screeninginstrumente auf der theoretischen Basis mathematischer Kompetenzentwicklungsmodelle zu konstruieren (vgl. Kuhn & Schwenk, 2018). Screenings, die sich leicht in die Unterrichtspraxis integrieren lassen, sind besonders relevant, da Lehrkräfte als primäre Informationsquelle für den akademischen Entwicklungsverlauf von Kindern ein wichtiger Bestandteil des diagnostischen Prozesses sind (Fischer, Rösch, Nuerk & Moeller, 2015). Doch die derzeit existierenden Screeningverfahren sind vorrangig auf das Kindergarten- und frühe Schulalter ausgerichtet (Landerl, Vogel & Kaufmann, 2022), obwohl bei einigen Kindern Rechenschwierigkeiten erst im Laufe der Grundschulzeit auftreten (Kohn et al., 2013). Zudem basieren die meisten Verfahren auf standardisiert gemessenen Testleistungen (Lorenz, 2012), weshalb die Anwendung teils zeitaufwendig, komplex und damit schwierig in den Lehrkräftealltrag integrierbar ist. Andere Instrumente wiederum wurden nicht auf ihre psychometrischen Eigenschaften hin untersucht, so dass keine Evidenz über die tatsächliche Vorhersagegenauigkeit vorliegt. Zudem werden basisnumerische Fertigkeiten teils vernachlässigt, obwohl diese den Ausgangspunkt der mathematischen Entwicklung bilden (z. B. Fischer, Rosch & Moeller, 2017) und auch am Ende der Grundschule gut zwischen Kindern mit und ohne RS differenzieren (Gaupp, Zoelch & Schumann-Hengsteler 2004). Ziel dieser Arbeit ist es daher, einen theoretisch fundierten Screeningfragebogen für Lehrkräfte zur Erfassung mathematischer Fertigkeiten (FERMAT) vorzustellen. Der ökonomisch anwendbare und in dieser Studie psychometrisch geprüfte Fragebogen soll Lehrkräfte bei der Früherkennung von Grundschulkindern mit einem Risiko für RS unterstützen.

Rechenschwierigkeiten

Rechenschwierigkeiten sind kein einheitlich definiertes Konstrukt. Nicht nur die Begrifflichkeiten (z. B. Rechenschwäche/-störung, Dyskalkulie), sondern auch die diagnostischen Kriterien variieren (DSM-5: American Psychiatric Association [APA], 2013; ICD-11: World Health Organization [WHO], 2020). Klinisch-diagnostische Vertreter_innen sind sich weitgehend einig, dass allgemein von Rechenschwierigkeiten die Rede ist, wenn die mathematische Leistung ohne das Vorliegen einer Intelligenzminderung (IQ > 70) und trotz angemessener Bildung mehr als eine Standardabweichung unter dem Durchschnitt liegt (z. B. Kuhn, 2017). Der davon abzugrenzende Begriff der RS, wird gemäß der S3-Leitlinie zur Diagnostik und Intervention bei Rechenstörung (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften [AWMF], 2018) gebraucht, wenn die mathematische Leistung in einem standardisierten Test trotz ausreichender Bildung und ohne Intelligenzminderung einem Prozentrang (PR) ≤ 7 entspricht. Im internationalen Raum wird häufig ein PR von ≤ 10 verwendet (z. B. Mazzocco, Feigenson & Halberda, 2011). Obwohl das Profil der RS selbst heterogen ist (z. B. Kißler, Schwenk & Kuhn, 2021), sind auf behavioraler Ebene einige Leitsymptome identifizierbar (z. B. Haberstroh & Schulte-Körne, 2019; AWMF, 2018). Diese manifestieren sich in beeinträchtigten basisnumerischen Fertigkeiten und defizitären Rechenfertigkeiten (z. B. Kuhn, 2017). Basisnumerische Fertigkeiten bilden die Grundlage für die weitere Ausbildung des rechnerischen Denkens und Manipulierens von Zahlen im frühen Schulalter (Ennemoser et al., 2015; Fischer et al., 2017) und sind prädiktiv für die mathematische Kompetenzentwicklung (z. B. Krajewski & Schneider, 2009). Die Fertigkeiten werden i. d. R. durch leicht zu lösende Aufgaben erfasst. Ein Beispiel hierfür ist das Größenvergleichsparadigma (z. B. Halberda & Feigenson, 2008), bei dem die Aufgabe darin besteht, aus zwei Mengen (z. B. Punkten oder Zahlen) möglichst schnell die numerisch größere auszuwählen, ohne zu zählen. Ein weiteres Beispiel sind Zahlenstrahlaufgaben, die die Fähigkeit testen, einen numerischen Wert auf einer visuellen Linie zu positionieren. Diesen Aufgabenparadigmen liegt die Annahme zugrunde, dass die Testleistungen Infomationen über angeborene Kernmechanismen der Zahlenverarbeitung generieren, die wiederum Rückschlüsse auf die domänenspezifischen Ursachen von RS ermöglichen (Andersson & Östergren, 2012). Daher sollte die theoretische Konstruktion eines Instruments zur Früherkennung von RS auf Entwicklungsmodellen zur Zahlenverarbeitung basieren und neben den Rechenfertigkeiten auch basisnumerische Fertigkeiten erfassen.

Die Entwicklung der Zahlenverarbeitung

Mehrere Modelle beschreiben die Entwicklung der Zahlenverarbeitung von Kindern (Fischer et al., 2017). Ausgangspunkt der Modelle ist die Annahme, dass die Zahlenverarbeitung auf präverbalen Kernsystemen beruht. Unterschieden werden das *Approximate Number System*, das für die ungefähre Erkennung und Schätzung von Mengen zuständig ist, sowie das *Object Tracking System,* das dem simultanen Erkennen von max. drei bis vier Objekten dient (Feigenson, Dehaene & Spelke, 2004).

Fischer et al. (2017) kontrastierten die bekanntesten deutschsprachigen Entwicklungsmodelle der Zahlbegriffsentwicklung und synthetisierten daraus acht *mathematische Kompetenzbereiche*: Kompetenzbereich 1) beschreibt den Ausgangspunkt der numerischen Entwicklung: die Fähigkeit *Mengen zu schätzen und zu vergleichen*. Darauf aufbauend entwickelt sich Kompetenzbereich 2) *Zahlwortreihe und Zählen*, wodurch ein ordinales Zahlverständnis erworben wird. Kompetenz 3) *Zahlensymbole lesen und schreiben* beschreibt die Fähigkeit arabische Zahlen zu schreiben und zu lesen. Kernkompetenz 4) *Zahlengröße verstehen* beschreibt das Verständnis darüber, dass jede Zahl auch für eine bestimmte Menge steht (kardinales Zahlverständnis). Basierend darauf entwickelt sich die Fähigkeit 5) *Zahlbeziehungen verstehen*, d.h. die Erkenntnis, dass Zahlen zueinander in Beziehung gesetzt werden können (relationale Zahlverständnis). Kompetenz 6) *Zahlenraumvorstellung*, bezieht sich auf die zunehmende Ausdifferenzierung des mentalen Zahlenstrahls, auf dem Zahlen entsprechend ihrer numerischen Größe lokalisiert werden. Die Kompetenzen 7) *Rechnen (Addition/Subtraktion)* und 8) *Stellenwertverständnis* entwickeln sich parallel zu den übrigen Kompetenzen und beschreiben keine separaten Entwicklungsstufen. Lösungsstrategien wie das Zählen beim Addieren und Subtrahieren ermöglichen Rückschlüsse auf die Entwicklungsstufen. Das Stellenwertverständnis beschreibt schließlich das Verständnis dafür, dass sich größere Zahlen aus der Zusammensetzung von Einern, Zehnern usw. ergeben.

Instrumente zur (Früh-)Erkennung von Rechenstörungen in der Grundschule

Es existieren bereits einige deutschsprachige Instrumente zur (Früh-)Erkennung von RS, die im Folgenden exemplarisch im Hinblick auf ihre Vor- und Nachteile dargestellt werden.

Das *Elementar Mathematische Basisinterview* (EMBI; Peter-Koop, Wollring, Spindeler & Grüßing, 2007) dient einer curricular orientierten qualitativen/förderdiagnostischen Einschätzung und ermöglicht eine gezielte Betrachtung der Entwicklung mathematischer Fertigkeiten. Allerdings ist die Durchführung des EMBI aufgrund seines standardisierten Interviewleitfadens mit 20 Seiten als zeitintensiv zu bewerten, was die Integrierbarkeit in den Alltag der Lehrkräfte erschwert.

Ökonomisch verwendbare Checklisten (z.B. Jacobs & Petermann, 2012) und Fragebögen (z.B. Dögnitz, 2022) für Lehrkräfte erscheinen auf den ersten Blick aufgrund ihrer inhaltlichen Konstruktion vielversprechend. Allerdings lässt sich aufgrund fehlender psychometrischer Angaben ihre Eignung hinsichtlich der Testgüte nicht abschließend beurteilen.

Screeninginstrumente wie das *Flensburger-Schulspiel* (FleSch; Clausen-Suhr & Walter, 2022) sind hingegen zwar prognostisch valide, basieren jedoch auf standardisierten Testleistungen und können daher nicht als Fremdbeurteilungsverfahren eingesetzt werden.

Daher war es das Ziel der vorliegenden Studie, ein standardisiertes, theoriebasiertes und ökonomisches Screeninginstrument mit angemessener Treffsicherheit für Lehrkräfte zur Früherkennung von RS zu konstruieren und psychometrisch zu validieren.

Fragebogenkonstruktion

Der FERMAT zielt nicht darauf ab, eine klinische RS (gemäß der ICD-11; WHO, 2020) festzustellen, sondern Kinder zu identifizieren, die einer detaillierteren diagnostischen Untersuchung zugewiesen werden sollten. Angelehnt an Opitz und Bern (2008) werden folgende Anforderungen an den FERMAT gestellt: 1) Der FERMAT soll zentrale mathematische Kompetenzen und häufig beobachtbare Leitsymptome von RS abbilden. 2) Lehrkräfte sollen den FERMAT ohne umfangreiche Schulung selbstständig ausfüllen, auswerten und interpretieren können. 3) Der Fragebogen soll möglichst wenige, aber inhaltlich und testtheoretisch informative Items beinhalten. 4) Das Testergebnis soll leicht interpretiert werden können und klare Handlungsempfehlungen ermöglichen. 5) Der FERMAT soll allgemein anerkannte psychometrische Gütekriterien (z.B. Reliabilität, Objektivität, Validität; American Educational Research Association, American Psychological Association & National Council for Measurement in Education, 2014) erfüllen und entsprechend Kinder mit einem erhöhten RS-Risiko identifizieren.

Der FERMAT (s. Fragebogen im elektronischen Supplement ESM1) setzt sich aus 19 Items zusammen. Je acht Items dienen der Einschätzung *basisnumerischer Fertigkeiten* und *Rechenfertigkeiten*. Drei weitere Items zielen auf eine globale Einschätzung ab. Die theoretische Konstruktion des FERMAT basiert auf den von Fischer et al. (2017)

synthetisierten numerischen Kompetenzen und wird durch häufig beobachtete Defizite bei RS und curricular relevante Themen ergänzt (s. Tab. 1 im ESM2). Nachfolgend werden die im FERMAT verwendeten Items und die damit erfassten Fertigkeitsbereiche dargestellt.

Basisnumerische Fertigkeiten

Das Verständnis der Zahlwortreihe und das damit vorrangig verbundene ordinale Zahlenverständnis wird über das Item *Zählen* (ZÄ) eingeschätzt. Beeinträchtigungen bei RS spiegeln sich häufig in Schwierigkeiten beim Vorwärts- und Rückwärtszählen (30, 31, 32 ...) und/oder im Zählen in Schritten (10, 12, 14 ...) wider (Gaupp et al., 2004). Das Item *Lesen und Schreiben von Zahlensymbolen* (ZLS) bezieht sich auf die Fähigkeit, Zahlensymbole korrekt zu lesen und zu schreiben. Diese Schlüsselkompetenz ist relevant für die Bewältigung visuell dargestellter Rechenaufgaben (Fischer et al., 2017). Bei RS sind häufig Transkodierfehler zu beobachten. Dabei werden verbal dargebotene Zahlen fehlerhaft aufgeschrieben (Kuhn, Raddatz, Holling & Dobel, 2013), bspw. in Form von Inversionsfehlern („vierunddreißig“ statt 43) oder Kompositionsfehlern („50015“ statt 515) (Zuber, Pixner, Moeller & Nuerk, 2009). Das Verständnis von Zahlengrößen und deren Repräsentation als Anzahl wird über das *Anzahlkonzept* (AK) erfasst. Kinder mit mangelndem kardinalen Zahlenverständnis interpretieren Zahlen vorrangig als Zählzahlen und erkennen nicht, dass diese auch Mengen darstellen können (••••• = „fünf“) (Kuhn, 2017). Dies führt u.a. zu Schwierigkeiten bei der Anwendung von Rechenstrategien (Kaufmann & Wessolowski, 2021). Das Verständnis für Zahlensymbole und die Orientierung im Zahlenraum wird über das Item *Zahlen ordnen und vergleichen* (ZOV) abgebildet. Typische Fehler zeigen sich beim Zahlenvergleich (835 > 1358, da 8 größer 1) oder auch darin, dass Vorgänger und Nachfolger einer Zahl falsch benannt werden (welche Zahl kommt vor 43? → 32). Um das häufig bei RS auftretende mangelnde Stellenwertverständnis zu berücksichtigen (Scherer & Moser Opitz, 2010), wurde das Item *Stellenwert* (STW) konstruiert, welches das Verständnis über das Dezimalsystem abdeckt. Um die *Zahlenraumvorstellung* (ZR) zu berücksichtigen, wird die Fähigkeit eingeschätzt, eine numerische Größe auf einem Zahlenstrahl zu positionieren. Kinder mit RS positionieren die Zahlen meist weniger präzise als Kinder ohne RS (Decarli et al., 2023). Die Zahlenraumvorstellung wird separat für den Zahlenraum bis zehn *(ZR-10)*, hundert *(ZR-100)* und tausend *(ZR-1000)* eingeschätzt.

Rechenfertigkeiten

Eingeschätzt werden die Fertigkeiten in den Grundrechenarten: *Addieren (AD)*, *Subtrahieren (SU)*, *Multiplizieren (MU)* und *Dividieren (DI)*, da Schwierigkeiten in diesen Bereichen zu den Leitsymptomen der RS zählen (WHO, 2020). Schwierigkeiten manifestieren sich häufig in Problemen beim Zehnerübergang oder durch einen beeinträchtigten Faktenabruf (Busch, Oranu, Schmidt & Grube, 2013). Dies zeigt sich darin, dass auch bei einfachen Rechenaufgaben wie 8 + 2 nicht auf Faktenwissen zurückgegriffen werden kann, sondern die Aufgabe immer wieder neu berechnet werden muss (Busch et al., 2013). Zur Beurteilung des relationalen Zahlverständnisses wurde das Item *Ergänzungsaufgabe* (*ER*, z.B. ? – 2 = 5 statt der 7 wird eine 3 eingetragen) konstruiert. Schwierigkeiten beim Lösen dieser Aufgabe zeigen sich typischerweise in einem mangelnden Verständnis für die Verwendung mathematischer Operationen bzw. Rechenzeichen. Das Lösen von *Textaufgaben (TA),* das u.a. curricular relevant ist, erfordert i.d.R. ein relationales Zahlenverständnis, da Kinder verstehen müssen, dass Zahlen in Beziehung zueinander gesetzt werden können (Fischer et al., 2017). Um den Inhalt einer Sachaufgabe zu erfassen, wird ein Verständnis für mathematische Operationen benötigt, ohne dieses werden die Zahlen willkürlich verknüpft (Kaufmann & Wessolowski, 2021). *Rechenstrategien (R-ST)* zählen nicht explizit zu den von Fischer et al. (2017) formulierten Kompetenzen, doch Kinder mit RS weisen teils eine verzögerte Strategieentwicklung auf (z.B. Torbeyns, Verschaffel & Ghesquière, 2004) und zeigen wenig bis gar keine Variabilität in ihren Rechenstrategien (z.B. Kuhn, 2017). Vor allem weniger ausgereifte Zählstrategien, z.B. verbales Zählen oder das Zählen mit den Fingern, sind bei RS zu beobachten (z.B. Kuhn 2017), weshalb das *zählende Rechnen (ZÄ-R)* erfasst wird.

Zielsetzung und Fragestellung

Für das Grundschulalter existiert bislang kein ökonomisches, psychometrisch validiertes Instrument zur Früherkennung von RS durch Lehrkräfte. Daher wird ein Fragebogen zur Erfassung mathematischer Fertigkeiten von Grundschulkindern der Klassenstufen zwei bis vier vorgestellt und hinsichtlich psychometrischer Eigenschaften geprüft. Untersucht werden die faktorielle Struktur des Fragebogens (Fragestellung 1[F1]), das psychometrische Messmodell auf Itemebene bzw. die Rasch-Konformität (F2) sowie die Güte der Identifikation von Kindern mit RS-Risiko (F3). Hinsichtlich F1 wird erwartet, dass sich die theoriebasiert entwickelte zweifaktorielle Struktur (basisnumerische Fertigkeiten versus Rechenfertigkeiten) des FERMAT empirisch bestätigt. In Zusammenhang mit F2 wird exploriert, ob die globale Modellpassung des FERMAT für ein ein- oder zweiparametrisches logistisches Testmodell spricht. F3 untersucht vorrangig, inwie-

fern eine zuverlässige Identifikation von Kindern mit einem Risiko für RS anhand der Lehrkräfteeinschätzung im FEMAT gelingt. Der FERMAT gilt als valide, wenn möglichst viele Risikokinder korrekt identifiziert werden, sowohl solche mit tatsächlichem Risiko für RS (richtig-positiv) als auch solche ohne ein Risiko für RS (richtig-negativ). Gleichzeitig sollten die Fehlentscheidungen (falsch-positiv und falsch-negativ) möglichst gering sein. Die Ergebnisse der Treffsicherheit sind von der Operationalisierung des vorherzusagenden Kriteriums (RS-Risiko) abhängig. Da in der Praxis und Forschung unterschiedliche Kriterien angewandt werden, wird die Risikoidentifikation in Anlehnung an gängige internationale (z. B. Mazzocco et al., 2011) und nationale (AWMF, 2018) diagnostische Kriterien exploriert. Kinder zählten zur RS-Risikogruppe, deren Leistung im Heidelberger Rechentest 1–4 (HRT 1–4; Haffner, Baro, Parzer & Resch, 2005) oder im CODY-Mathetest für die 2. bis 4. Klasse (CODY-M 2–4; Kuhn, Schwenk, Raddatz, Dobel & Holling, 2017) einem PR von ≤ 10 oder ≤ 7 entsprach. Während der HRT 1–4 (über die Skala Rechenoperationen) insbesondere die Rechenfertigkeiten in den Blick nimmt, fokussiert der CODY-M 2–4 (Kuhn et al., 2017) die basisnumerischen Fertigkeiten.

Methode

Stichprobe

Die Erhebung erfolgte zu Beginn des Schuljahres an regulären Grundschulen in Nordrhein-Westfalen. Ein positives Ethikvotum sowie die Einverständniserklärungen liegen vor. Die N = 377 Schüler_innen aus n = 33 Schulklassen (weiblich: n = 197 [52.25 %]) verteilen sich auf drei Jahrgangsstufen wie in Tabelle 1 dargestellt.

Design

Über die Dauer von drei Schulstunden bearbeiteten die Kinder Aufgaben aus vier verschiedenen Testverfahren. Die Verfahren wurden in einem ausbalancierten Design vorgegeben, sodass die Reihenfolge zwischen verschiedenen Klassen variierte und jedes Testverfahren an jeder Position vorkam. Während die Testung mittels papierbasierter Instrumente im Klassenverband stattfand, wurde der computerbasierte CODY-M 2–4 Test (Kuhn et al., 2017) in kleineren Gruppen von max. 13 Kindern durchgeführt. Um sicherzustellen, dass die Tests standardisiert durchgeführt werden, wurden die Aufgaben durch geschulte Testleitungen instruiert. Details zu den Instruktionen können den jeweiligen Testmanualen entnommen werden. Die Instruktionen zum unter Aufsicht selbstadministrierten CODY-M 2–4 Test erhielten die Kinder via Kopfhörer, während die Instruktion der übrigen Testverfahren von der Testleitung im Klassenverband vorgelesen wurde.

Die Lehrkräfte schätzten die mathematischen Fertigkeiten ihrer Schüler_innen nach sorgfältiger Durcharbeitung der Instruktion mittels des FERMAT ein. Die Testleitung stand den Lehrkräften im Vorfeld für Rückfragen zur Verfügung. Alle Instruktions- und Ausfüllhinweise (inkl. Beispiele) sind auf dem FERMAT-Fragebogen vermerkt.

Die Rechenfertigkeiten der Kinder wurden mittels des HRT 1–4 erfasst (Haffner, Baro, Parzer & Resch, 2005). Der HRT 1–4 umfasst 11 Untertests, die sich auf zwei Subskalen verteilen: 1) *Rechenoperationen* (6 Untertests, r_{tt} = .93): Addition (z. B. 2 + 4), Subtraktion (z. B. 10–4), Multiplikation (z. B. 2 · 5), Division (z. B. 10 : 5), Ergänzungsaufgaben (z. B. __ + 5 = 12), Größer-Kleiner-Vergleiche (z. B. 4 __ 50) und 2) *numerisch-logische und räumlich-visuelle Fähigkeiten* (5 Untertests, r_{tt} = .87): Zahlenreihen (z. B. 10 8 6 _ _), Längenschätzen, Würfelzählen, Mengenzählen,

Tabelle 1. Demografische Angaben und Testergebnisse in Mathematik, Intelligenz und Leseflüssigkeit gruppiert nach Klassenstufen

Details	Klassenstufe 2	Klassenstufe 3	Klassenstufe 4
n (Jungen)[1,2]	172 (81)	154 (73)	50 (25)
	M (SD)	*M (SD)*	*M (SD)*
Alter in Monaten[2]	98.01_a (6.64)	110.66_b (6.88)	121.31_c (6.62)
HRT 1 – 4	101.81_a (14.31)	99.05_a (15.33)	96.68_a (15.75)
CODY-M-2-4	100.11_a (15.33)	99.74_a (14.74)	100.43_a (14.93)
CFT 1-R/20-R	101.65_a (14.86)	101.09_a (14.06)	90.92_b (15.39)
SLS 1 – 4	99.56_a (15.05)	100.24_a (14.35)	100.78_a (16.95)

Anmerkungen: [1] $\chi^2(2) = 0.13559$, p = .935. [2] Teils fehlende Alters- und Geschlechtsangaben. Alle Werte sind Normwerte. Skalierung für alle Instrumente (M: 100, SD: 15). HRT 1 – 4 (Rechenfertigkeiten); CODY-M-2-4 (vorrangig basisnumerische Fertigkeiten); CFT (Intelligenz), CFT 1-R für die Klassenstufen zwei und drei, CFT 20-R für Klassenstufe vier; SLS 1 – 4 (Leseflüssigkeit). Alle Gruppenvergleiche wurden nach Tukey korrigiert (α = .05). Kleinbuchstaben, die von zwei Gruppen nicht geteilt werden, weisen auf statistisch signifikante Mittelwertunterschiede im Post-hoc-Vergleich hin.

Zahlenverbinden. Jeder Subtest besteht aus mind. 10 und höchstens 40 Aufgaben, die in der Reihenfolge mit zunehmender Schwierigkeit angeordnet sind. Pro Subtest sollten innerhalb von zwei Minuten (Min.) so viele Aufgaben wie möglich korrekt gelöst werden. Als Testergebnis wurde die Gesamtpunktzahl beider Subskalen kombiniert, wobei die Aufgaben Multiplikation, Division, Zahlenverbinden und Würfelzählen nicht berücksichtigt wurden (Haffner et al., 2005).

Die basisnumerischen Fertigkeiten der Kinder wurden mithilfe des CODY-M 2–4 erfasst (Kuhn et al., 2017). Die Subskala *basale Zahlenverarbeitung* (r_{tt} = .72) besteht aus drei Untertests: Punkte Zählen (18 Aufgaben), Mengenvergleich symbolisch und gemischt (je 24 Aufgaben). Die Testwerte basieren auf reaktionszeitbasierten Effizienzmaßen. Die vier Untertests Zahlendiktat (8 Aufgaben), Zahlensteine (max. 140 Aufgaben), Fehlende Zahl (16 Aufgaben) und Zahlenstrahl (23 Aufgaben) lassen sich der Subskala *komplexe Zahlenverarbeitung* (r_{tt} = .76) zuordnen. Die Testergebnisse der ersten drei Aufgaben basieren auf der Anzahl richtig und falsch gelöster Aufgaben, während beim Zahlenstrahl die mittlere Abweichung von der Zielzahl als Testergebnis dient. Die Subskala *Rechnen* (r_{tt} = .85) umfasst vier Untertests: Addition, Subtraktion (je sieben Aufgaben), Multiplikation und Platzhalteraufgaben (je vier Aufgaben). Als Testscore wird jeweils die Anzahl der Richtigantworten verwendet. Die Skala *visuell-räumliches Arbeitsgedächtnis* (r_{tt} = .61) besteht aus dem Test Matrixspanne, der 16 Aufgaben beinhaltet. Als Testscore wird die Anzahl vollständig korrekt reproduzierter Muster verwendet. Als Testergebnis (Gesamttest: r_{tt} = .88) wurden die Ergebnisse der vier Subskalen kombiniert (Kuhn et al., 2017). Die Bearbeitungszeit des CODY-Tests variierte je nach individueller Lösungsgeschwindigkeit zwischen 25 und 35 Min.

Zur Erfassung der allgemeinen Intelligenz der Zweit- und Drittklässler_innen wurde die Kurzform der *Grundintelligenztest Skala 1-Revision* (CFT 1-R; Weiß & Osterland, 2013) eingesetzt, die aus drei Untertests besteht: Reihen fortsetzen, Klassifikationen und Matrizen *(*α = .95). Die Viertklässler_innen wurden anhand von vier Untertests der *Grundintelligenztest Skala 2-Revision* (CFT 20-R; Weiß, 2006) getestet: Reihen fortsetzen, Klassifikationen, Matrizen und topologische Schlussfolgerungen (r_{tt} = .80).

Zur Beurteilung der Leseflüssigkeit wurde das *Salzburger Lese-Screening* (SLS 1–4; Mayringer & Wimmer, 2003) verwendet. Den Kindern wurde eine Liste mit 84 einfachen Sätzen (z. B. „Bananen sind rot“) vorgelegt (r_{tt} = .87-.90). Innerhalb von drei Min. sollten möglichst viele Sätze auf ihre Korrektheit hin beurteilt werden. Das Testergebnis (Lesequotient [LQ]) ist wie der Intelligenzquotient skaliert (*M*: 100, *SD*: 15) und basiert auf der Gesamtzahl richtig bewerteter Sätze.

Zur Früherkennung von Kindern mit einem RS-Risiko wurde der FERMAT, basierend auf den Beobachtungen der Lehrkräfte der letzten drei Unterrichtsmonate, genutzt. Die 19 einzuschätzenden Items gliedern sich in drei Bereiche: 1) *Basisnumerische Fertigkeiten (FERMAT-BF)*, 2) *Rechenfertigkeiten (FERMAT-RF)* (je acht Items, s. Abschnitt Fragebogenkonstruktion) und 3) *Ergänzungsfragen* (drei Items zur globalen Einschätzung). Ein beigefügter Erklärungsbogen illustriert typische Schwierigkeiten, die bei Kindern mit RS beobachtbar sind (s. Fragebogen im ESM1). Der FERMAT kann auf einem DIN-A4-Bogen ausgefüllt werden. Ein Item wird angekreuzt, wenn ein Kind regelmäßig Schwierigkeiten in einer beschriebenen Fertigkeit hat. Themen, die noch nicht im Unterricht behandelt wurden, können ausgelassen werden. Der Gesamtscore ergibt sich aus der Summe der gesetzten Kreuze. Die Ergänzungsfragen fließen nicht in den Testscore ein. Pro Klasse beträgt die Bearbeitungszeit für die Lehrkräfte ca. 10 bis 15 Min.

Ergebnisse

Häufigkeitsverteilungen für Rechenstörungen

Von 377 Kindern erfüllten 45 (11.94 %) im HRT 1–4 das Kriterium eines RS-Risikos (PR ≤ 10). Davon erfüllten 32 (71.11 %) Kinder das strengere Kriterium und erreichten nur einen PR ≤ 7. Ermittelt durch den CODY-M 2–4 erfüllten 40 der 377 Kinder (10.61 %) das Kriterium eines RS-Risikos (PR ≤ 10), wobei davon 31 (77.5 %) nur einen PR ≤ 7 erreichten. Von den Kindern, die im HRT 1 – 4 Minderleistungen (PR ≤ 10) zeigten, wiesen 40 % zusätzlich Schwierigkeiten (PR ≤ 10) im Lesen auf, wobei 13 % gravierende Minderleistungen (PR ≤ 2 entspricht auf der IQ-Metrik etwa einem Wert von 70) zeigten. Etwa 11 % erreichten einen IQ von ≤ 70. Von den Kindern ohne Minderleistungen (HRT 1 – 4, PR > 10), zeigten ca. 8 % erwartungswidrige Leseleistungen (PR ≤ 10), wobei etwa 1 % nur einen PR ≤ 2 erreichte. 1 % der Kinder ohne Beeinträchtigungen im Rechnen erreichten einen IQ-Wert von ≤ 70. Die Kinder wurden in dieser Studie nicht ausgeschlossen, da diese zusätzliche Beeinträchtigung für die reine Identifikation von Kindern mit einem RS-Risiko zunächst irrelevant ist.

Psychometrische Eigenschaften

F1: Die postulierte Struktur des theoriebasiert entwickelten FERMAT wurde mittels einer konfirmatorischen Faktorenanalyse für kategoriale Daten untersucht (z. B. Wirth & Edwards, 2007). Die Items MU, DI und ZR-1000 wurden von

den Analysen ausgeschlossen, da nicht sichergestellt werden konnte, dass alle Kinder in diesen Kompetenzbereichen bereits formal unterrichtet wurden. Der Likelihood-Ratio-Test (LRT) des Vergleichs eines einfaktoriellen mit einem zweifaktoriellen Modell bevorzugt das zweifaktorielle Modell: χ^2 (1) = 6.7114, p < .01. Die Fitindizes sprechen, interpretiert nach den Richtwerten von Hu und Bentler (1999), für eine gute Passung: χ^2 = 76.900, df = 64, p < .129; CFI = .997; TLI = .995; RMSEA = .024). Da die beiden Faktoren (FERMAT-BF und -RF) hoch korreliert sind r = .93 [.87; .98], wurde in den nachfolgenden Analysen neben den Subskalen auch die Gesamtskala berücksichtigt.

Als Maß für die interne Konsistenz wurde McDonald's Omega (McDonald, 1999) herangezogen. Der FERMAT (ω = .91) und die Subskalen (FERMAT-BF, ω = .85; FERMAT-RF, ω = .91) erwiesen sich als hoch reliabel.

F2: Zur Exploration der Rasch-Konformität wurde die globale Modellpassung des Rasch-Modells (one parameter logistic model/1PL-Modell; Rasch, 1960) und eines zweiparametrigen IRT-Modells (2PL-Modell; Birnbaum, 1968), welches zusätzlich zur Itemschwierigkeit eine itemspezifische Trennschärfe schätzt, verglichen. Der LRT ergab eine bessere Passung für das 2PL-Modell, während das informationstheoretische Fit-Maß des BIC für den FERMAT-Gesamtscore und die Subskala FERMAT-BF eher einen besseren Fit für das Rasch-Modell nahelegte (Tab. 2). Da die Schätzungen der Personenparameter verschiedener unidimensionaler IRT-Modelle (1PL/2PL) sehr hoch korrelieren (> 0.9) (von Davier, 2016), wird der FERMAT im Folgenden Rasch-konform ausgewertet. Die Itemschwierigkeiten für den FERMAT-Gesamtscore reichten von 1.993 (SE = 0.219) bis 6.309 (SE = 0.423) (Tab. 2 im ESM3).

F3: Um die Treffsicherheit von Kindern mit einem RS-Risiko mittels des FERMAT zu bestimmen, wurde ein klassifikatorischer Ansatz verwendet. Mit Receiver Operating Characteristic Curve Analysen wurde die Beziehung zwischen Sensitivität (SN) und Spezifität (SP) mit verschiedenen Cut-Off-Werten eruiert. Als Cut-Off-Werte wurden die Werte herangezogen, die basierend auf der max. Trennschärfe identifiziert wurden: die datenbasierten Cut-Off-Werte repräsentieren den Punkt, an dem die max. Differenzierung zwischen Kindern mit und ohne Anzeichen für RS erreicht wird (Tröster, 2009). Abhängig von den Kriterien zur Klassifikation des RS-Risikos (HRT 1–4, CODY-M 2–4, PR ≤ 7 und PR ≤ 10) erreichten der Gesamtscore bzw. die Subskalen des FERMAT ihre max. Trennschärfe bei Cut-Off-Werten von ein bis vier Testpunkten. Basierend darauf sollten Kinder, je nach Kriterium, bereits beim Vorliegen von Schwierigkeiten in einem der im FERMAT erfassten Fertigkeiten einer weiterführenden Diagnostik zugeteilt werden.

Anschließend wurden die Testergebnisse des FERMAT (positiv = Risiko für RS/negativ = kein Risiko für RS) und die Ergebnisse der standardisierten Tests (HRT 1–4 und CODY-M 2–4), d.h. das tatsächliche Vorliegen eines RS-Risikos (vorhanden/nicht vorhanden) in einer Vierfeldertafel zusammengefasst. Die Spalten der Tabelle geben die Häufigkeiten des Auftretens vier möglicher Kombinationen an: die richtigen Vorhersagen (richtig-positiv bzw. richtig-negativ) und die falschen Vorhersagen (falsch-positiv bzw. falsch-negativ).

Für den FERMAT-Gesamtscore sowie für die beiden Subskalen FERMAT-BF und FERMAT-RF wurden folgende Gütekriterien berechnet: SN und SP, die Positive und Negative Korrektheit (PK und NK), der AUC-Wert (Area Under the Curve), der prävalenzunabhängige Youden-Index sowie der RATZ-Index, der den Relativen Anstieg der Trefferquote gegenüber der Zufallstrefferquote beschreibt (Marx,

Tabelle 2. Globale Modellpassung des 1PL und 2PL-Modells

	AIC	SABIC	HQ	BIC	logLik	χ^2	*df*	*p*
FERMAT-Gesamtscore								
1PL	1983.531	1994.164	2005.383	2038.583	-977.766			
2PL	1942.876	1962.623	1983.457	2045.115	-945.438	64.655	12	.001
FERMAT-BF								
1PL	797.316	803.392	809.802	828.774	-390.658			
2PL	787.430	798.063	809.282	842.482	-379.715	21.886	6	.001
FERMAT-RF								
1PL	1347.403	1352.719	1358.328	1374.928	-666.701			
2PL	1309.992	1319.106	1328.722	1357.179	-642.996	47.41	5	.001

Anmerkungen: 1PL = one parameter logistic model/1PL-Modell; 2PL = two parameter logistic model/2PL-Modell; AIC = Akaike Information Criterion; SABIC = Sample-Size Adjusted Bayesian Information Criterion; HQ = Hannan-Quinn Criterion; BIC = Bayesian Information Criterion; logLik = Log-Likelihood-Wert; χ^2 = Wert des Likelihood-Ratio-Tests.

Tabelle 3. Ergebnisse der ROC-Analysen

	HRT 1 – 4 PR ≤ 10	HRT 1 – 4 PR ≤ 7	CODY-M 2 – 4 PR ≤ 10	CODY-M 2 – 4 PR ≤ 7
FERMAT-Gesamtscore				
AUC [KI]	.835 [.765, .905]	.802 [.714, .890]	.795 [.706, .884]	.828 [.733, .924]
Cut-Off (YI)	2	2	4	4
YI	.546	.503	.489	.547
SN	.698	.677	.576	.640
SP	.848	.825	.914	.907
PK	.395	.276	.413	.348
NK	.952	.963	.953	.970
FP	46	55	27	30
FN	13	10	14	9
RATZ-Index	.613	.587	.511	.585
FERMAT-BF				
AUC [KI]	.733 [.655, .812]	.722 [.629, .815]	.726 [.634, .818]	.774 [.670, .877]
Cut-Off (YI)	1	1	1	2
YI	.446	.432	.415	.500
SN	.545	.548	.529	.538
SP	.901	.884	.886	.962
PK	.429	.304	.321	.519
NK	.936	.955	.949	.965
FP	32	39	38	13
FN	20	14	16	12
RATZ-Index	.464	.467	.445	.502
FERMAT-RF				
AUC [KI]	.829 [.758, .900]	.793 [.704, .882]	.803 [.729, .878]	.825 [.747, .903]
Cut-Off (YI)	2	3	1	1
YI	.544	.529	.527	.563
SN	.682	.625	.821	.867
SP	.862	.904	.707	.696
PK	.411	.392	.256	.208
NK	.951	.961	.970	.983
FP	43	31	93	99
FN	14	12	7	4
RATZ-Index	.600	.562	.723	.795

Anmerkungen: AUC = Area Under the Curve; KI = 95%-Konfidenzintervall; YI = Youden-Index; SN = Sensitivität; SP = Spezifität; PK = Positive Korrektheit; NK = Negative Korrektheit; FP = Falsch-Positiv; FN = Falsch-Negativ; RATZ-Index = Relativer Anstieg der Trefferquote gegenüber der Zufallstrefferquote.

1992). Alle Youden-Indizes waren positiv und sind damit als gut zu bewerten (Youden, 1950). Alle AUC-Werte lagen nach Backhaus, Erichson, Plinke und Weiber (2018) im guten bis sehr guten Bereich und bestätigen damit, dass der FERMAT in der Lage ist, zwischen Kindern mit und ohne Anzeichen eines Risikos für RS zu unterscheiden.

Die SN des *FERMAT-Gesamtscores* variierte je nach Kriterium und lag zwischen 57.6 % und 69.8 % (Tab. 3). D.h. durch den FERMAT werden max. ca. 70 % der Kinder mit testdiagnostischen Anzeichen für RS korrekt klassifiziert. Der höchste positive prädiktive Wert wurde für das Kriterium CODY-M 2–4 PR ≤ 10 ermittelt und betrug 0.413, was bedeutet, dass ca. 40 % der im FERMAT als auffällig erkannten Kinder auch auffällige Leistungen im CODY-M 2–4 zeigten. Die ermittelten RATZ-Indizes bewegen sich in einem Bereich von 51.1 % bis 61.3 %. Werte über 60 % zeigen eine deutliche Verbesserung der Vorhersagegenauigkeit im Vergleich zu einer rein zufälligen Zuordnung (Marx, 1992). Hingegen wurden mind. etwa 80 % der Kinder ohne Anzeichen für RS korrekt durch den FERMAT erkannt. Der niedrigste negative prädiktive Wert wurde für das Kriterium HRT 1–4 PR ≤ 10 ermittelt und lag bei 0.952, d.h. ca. 95 % der Kinder, die im FERMAT als unauffällig identifiziert wurden, erzielten auch im standardisierten Leistungstest (HRT 1–4 PR ≤ 10) unauffällige Ergebnisse (Tab. 3).

Für den Teilbereich *FERMAT-BF* lag die SN je nach Kriterium zwischen 52.9 % und 54.8 %. Somit wurden max. etwa 55 % der Kinder korrekt klassifiziert. Die ermittelten RATZ-Indizes lagen im Bereich von 44.5 % bis 50.2 %. Der niedrigste positive prädiktive Wert zeigte, dass mind. ca. 30 % der Kinder, die im Teilbereich FERMAT-BF als rechenwach identifiziert wurden, tatsächlich Minderleistungen im Rechnen zeigten (HRT 1–4 PR ≤ 7). Der höchste positive prädiktive Wert wurde für das Kriterium CODY-M 2–4 PR ≤ 7 ermittelt und betrug 51.9 %. Der niedrigste negative prädiktive Wert zeigt, dass ca. 94 % der Kinder, die als unauffällig identifiziert wurden, auch im standardisierten Leistungstest (HRT 1–4 PR ≤ 10) unauffällige Ergebnisse erzielten (Tab. 3).

Für den Teilbereich *FERMAT-RF* lag die SN je nach Kriterium zwischen 62.5 % und 86.7 %. Damit wurden ca. 63 % bis 87 % der Kinder korrekt identifiziert. Die ermittelten RATZ-Indizes lagen in einem Bereich von 56.2 % bis 79.5 %. Der niedrigste positive prädiktive Wert zeigt, dass etwa ein Fünftel der Kinder, die im Teilbereich FERMAT-RF als auffällig identifiziert wurden, tatsächlich Minderleistungen im Rechnen aufwiesen (CODY-M 2–4 PR ≤ 7). Der höchste positive prädiktive Wert wurde für das Kriterium HRT 1–4 PR ≤ 10 ermittelt und betrug ca. 41 %. Der niedrigste negative prädiktive Wert zeigte, dass ca. 95 % der Kinder, die als unauffällig identifiziert wurden, auch im HRT 1–4 (PR ≤ 10) unauffällige Ergebnisse erzielten (Tab. 3).

Diskussion

Anfängliche Rechenschwierigkeiten sollten früh erkannt werden (AWMF, 2018). Die meisten Screeninginstrumente konzentrieren sich daher auf den Vorschulbereich und das frühe Schulalter (Landerl et al., 2022), allerdings entwickeln einige Kinder erst im Verlauf der Grundschulzeit Schwierigkeiten (z.B. Kohn et al., 2013). Für diese Zielgruppe fehlen Screeninginstrumente mit angemessener Treffsicherheit, die Lehrkräfte zeiteffizient in ihren Unterricht integrieren können. Die vorliegende Arbeit schließt an diese Forschungslücke durch die Entwicklung des FERMAT an und konzentriert sich auf die Überprüfung der psychometrischen Eigenschaften des Instruments. Im Folgenden werden die Befunde entlang der Forschungsfragen diskutiert.

F1: Beeinträchtigungen in der Basisnumerik zählen neben Defiziten in den Rechenfertigkeiten zu den Kernsymptomen von RS (z.B. Busch et al., 2013). Daher berücksichtigt der FERMAT im Vergleich zu anderen Instrumenten zur (Früh-)Erkennung nicht nur Rechenfertigkeiten, sondern auch basisnumerische Fertigkeiten. Diese theoretische zweidimensionale Struktur wurde empirisch bestätigt, wobei eine Differenzierung hinsichtlich der Risikoidentifikation mittels des FERMAT nicht sinnvoll ist, denn unabhängig davon, in welchem Teilbereich die Schwierigkeiten von der Lehrkraft eingeschätzt werden, sollte die weitere diagnostische Abklärung beide Teilbereiche berücksichtigen, um geeignete Interventionen auszuwählen (AWMF, 2018; Kuhn & Schwenk, 2018). Dennoch unterstützt der FERMAT bei der Auswahl geeigneter Tests. Werden im FERMAT z.B. vorrangig Schwierigkeiten in der Basisnumerik festgestellt, empfiehlt es sich, in der weiteren Diagnostik den Schwerpunkt auf diesen Bereich zu legen.

F2: Basierend auf dem LRT wurde die globale Modellpassung des FERMAT besser durch das 2PL- als durch das 1PL-Modell abgebildet. Da die Schätzungen der Personenparameter verschiedener unidimensionaler IRT-Modelle hoch korrelieren (von Davier, 2016) und informationstheoretische Maße (BIC) das überwiegend 1PL-Modell bevorzugen, wurden in der vorliegenden Studie Rasch-konform variierende Schwierigkeitsparameter, nicht aber unterschiedliche Trennschärfen berücksichtigt. Die Rasch-Konformität sollte in Folgestudien klassenstufenspezifisch erneut überprüft werden.

F3: Durch den FERMAT werden Kinder mit einem erhöhten RS-Risiko mit angemessener Treffsicherheit identifiziert. Ca. 70 % der Kinder, die unterdurchschnittliche Leistungen (PR ≤ 10) im HRT 1–4 zeigten, wurden durch den FERMAT (Gesamtscore) identifiziert. Etwa 55 % der Kinder, die Minderleistungen im HRT 1–4 (PR ≤ 7) zeigten, wurden durch die Subskala FERMAT-BF identifiziert. Durch die Subskala FERMAT-RF konnten fast 90 % der Kinder, die im CODY-M 2–4 einen PR ≤ 7 erreichten, rich-

tig erkannt werden. Dass die Treffsicherheit der Subskala FERMAT-RF höher ausfiel, könnte mit der diagnostischen Kompetenz der Lehrkräfte in Zusammenhang stehen. Der Mathematikunterricht in der Grundschule schließt an die frühen mathematischen Alltagserfahrungen der Kinder an, doch der Fokus liegt auf der Vermittlung mathematischer Inhalte. Daher wäre es plausibel, dass die Lehrkräfte die Rechenfertigkeiten akkurater einschätzen als die basisnumerischen Fertigkeiten. Der FERMAT könnte dazu beitragen, das Bewusstsein der Lehrkräfte für Schwierigkeiten in der Basisnumerik zu schärfen, die selbst in der weiterführenden Schule noch substanziell mit der Mathematikleistung assoziiert sind (Ennemoser, Krajewski & Schmidt, 2011). Studien in anderen Bereichen zeigten, dass fragebogengestützte Beurteilungsverfahren zuverlässigere Einschätzungen ermöglichen (Begeny, Eckert, Montarello & Storie 2008) und explizite Informationen über den zu beurteilenden Bereich die Genauigkeit der Beurteilungen verbessern können (Demaray & Elliott, 1998).

Die datenbasierten Cut-Off-Werte des FERMAT reichten von einem bis vier Testpunkten, je nach Betrachtung des Gesamtwerts oder der Subskalen. Um schwerwiegendere falsch-negative Ergebnisse zu vermeiden, empfiehlt sich die Nutzung eines konservativeren Cut-Off-Wertes von zwei. Denn auch wenn ein falsch-positiver Befund negative Auswirkungen wie Stigmatisierung oder schulische Ängste haben kann (Fischbach et al., 2013; Tröster, 2009), sind falsch-negative Ergebnisse im Kontext von Lernschwierigkeiten weitaus gravierender. Setzen sich anfängliche Schwierigkeiten in einer negativen Abwärtsspirale fort und führen zu sekundären Auffälligkeiten, sind bspw. rechenspezifische Interventionen weniger wirksam (Herzog & Casale, 2022). Weist ein Kind in zwei der durch den FERMAT abgedeckten Fertigkeiten Schwierigkeiten auf, ergibt sich eine SN von 69.8 %, d. h. ca. 70 % aller Kinder, die im HRT 1–4 Minderleistungen aufwiesen (PR ≤ 10), ließen sich durch den FERMAT korrekt identifizieren. Durch die frühe Risikoidentifikation können Lehrkräfte mit einzelnen Kindern weitere Tests durchführen (eine Übersicht über geeignete Test- und Förderverfahren bietet bspw. das LONDI-Hilfssystem; Schulte-Körne & Hasselhorn, 2022), um die Erfordernis weiterer Maßnahmen zur Abklärung einer klinisch relevanten RS zu prüfen und um auf dieser Basis adaptive Lehr-Lernsituationen (Fischer et al., 2017) bzw. Fördermaterialien (LONDI-Hilfssystem; Schulte-Körne & Hasselhorn, 2022) bereit zu stellen.

Einige Kinder wiesen unterdurchschnittliche Leistungen in der Leseflüssigkeit und Intelligenz auf. Obgleich diese zusätzliche Beeinträchtigung für die RS-Risikoidentifikation zunächst irrelevant ist, belegen die Befunde erneut, dass einige Kinder multiple schulische Schwierigkeiten aufweisen (z. B. Fischbach et al., 2013). Dies unterstreicht die Relevanz der Berücksichtigung weiterer Ein- und Ausschlusskriterien in einer nachfolgenden Diagnostik. Des Weiteren konnte beobachtet werden, dass etwa 8 % der Kinder zwar keine Schwierigkeiten im Rechnen, jedoch im Lesen aufwiesen, weshalb auch Screening-Instrumente für die Früherkennung von Lese- und Rechtschreibstörungen sinnvoll erscheinen.

Wiederum wurden ca. 85 % der Kinder, die keine Schwierigkeiten (HRT 1–4, PR > 10) aufwiesen, im FERMAT als unauffällig identifiziert und damit ebenfalls richtig erkannt. Die Anzahl der fälschlicherweise als rechenschwach identifizierten Kinder, beläuft sich auf ca. 15 %. Auf Basis qualitativer Daten oder förderdiagnostischer Ansätze könnte zukünftig untersucht werden, ob sich ein systematisches Muster bezüglich der Fehlklassifikation identifizieren lässt. Da das Testergebnis im FERMAT keine Diagnose mit lerntherapeutischer Intervention, sondern lediglich eine weiterführende Diagnostik anregt, kann die Treffsicherheit des FERMAT aus beiden Perspektiven als ausreichend bewertet werden.

Limitationen

Die ermittelten Cut-Off-Werte basieren auf der max. Trennschärfe. Künftige Studien sollten den Einfluss der Cut-Off-Werte im Hinblick auf weitere Gütekriterien (z. B. Aufdeckungsrate) kontrastieren. Obwohl eine geringe positive Korrektheit bei hoher SN und ausreichenden RATZ-Indizes nicht unüblich ist, ist eine Verbesserung der positiven Korrektheit anzustreben. Stichprobenspezifische Effekte, insbesondere bei der positiven und negativen Korrektheit, sind möglich, da sie nicht die Testeigenschaften widerspiegeln, sondern von der Grundquote der Stichprobe abhängen (Tröster, 2009). Zukünftige Studien sollten sicherstellen, dass die Kinder bereits formal in allen vom FERMAT erfassten Kompetenzen unterrichtet wurden, um das Ausschließen von Items in den Analysen zu vermeiden. Die prognostische Validität sollte durch zwei unabhängige Messzeitpunkte ergänzend überprüft werden. Ob andere Kriterien zur Klassifikation von RS die Genauigkeit, mit der der FERMAT Kinder mit RS-Risiko identifiziert beeinflusst und ob die Ergebnisse in Abhängigkeit der Klassenstufen variieren, sollte künftig ebenfalls getestet werden. Da die Daten aus einem Bundesland stammen, könnte die Generalisierbarkeit der Ergebnisse eingeschränkt sein. Aufgrund der freiwilligen Teilnahme könnten selektiert Lehrkräfte an der Befragung teilgenommen haben, die sich in ihrer Einschätzung über die mathematischen Fertigkeiten ihrer Schüler_innen sicher fühlten. Künftig sollten Informationen erhoben werden, die einen Einfluss auf die Lehrkräfteeinschätzung und damit auf die Güte und Bearbeitungszeit des FERMAT haben könnten (z. B. Berufserfahrung, diagnostische Kompetenz etc.).

Relevanz für die Praxis

Der FERMAT kann Kinder mit einem erhöhten Risiko für RS ökonomisch und valide identifizieren. Die Anwendung erfordert keine umfangreiche Schulung. Alle Ausfüllhinweise inkl. Beispiele für die typisch zu beobachtenden Schwierigkeiten bei RS sind auf dem Fragebogen vermerkt. Die Auswertung erfolgt durch eine einfache Summenbildung, dessen Interpretation eindeutig ist: A) Weist ein Kind in weniger als zwei Bereichen Schwierigkeiten auf, liegt kein erhöhtes RS-Risiko vor. B) Weist ein Kind in mind. zwei Bereichen Schwierigkeiten auf, sollten die mathematischen Fertigkeiten des Kindes in einem standardisierten Test überprüft werden. Der ggf. ins Digitale übertragbare FERMAT kann somit dazu beitragen, dass personelle Ressourcen im schulischen Setting optimal genutzt und Kinder mit einem RS-Risiko frühzeitig erkannt und gefördert werden.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1024/2235-0977/a000456.

ESM1. Fragebogen zur Erfassung mathematischer Fertigkeiten (FERMAT).

ESM2. Tabelle E1. Verortung der numerischen Kompetenzbereiche des FERMAT im Kompetenzentwicklungsmodell nach Fischer et al. (2017).

ESM3. Tabelle E2. Itemschwierigkeiten und Standardfehler des FERMAT-Gesamtscores basierend auf dem 1 PL-Modell.

Literatur

American Educational Research Association, American Psychological Association & National Council for Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

American Psychiatric Association (APA). (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American psychiatric association. Verfügbar unter https://doi.org/10.1176/appi.books.9780890425596

Andersson, U. & Östergren, R. (2012). Number magnitude processing and basic cognitive functions in children with mathematical learning disabilities. *Learning and Individual Differences, 22*(6), 701 – 714. https://doi.org/10.1016/j.lindif.2012.05.004

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) (2018). *S3-Leitlinie: Diagnostik und Behandlung der Rechenstörung*. Verfügbar unter: https://www.awmf.org/leitlinien/detail/ll/028-046.html

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2018). *Multivariate Analysemethoden*. Berlin: Springer.

Begeny, J. C., Eckert, T. L., Montarello, S. A. & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judg ments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 43 – 55. https://doi.org/10.1037/1045-3830.23.1.43

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), Statistical Theories of Mental Test Scores (S. 397 – 479). Frankfurt a. M.: Addison-Wesley.

Busch, J., Oranu, N., Schmidt, C. & Grube, D. (2013). Rechenschwäche im Grundschulalter: Reduzierte Verfügbarkeit basalen arithmetischen Faktenwissens und Belastung des Arbeitsgedächtnisses bei Drittklässlern. *Lernen und Lernstörungen, 2*, 217 – 227. https://doi.org/10.1024/2235-0977/a000043

Butterworth, B., Varma, S. & Laurillard, D. (2011). Dyscalculia: From brain to education. *Science, 332*(6033), 1049 – 1053. https://doi.org/10.1126/science.1201536

Clausen-Suhr, K. & Walter, J. (2022). Entwicklung und Evaluation eines Screening-Verfahrens zur Prognose von Rechenschwierigkeiten in der Grundschule. Das Flensburger Schulspiel (FleSch). *Lernen und Lernstörungen, 11*, 125 – 138. https://doi.org/10.1024/2235-0977/a000373

Von Davier, M. (2016). Rash model. In W. J. van der Linden (Hrsg.), *Handbook of Item Response Theory* (Vol 1, S. 31 – 48). Boca Raton, FL, USA: Chapman and Hall/CRC. https://doi.org/10.1201/9781315374512

Decarli, G., Sella, F., Lanfranchi, S., Gerotto, G., Gerola, S., Cossu, G. et al. (2023). Severe Developmental Dyscalculia Is Characterized by Core Deficits in Both Symbolic and Nonsymbolic Number Sense. *Psychological Science, 34*(1), 8 – 21. https://doi.org/10.1177/09567976221097947

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*(1 – 2), 1 – 42. https://doi.org/10.1016/0010-0277(92)90049-N

Demaray, M. K. & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8 – 24. https://doi.org/10.1037/h0088969

Dögnitz, S. (2022). *Diagnostik von besonderen Rechenschwierigkeiten in der Sekundarstufe I*. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-40071-2_9

Ennemoser, M., Krajewski, K. & Schmidt, S. (2011). Entwicklung und Bedeutung von Mengen-Zahlen-Kompetenzen und eines basalen Konventions- und Regelwissens in den Klassen 5 bis 9. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 228 – 242. https://doi.org/10.1026/0049-8637/a000055

Ennemoser, M., Sinner, D. & Krajewski, K. (2015). Kurz- und langfristige Effekte einer entwicklungsorientierten Mathematikförderung bei Erstklässlern mit drohender Rechenschwäche. *Lernen und Lernstörungen, 4*, 43 – 59. https://doi.org/10.1024/2235-0977/a000091

Feigenson, L., Dehaene, S. & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307 – 314. https://doi.org/10.1016/j.tics.2004.05.002

Fischbach, A., Schuchardt, K., Brandenburg, J., Klesczewski, J., Balke-Melcher, C., Schmidt, C. et al. (2013). Prävalenz von Lernschwächen und Lernstörungen: Zur Bedeutung der Diagnosekriterien. *Lernen und Lernstörungen, 2*, 65 – 76. https://doi.org/10.1024/2235-0977/a000035

Fischer, U., Rösch, S. & Moeller, K. (2017). Diagnostik und Förderung bei Rechenschwäche. *Lernen und Lernstörungen, 6*, 25 – 38. https://doi.org/10.1024/2235-0977/a000160

Fischer, U., Rösch, S., Nuerk, H. C. & Moeller, K. (2015). Erkennen von Rechenschwäche durch LehrerInnen und Testungen im Klassenverband. *Lernen und Lernstörungen, 4*, 269 – 282. https://doi.org/10.1024/2235-0977/a000116

Fusion, K. C. (2012). *Children's counting and concepts of number*. Wiesbaden: Springer.

Gaupp, N., Zoelch, C. & Schumann-Hengsteler, R. (2004). Defizite numerischer Basiskompetenzen bei rechenschwachen Kindern der 3. und 4. Klassenstufe. *Zeitschrift für Pädagogische Psychologie, 18*, 31 – 42. https://doi.org/10.1024/1010-0652.18.1.31

Grosche, M. & Huber, C. (2012). Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik, 8*(12), 312 – 322.

Haberstroh, S. & Schulte-Körne, G. (2019). The Diagnosis and Treatment of Dyscalculia. *Deutsches Ärzteblatt international, 116*(7), 107 – 114. https://doi.org/10.3238/arztebl.2019.0107

Haffner, J., Baro, K., Parzer, P. & Resch, F. (2005). *Heidelberger Rechentest (HRT 1 – 4)*. Göttingen: Hogrefe.

Halberda, J. & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology, 44*(5), 1457 – 1465. https://doi.org/10.1037/a0012682

Herzog, M. & Casale, G. (2022). The effects of a computer-based mathematics intervention in primary school students with and without emotional and behavioral difficulties. International Electronic Journal of Elementary Education, 14(3), 303 – 317. https://iejee.com/index.php/IEJEE/article/view/1730

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1 – 55. https://doi.org/10.1080/10705519909540118

Jacobs, C. & Petermann, F. (2012). *Diagnostik von Rechenstörungen*. Göttingen: Hogrefe.

Kaufmann, S. & Wessolowski, S. (2021). *Rechenstörungen: Diagnose und Förderbausteine* (5. Aufl.). Hannover: Klett Kallmeyer.

Kißler, C., Schwenk, C. & Kuhn, J. T. (2021). Two dyscalculia subtypes with similar, low comorbidity profiles: A mixture model analysis. Frontiers in Psychology, 12, 589506. https://doi.org/10.3389/fpsyg.2021.589506

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, 77 – 89. https://doi.org/10.1024/2235-0977/a000032

Krajewski, K. & Schneider, W. (2009). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: Findings from a 3-year-longitudinal study. *Journal of Experimental Child Psychology, 103*(4), 516 – 531. https://doi.org/10.1016/j.jecp.2009.03.009

Kuhn, J.-T. (2017). Rechenschwäche – eine interdisziplinäre Einführung. In A. Fritz, S. Schmidt & G. Ricken (Hrsg.), *Handbuch Rechenschwäche: Lernwege, Schwierigkeiten und Hilfen bei Dyskalkulie* (3. Aufl., S. 14 – 29). Weinheim: Beltz.

Kuhn, J.-T., Raddatz, J., Holling, H. & Dobel, C. (2013). Dyskalkulie vs. Rechenschwäche: Basisnumerische Verarbeitung in der Grundschule. *Lernen und Lernstörungen, 2*, 229 – 247. https://doi.org/10.1024/2235-0977/a000044

Kuhn, J.-T., & Schwenk, C. (2018). Onlinebasierte Diagnostik mathematischer Kompetenzen: Möglichkeiten und Grenzen. Lernen und Lernstörungen, 7, 231 – 235. https://doi.org/10.1024/2235-0977/a000232

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

Landerl, K., Vogel, S. & Kaufmann, L. (2022). *Dyskalkulie: Modelle, Diagnostik, Intervention* (4. überarb. und erw. Auflage). *UTB: Bd. 3066*. München: Ernst Reinhardt. https://doi.org/10.36198/9783838557342

Lorenz, J. H. (2012). *Kinder begreifen Mathematik: Frühe mathematische Bildung und Förderung. Entwicklung und Bildung in der Frühen Kindheit*. Stuttgart: Kohlhammer.

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

Mayringer, H. & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufen 1 – 4 (SLS 1 – 4)*. Bern: Huber.

Mazzocco, M. M., Feigenson, L. & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child development, 82*(4), 1224 – 1237.https://doi.org/10.1111/j.1467-8624.2011.01608.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence.

Opitz, E. M. & Bern, P. H. (2008). *Rechenschwäche erfassen: Screening für die Schuljahre 4 – 8*. Universitätsbibliothek Dortmund. https://eldorado.tu-dortmund.de/bitstream/2003/31691/1/141.pdf

Parsons, S. & Bynner, J. (2005). *Does numeracy matter more?* London: NRDC.

Peter-Koop, A., Wollring, B., Spindeler, B. & Grüßing, M. (2007). *Elementarmathematisches Basisinterview*. Offenburg: Mildenberger.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Scherer, P. & Moser Opitz, E. (2010). *Fördern im Mathematikunterricht der Primarstufe*. Heidelberg: Springer.

Schulte-Körne, G. & Hasselhorn, M. (2022). *LONDI-Hilfssystem. Unterstützung bei Diagnostik und Förderung von Lernstörungen bei Kindern*. Verfügbar unter: https://hilfssytem.londi.de/

Selter, C., Walter, D., Heinze, A., Brandt, J. & Jentsch, A. (2020). Mathematische Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, N., C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 25 – 56). Münster; New York: Waxmann. https://doi.org/10.31244/9783830993193

Torbeyns, J., Verschaffel, L. & Ghesquière, P. (2004). Strategy development in children with mathematical disabilities: Insights from the choice/no-choice method and the chronological-age/ability-level-match design. *Journal of Learning Disabilities, 37*(2), 119 – 131. https://doi.org/10.1177/00222194040370020301

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter. Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

Vaughn, S. & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction. The promise and potential problems. *Learning Disabilities Research & Practice, 18*(3), 137 – 146. https://doi.org/10.1111/1540-5826.00070

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 – Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest – Revision (WS/ZFR)*. Göttingen: Hogrefe.

Weiß, R. H. & Osterland, J. (2013). *Grundintelligenztest Skala 1 – Revision (CFT 1-R)*. Göttingen: Hogrefe.

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58 – 79. https://doi.org/10.1037/1082-989X.12.1.58

World Health Organization (WHO) (2020). International Classification of Diseases for Mortality and Morbidity Statistics (11th revision). Retrieved from https://icd.who.int/browse11/l-m/en

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32 – 35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Zuber, J., Pixner, S., Moeller, K. & Nuerk, H.-C. (2009). On the language specificity of basic number processing: Transcoding in a language with inversion and its relation to working memory ca-

pacity. *Journal of Experimental Child Psychology, 102*(1), 60 – 77. https://doi.org/10.1016/j.jecp.2008.04.003

Historie
Manuskript eingereicht: 05.12.2023
Manuskript angenommen: 04.07.2024
Onlineveröffentlichung: 25.09.2024

Danksagung
Wir danken allen Kindern und Lehrkräften, die an der Studie teilgenommen haben.

Förderung
Diese Arbeit wurde finanziell durch das Bundesministerium für Bildung und Forschung unterstützt (Projektnummer: 01GJ1302). Open-Access-Veröffentlichung ermöglicht durch die Technische Universität Dortmund.

ORCID
Sarah Lamb
https://orcid.org/0000-0003-4619-3269
Jörg-Tobias Kuhn
https://orcid.org/0000-0002-4399-9569

Sarah Lamb
Methoden der empirischen Bildungsforschung
Fakultät Rehabilitationswissenschaften
Technische Universität Dortmund
Emil-Figge-Straße 50
44227 Dortmund
Deutschland
sarah.lamb@tu-dortmund.de

Anhang C: Studie III

Lamb, S., Schulz, A.-K., & Kuhn, J.-T. (2025). Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule. *Empirische Sonderpädagogik, 17*(1), 35–49. https://doi.org/10.25656/01:34364

Lamb, Sarah; Schulz, Ann-Katrin; Kuhn, Jörg-Tobias

Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule

Empirische Sonderpädagogik 17 (2025) 1, S. 35-49

Empirische Sonderpädagogik, 2025.17:35-49
DOI https://doi.org/10.2440/003-0039
ISSN 1869-4845 (Print) · ISSN 1869-4934 (ebook)

Einflussfaktoren der Lehrkräfteeinschätzung mathematikbezogener Schwierigkeiten in der Grundschule

Sarah Lamb, Ann-Katrin Schulz & Jörg-Tobias Kuhn

Technische Universität Dortmund

Zusammenfassung
Lehrkräfte stehen vor der Hausforderung, die Lernvoraussetzungen ihrer Schüler*innen (SuS) zutreffend einzuschätzen. Der Zusammenhang zwischen Lehrkräfteeinschätzungen und SuS-Leistungen in standardisierten Tests ist substanziell, doch die Urteilsgenauigkeit zwischen den Lehrkräften variiert. Zudem können urteilsrelevante sowie -irrelevante SuS-Merkmale das Lehrkräfteurteil beeinflussen. Bei der Einschätzung mathematikbezogener Schwierigkeiten können mathematikspezifische und -unspezifische SuS-Merkmale wie die Lesefertigkeit in das Lehrkräfteurteil einfließen. Zudem gelingt es Lehrkräften weniger gut, SuS im unterdurchschnittlichen Leistungsbereich präzise zu beurteilen, wodurch Förderbedarfe übersehen werden können. Basierend auf den Daten von $N = 377$ Grundschüler*innen und $N = 33$ Lehrkräften, wurde mittels generalisierter linearer gemischter Modelle (GLMM) untersucht, ob die Lehrkräfteeinschätzung über mathematische Schwierigkeiten von den tatsächlichen mathematikspezifischen Leistungen der SuS abhängt, oder ob dabei auch mathematikunspezifische SuS-Merkmale eine Rolle spielen. Zudem wurde geprüft, ob die Urteilsstrenge bzw. -milde zwischen den Lehrkräften variiert und ob das Ausmaß, in dem die SuS-Merkmale herangezogen werden, lehrkräfteabhängig ist. Die Ergebnisse zeigten, dass die tatsächliche mathematische Leistung der SuS prädiktiv für die Lehrkräfteeinschätzung war. Während die Einschätzung basisnumerischer Schwierigkeiten ausschließlich auf mathematikspezifischen Informationen beruhte, bezogen die Lehrkräfte bei der Einschätzung von Schwierigkeiten in den Rechenfertigkeiten auch die Intelligenz und Lesefertigkeit der SuS ein. Das Geschlecht der SuS hatte keinen Einfluss. Urteilsvariationen wiesen auf interindividuelle Unterschiede in der Urteilsstrenge hin, die Gewichtung verschiedener SuS-Merkmale variierte jedoch nicht. Insgesamt zeigen die Ergebnisse, dass die Lehrkräfteeinschätzung über die mathematischen Schwierigkeiten ihrer SuS vorrangig von deren tatsächlichen mathematischen Leistung bestimmt wird. Die Ergebnisse bieten wichtige Erkenntnisse für die Früherkennung von Rechenschwierigkeiten, die für inklusive Schul- und Unterrichtsprozesse relevant sind.

Schlagwörter: Diagnostische Kompetenz, Lehrkräfte, Einflussfaktoren, Rechenschwierigkeiten

Factors influencing teachers' assessments of math-related difficulties in primary school

Summary
Teachers face the challenge of accurately assessing the learning requirements of students. Even if the correlation between teachers' assessments and students' performance in standardized tests is substantial, diagnostic accuracy varies between teachers. Additionally, both relevant and irrelevant student characteristics, such as reading skills can influence the teachers' assessment of mathematical difficulties. Further, teachers are less successful at accurately assessing students in the below-average performance range, therefore learning difficulties can be overlooked. Based on data from *N* = 377 elementary school students and N = 33 teachers, generalized linear mixed models (GLMM) were used to investigate whether teachers' assessments of students' mathematical difficulties depend primarily on their actual mathematics-specific performance or whether -unspecific characteristics also play a role. We also examined whether the strictness or leniency varied between teachers and if the extent to which student characteristics are taken into account is teacher-dependend. The findings indicated that students' actual mathematics performance was predictive of the teachers' assessments. While the teachers' assessments of basic numerical difficulties were based exclusively on mathematics-specific information, teachers incorporated intelligence and reading skills when assessing difficulties in arithmetic performance. Students' gender had no influence. Variations in the teachers' assessments pointed to inter-individual differences in strictness. The weighting of different student characteristics did not vary. Overall, the findings indicate that teachers' assessments of students' mathematical difficulties are primarily determined by their actual mathematical performance. These results provide important insights for the early detection of mathematical difficulties, which are relevant for inclusive school and teaching practices.

Keywords: diagnostic competence, teachers, influencing factors, mathematical difficulties

Die Leistungsbewertung von Schüler*innen (SuS) gehört zum Lehrkräftealltag (Karst et al., 2014). Obwohl Lehrkräfteeinschätzungen generell recht präzise sind (Kaufmann, 2020; Südkamp et al., 2012), gelingt die Beurteilung leistungsschwacher SuS weniger gut (z.B. Lorenz, 2011; Wagner, 2024). Dadurch besteht die Gefahr, dass Minderleistungen übersehen werden und keine angemessenen Interventionen folgen. Gerade in aufeinander aufbauenden Unterrichtsfächern wie der Mathematik, in denen die Kompetenzentwicklung kumulativ verläuft (Salaschek et al., 2014), haben unentdeckte Lernschwierigkeiten massive Folgen für den weiteren Lern- und Entwicklungsverlauf der SuS (Schrader, 2013). Ein Viertel der SuS weist am Ende der Grundschulzeit unterdurchschnittliche mathematische Kompetenzen auf (Selter et al., 2020). Etwa 5 % der SuS entwickeln eine persistierende Lernstörung (Fischbach et al., 2013). Auch wenn Rechenschwierigkeiten (RES) nicht einheitlich definiert sind, lassen sie sich durch charakteristische Schwierigkeiten präzisieren. Dazu zählen Beeinträchtigungen in den basalen (z.B. Mengenvergleiche und Aufgaben im einstelligen Bereich) und komplexen (z.B. Zahlenraumvorstellung im mehrstelligen Bereich) basisnumerischen Fertigkeiten (z.B. Lamb et al., 2024 a), die den Ausgangspunkt für die mathematische Kompetenzentwicklung bilden (Fischer et al., 2017). Weiter zeigen sich Defizite in der

Merkfähigkeit mathematischer Fakten (z.B. Einmaleins) sowie in den Rechenfertigkeiten und -strategien (z.B. Grundrechenarten; APA, 2013). Es ist die Aufgabe der Lehrkräfte, Lernschwierigkeiten frühzeitig zu erkennen (Hesse & Latzko, 2017). Dazu müssen die Leistungen der SuS zutreffend beurteilt werden (Karst, 2012). Bei der Beurteilung mathematischer Schwierigkeiten sollten mathematikspezifische SuS-Merkmale, darunter Rechen- und basisnumerische Fertigkeiten, das Lehrkräfteurteil bestimmen (Fischer et al., 2015). Domänengenerelle SuS-Merkmale, die nicht immanenter Bestandteil der mathematischen Fertigkeiten sind (z.B. Kuhn et al., 2019), sollten nicht substanziell in die Beurteilung einfließen. Doch Lehrkräfte tendieren dazu, die Ausprägung eines SuS-Merkmals (z.B. Intelligenz als zentrales kognitives Merkmal für gute Lernleistungen) als Indiz für ein anderes Merkmal (z.B. Mathematikleistung) heranzuziehen (z.B. Fischer et al., 2015; Kaiser et al., 2015; Mack et al., 2023). Zudem existieren Unterschiede in der Urteilsgenauigkeit und -strenge zwischen den Lehrkräften (z.B. Hesse & Latzko, 2017; Karing & Artelt, 2014; Südkamp et al., 2012). Daher wird in der vorliegenden Studie untersucht, welche mathematikspezifischen und -unspezifischen SuS-Merkmale prädiktiv für die Lehrkräfteeinschätzung von Schwierigkeiten in den mathematischen Fertigkeiten sind und ob sich die Lehrkräfte in ihren Einschätzungen unterscheiden.

Diagnosekompetenz von Lehrkräften

Die Diagnosekompetenz ist elementar für das Handeln von Lehrkräften (Baumert & Kunter, 2006) und eine Schlüsselkompetenz für inklusive Schul- und Unterrichtsprozesse (z.B. Schäfer & Rittmeyer, 2015). Das mehrdimensionale Konstrukt setzt sich aus verschiedenen Kompetenzbereichen und -facetten zusammen (z.B. Brunner et al., 2011). Im Kern handelt es sich dabei um die fachspezifische unterrichtliche Fähigkeit SuS zutreffend zu beurteilen und die Lernanforderungen korrekt einzuschätzen (Karst, 2012). Dieses Verständnis ist gleichbedeutend mit dem Konstrukt der Urteilsgenauigkeit. Für den Bereich der Mathematik setzt dies u.a. Wissen über das mathematische Denken von SuS und Wissen über Leistungsbeurteilungen voraus (z.B. Brunner et al., 2011). Schrader (1989) unterteilt die Diagnosekompetenz von Lehrkräften in drei Komponenten : 1) die Niveaukomponente gibt Auskunft darüber, ob Lehrkräfte dazu neigen, die zu beurteilenden SuS-Merkmale zu über- oder zu unterschätzen. Dazu wird die Differenz zwischen dem Lehrkrafturteil und der SuS-Leistung berechnet. 2) Die Differenzierungskomponente gibt an, ob Lehrkräfte die Leistungsunterschiedlichkeit von SuS in ihrer gesamten Bandbreite wahrnehmen oder unter- bzw. überdifferenzieren. Dazu wird das Verhältnis der Streuungen zwischen den Lehrkräfteeinschätzungen und den SuS-Merkmalen bestimmt. Die 3) Rangordnungskomponente als originärer Kennwert diagnostischer Kompetenz, gibt Aufschluss darüber, wie akkurat Lehrkräfte die Rangordnung der Leistung verschiedener SuS einschätzen können (Schrader & Helmke, 1987). Dazu wird der statistische (klassenspezifische) Zusammenhang zwischen den Lehrkräfteeinschätzungen (z.B. mathematische Fertigkeiten) und den objektiv gemessenen SuS-Merkmalen (z.B. Leistungstest) berechnet. Meta-Analysen, die die Urteilsgenauigkeit anhand von Korrelationen untersuchten, zeigten, dass die Einschätzungen der Lehrkräfte recht präzise sind (Kaufmann, 2020; Südkamp et al., 2012). Dennoch variiert die Urteilsgenauigkeit zwischen Lehrkräften (Lorenz 2011; Südkamp et al. 2012). Grund dafür sind verschiedene Faktoren, die Einfluss auf das Urteil haben können (z.B. Wagner, 2024).

Einflussfaktoren auf die Diagnosekompetenz

In dem heuristischen Modell der Akkuratheit diagnostischer Urteile von Lehrkräften (Südkamp et al., 2012) werden die Einfluss-

faktoren auf vier Bereiche zurückgeführt, die sich gegenseitig beeinflussen aber auch direkt auf die Urteilsgenauigkeit wirken können: Merkmale der Lehrkraft (z.B. Berufserfahrung), der SuS (z.B. Motivation), des Tests (z.B. Länge des Tests) und des Urteils (z.B. Bewertungsskala).

Kaiser et al. (2015) zeigten, dass höhere allgemeine kognitive Fähigkeiten und bessere Deutschleistungen mit positiveren Einschätzungen der Mathematikleistung durch Lehrkräfte einhergingen. Mack et al. (2023) stellten fest, dass die Lehrkräfteeinschätzungen in Mathematik und Deutsch signifikant durch die Leistungen in der jeweils anderen Domäne vorhergesagt werden. Diese Effekte lassen sich durch kognitive Verzerrungen, wie dem logischen Fehlschluss oder Halo-Effekt, erklären (z.B. Helmke, 2017). Dadurch besteht die Gefahr, dass SuS mit RES, die gute Deutschleistungen oder hohe kognitive Fähigkeiten zeigen, übersehen werden. Hinzu kommt, dass Lehrkräfte leistungsstarke SuS tendenziell unter- und leistungsschwache SuS häufiger überschätzen bzw. weniger akkurat beurteilen (z.B. Lorenz, 2011; Wagner, 2024), wodurch sich in der Konsequenz Unterschiede zwischen leistungsstarken und -schwachen SuS verstärken könnten.

Negative mathematikbezogene Stereotype zu Ungunsten von Schülerinnen (z.B. Holder & Kessels, 2017) können zudem dazu führen, dass Schüler von Lehrkräften als mathematisch begabter eingeschätzt werden (z.B. Lorenz, 2011; Mack et al., 2023). Damit beruht das Lehrkräfteurteil auf urteilsrelevanten sowie -irrelevanten Informationen.

Ein möglicher Grund dafür, dass Lehrkräfte leistungsgleiche SuS unterschiedlich einschätzen, ist der Strenge- oder Mildeeffekt. Beim Strengeeffekt bewerten Lehrkräfte Leistungen systematisch strenger, während beim Mildeeffekt Leistungen tendenziell weniger streng bewertet werden (z.B. Helmke, 2017; Hesse & Latzko, 2017).

Daher fokussiert sich diese Studie auf Einflussfaktoren und Lehrkrafteffekte, die prädiktiv für die Lehrkräfteeinschätzung mathematischer Schwierigkeiten von Grundschulkindern sein können.

Fragestellungen

Die Entwicklung und Ausbildung basaler basisnumerischer Fertigkeiten ist vorrangig mit dem Vorschulalter assoziiert (Fischer et al., 2017). Im Grundschulalter rückt die Ausbildung von Rechenfertigkeiten und -strategien in den Fokus. Daher wäre zu erwarten, dass Rechenfertigkeiten (z.B. Addition) sowie komplexere basisnumerische Fertigkeiten (z.B. Transkodieraufgaben) prädiktiv bedeutsamer für das Lehrkräfteurteil sind, während basalere basisnumerische Fertigkeiten (z.B. Mengenvergleiche) aufgrund ihrer geringeren Relevanz im Unterricht auch in geringerem Ausmaß in das Lehrkräfteurteil einfließen. Bei der Einschätzung mathematischer Schwierigkeiten sollten die Rechen- und basisnumerischen Fertigkeiten der SuS das Lehrkräfteurteil bestimmen. Gemäß der Studienlage (z.B. Fischer et al., 2015; Kaiser et al., 2015; Mack et al., 2023) wird jedoch erwartet, dass die Lehrkräfte auch die Lesefertigkeit und Intelligenz ihrer SuS mit in das Urteil einfließen lassen. Möglicherweise neigen Lehrkräfte auch dazu das Geschlecht der SuS einzubeziehen und Schüler als mathematisch begabter einzuschätzen als Schülerinnen (z.B. Holder & Kessels, 2017; Mack et al., 2023). Daher untersucht die erste Forschungsfrage (FF1), *inwieweit Lehrkräfte mathematikspezifische und -unspezifische SuS-Merkmale bei der Einschätzung mathematischer Schwierigkeiten berücksichtigen*. Dazu schätzten die Lehrkräfte die Schwierigkeiten ihrer SuS in der Basisnumerik und den Rechenfertigkeiten mittels des Fragebogens zur Erfassung mathematischer Fertigkeiten (FERMAT; Lamb et al., 2024 b) ein. Der Screeningfragebogen adressiert die Diagnosekompetenz. Die mathematikspezifischen und -unspezifischen SuS-Merkmale wurden über standardisierte Tests erhoben.

Erwartet wird, dass die Urteilsstrenge bzw. -milde zwischen den Lehrkräften variiert (Hesse & Latzko, 2017). Karing und Artelt (2014) zeigten, dass Lehrkräfte die Mathematikleistung ihrer SuS unterschiedlich präzise einschätzten. Sogar innerhalb einer Domäne (z.B. Mathematik) kann die individuelle Urteilsgenauigkeit variieren (Mack et al., 2023; Kolovou et al., 2021). Das Ausmaß, in dem mathematikspezifische und -unspezifische SuS-Merkmale herangezogen werden, könnte somit lehrkräfteabhängig sein. Daher untersucht die zweite Forschungsfrage (FF2), *a) ob die Urteilsstrenge bzw. -milde zwischen den Lehrkräften variiert und b) ob das Ausmaß, in dem mathematikspezifische und -unspezifische SuS-Merkmale herangezogen werden, lehrkräfteabhängig ist.*

Methode

Stichprobe

Erhoben wurden die Daten von *N* = 377 SuS und *N* = 33 Grundschullehrkräften aus 33 regulären Schulklassen in NRW. Die vorliegende Studie knüpft an die Arbeit von Lamb et al. (2024 b) an und bezieht sich auf dieselbe Stichprobe. Die SuS verteilen sich auf drei Jahrgangsstufen (Tab. 1). Die Testung der SuS erfolgte im Klassen- und Einzelsetting. Ein positives Ethikvotum sowie die Einverständniserklärungen liegen vor.

Messinstrumente

Mithilfe des computeradministrierten CODY-Mathetests 2 – 4 (CODY-M 2 – 4; Kuhn et al., 2017) wurden die basalen und komplexen basisnumerischen Fertigkeiten der SuS erhoben. Die Subskala „Basale Zahlenverarbeitung" (r_{tt} =.72) besteht aus drei Untertests: Punkte zählen, symbolischer

Tabelle 1

Demografische Angaben und deskriptive Kennwerte

	Jahrgangsstufe		
	2	**3**	**4**
n (Jungen)[1]	172 (81)	154 (73)	50 (25)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
Alter in Monaten	98.01_a (6.64)	110.66_b (6.88)	121.31_c (6.62)
Basale und komplexe basisnumerische Fertigkeiten			
CODY-M 2 – 4 (BZV)	99.41_a (14.88)	100.15_a (15.53)	101.58_a (13.86)
CODY-M 2 – 4 (KZV)	100.28_a (15.77)	99.35_a (14.12)	101.02_a (15.11)
Rechenfertigkeiten			
HRT 1 – 4	102.63_a (14.37)	98.24_b (14.71)	96.34_b (16.86)
Lesefertigkeit			
SLS 1 – 4	99.56_a (15.05)	100.24_a (14.35)	100.78_a (16.95)
Intelligenz			
CFT 1-R / 20-R	101.65_a (14.86)	101.09_a (14.06)	90.92_b (15.39)

Anmerkungen. Alle Werte sind Normwerte. Skalierung der Erhebungsinstrumente (*M*: 100, *SD*: 15); [1]eine fehlende Geschlechtsangabe, $\chi^2(2) = 0.136$, $p = .935$; basale Zahlenverarbeitung (BZV); komplexe Zahlenverarbeitung (KZV); alle Gruppenvergleiche wurden nach Tukey korrigiert (ausgehend von einem Signifikanzniveau von $\alpha = .05$); Kleinbuchstaben, die von zwei Gruppen nicht geteilt werden, weisen auf statistisch signifikante Mittelwertunterschiede zwischen diesen Gruppen im Post-hoc-Vergleich (Jahrgangsstufe) hin.

und gemischter Mengenvergleich. Die vier Untertests Zahlendiktat, -steine und -strahl sowie Fehlende Zahl bilden die Skala „Komplexe Zahlenverarbeitung“ (r_{tt} =.76).

Mittels des Heidelberger Rechentests (HRT 1 – 4; Haffner et al., 2005) wurden die Rechenfertigkeiten erfasst. Eingesetzt wurden vier Subtests der Skala „Rechenoperationen“ (RO, r_{tt} = .93): Addition (z.B. 6 + 2), Subtraktion (z.B. 8 – 2), Ergänzungsaufgaben (z.B.__+ 7 = 12) und Größer-Kleiner-Vergleiche (z.B. 5__61).

Die Leseflüssigkeit wurde mittels des Salzburger Lese-Screenings (SLS 1 – 4; Mayringer & Wimmer, 2003) erhoben. Dazu wurde den SuS eine Liste mit 84 einfachen Sätzen vorgelegt. Diese mussten so schnell wie möglich gelesen und auf ihre Richtigkeit hin beurteilt werden (r_{tt} = .87–90).

Die Intelligenz der Zweit- und Drittklässler*innen wurde durch die Grundintelligenzskala 1 (CFT 1-R; Weiß & Osterland, 2012) erhoben. Verwendet wurden drei Untertests: Reihen fortsetzen, Klassifikationen und Matrizen (α = .95). Die SuS der vierten Klasse bearbeiteten vier Untertests der Grundintelligenzskala 20-Revision (CFT 20-R; Weiß, 2006): Reihen fortsetzen, Klassifikationen, Matrizen und topologische Schlussfolgerungen (r_{tt} = 0.80).

Die Lehrkräfte schätzten die mathematischen Schwierigkeiten ihrer SuS mittels des FERMAT ein (Lamb et al., 2024 b). Die Subskala „Basisnumerische Fertigkeiten“ (FERMAT-BF, ω = .85) umfasst acht zu bewertende Fertigkeiten: Zählen, das Lesen und Schreiben von Zahlen, Mengenverständnis, Zahlen ordnen und vergleichen, Stellenwertverständnis sowie die Zahlenraumvorstellung bis 10, bis 100 und bis 1000. In der Subskala „Rechenfertigkeiten“ (FERMAT-RF, ω = .91, acht Items) werden Schwierigkeiten in den vier Grundrechenarten, Ergänzungsaufgaben, Textaufgaben, Rechenstrategien und Zählen als Rechenstrategie eingeschätzt. Da nicht sichergestellt werden konnte, dass die Themen Multiplikation, Division und der Zahlenraum bis 1000 bereits in allen Klassen Bestandteil des Unterrichts waren, wurden diese Items von den Analysen ausgeschlossen. Die Subskala FERMAT-BF wurde daher aus sieben und die Skala FERMAT-RF aus sechs Items gebildet. Die Gesamtskala des FERMAT, die sich aus den beiden Teilbereichen zusammensetzt, wurde nicht untersucht, da diese mit der Subskala FERMAT-RF hoch korrelierte, r = .97, p < .001. Die Lehrkraft kreuzte ein Item an (Itemscore = 1), wenn SuS regelmäßig in den letzten drei Unterrichtsmonaten Schwierigkeiten in Bezug auf die beschriebenen Fertigkeiten zeigten. Die Einschätzung durch die Lehrkräfte erfolgte analytisch, nicht holistisch, und wurde durch klare Instruktionen und praktische Beispiele unterstützt (Abb. 1). Höhere Werte im FERMAT (Summe der Itemscores) deuten somit auf größere Schwierigkeiten der SuS in der Basisnumerik (FERMAT-BF: n = 367, M = 0.31, SD = 0.94, Min = 0, Max = 7) und/oder den Rechenfertigkeiten hin (FERMAT-RF: n = 356 M = 0.94, SD = 1.63, Min = 0, Max = 6).

Statistische Analysen

Die statistischen Analysen erfolgten mit R (Version 4.3.1; R Core Team, 2023). Um den Einfluss mathematikspezifischer und -unspezifischer SuS-Merkmale auf die Lehrkräfteeinschätzung von Schwierigkeiten in den mathematischen Fertigkeiten zu untersuchen (FF1) und dabei Lehrkrafteffekte zu berücksichtigen (FF2), wurden Mehrebenenmodelle spezifiziert. Auf der ersten Ebene (Level 1) wurden die mathematikspezifischen (basale und komplexe basisnumerische Fertigkeiten sowie Rechenfertigkeiten) und -unspezifischen (Lesefertigkeit, Intelligenz und Geschlecht) SuS-Merkmale betrachtet. Auf der zweiten Ebene (Level 2) wurden die Lehrkräfte berücksichtigt. Ergänzend wurde geprüft, ob die Jahrgangsstufe prädiktiv für die Einschätzung der Lehrkräfte ist. Dazu wurde diese als fester Level-2-Prädiktor betrachtet. Als Kriterien wurden die beiden Teilbereiche Basisnumerik und Rechenfertigkeiten, in die sich

Abbildung 1.
Adaptierter Auszug aus dem FERMAT (Lamb et al., 2024 b)

Fragebogen zur Erfassung der mathematischen Fertigkeiten (FERMAT)

Dieser Fragebogen bezieht sich auf Ihre Beobachtungen zu den **mathematischen Fertigkeiten Ihrer Schüler*innen**. Bitte machen Sie für jede*n Schüler*in im Fragebogen ein Kreuz, wenn Sie bei diesem*r Schüler*in die im **Erklärungsbogen** genannten **typischen Schwierigkeiten regelmäßig beobachten.** Bitte beziehen Sie sich dabei auf den Zeitraum **der letzten drei Unterrichtsmonate**. Falls etwas noch nicht im Unterricht behandelt wurde (z.B. Zahlenraum bis 1000, Division), streichen Sie bitte die entsprechende Spalte durch.

Wann muss im Fragebogen ein Kreuz gemacht werden?
Der*die Schüler*in zeigt bei der in der Spalte beschriebenen Fertigkeit **mindestens eine der typischen Schwierigkeiten** (s.u.) und/oder **andere** Schwierigkeiten **regelmäßig.**

Fertigkeiten	Auszug aus dem Erklärungsbogen
Basisnumerische Fertigkeiten (BF)	
AK = Anzahlkonzept ☐	Vorstellung von Zahlen als Mengen (kardinaler Zahlenbegriff) **Typische Schwierigkeiten:** Unpräzises Mengenschätzen, Zahlen werden „mengenlos" (nur als „Zählzahl") gedacht
ZOV = Zahlen ordnen und vergleichen ☐	Orientierung im Zahlenraum (Zahlenvergleiche/Zahlenordnen, Vorgänger/Nachfolger) **Typische Schwierigkeiten:** Falsche Zahlenvergleiche/falsches Zahlenordnen (z.B. 975 > 1215, da 9 > 1), Vorgänger/Nachfolger werden falsch benannt (z.B. „welche Zahl kommt vor der 34?" → „23")
Rechenfertigkeiten (RF)	
TA = Textaufgaben ☐	„Vier Hunde liegen im Korb. Zwei kommen noch dazu und drei gehen weg. Wie viele Hunde sind noch im Korb?" **Typische Schwierigkeiten:** Modellieren/Mathematisieren gelingt nicht (Text kann nicht in Rechenaufgabe „übersetzt" werden), wesentliche Informationen zum Lösen der Aufgabe können nicht erkannt werden
R-ST = Rechenstrategien ☐	Verwendung von effizienten Rechenstrategien/operationales Verständnis *Kommutativgesetz:* Der*die Schüler*in weiß, dass 4 + 9 = 13 ist, also muss 9 + 4 auch = 13 sein *n + 1 / n – 1 Prinzip:* Der*die Schüler*in weiß, dass 23 + 44 = 67 ist, also muss 23 + 45 = 68 sein (analog bei Subtraktion) *n · 10 / n + 10-Prinzip:* Der*die Schüler*in weiß, dass 26 + 72 = 98 ist, dann ist auch 260 + 720 = 980 (oder: 5 + 3 = 8 → 15 + 3 = 18) *Inversions-/Umkehrprinzip:* Wenn 46 + 27 = 73, dann ist 73 – 27 = 46 (auch: 11 – 3 = 8 → 11 – 8 = 3) **Typische Schwierigkeiten:** Der*die Schüler*in verwendet keine/extrem selten Rechenstrategien

der FERMAT gliedert, untersucht. Die kontinuierlichen Prädiktorvariablen auf Level-1 wurden basierend auf den klassenspezifischen Normwerten für die Gesamtstichprobe z-standardisiert. Bei den Kriterien handelt es sich um Rohwerte, die nicht z-standardisiert wurden.

Voranalysen. Erwartungsgemäß wählten die Lehrkräfte mit Abstand am häufigsten die Antwortkategorie 0 (= keine Schwierigkeiten) aus. Daraus resultierte eine extrem schiefe Verteilung der FERMAT-Testscores: FERMAT-BF: $g_1 = 4.05$, $z = 15.27$, $p < .001$; FERMAT-RF: $g_1 = 1.89$, $z = 10.25$, $p < .001$ (D'Agostino, 1970). Daher wurde auf eine lineare Modellierung verzichtet und stattdessen geprüft, ob die Datenstruktur besser durch eine Poisson- oder negative Binomialverteilung repräsentiert wird und ob die ‚zero inflation' statistisch zu berücksichtigen ist. Verwendet wurde das R-Paket „glmTMB" (Brooks et al., 2017). Die generalisiert linearen gemischten Modelle (GLMM; z.B. Fahrmeir et al., 2007) mit den vier verschiedenen Verteilungsannahmen (Poisson-Modell und negatives Binomialmodell jeweils mit und ohne zero inflation) wurden mit einem Likelihood-Ratio-Test und anhand von Fit-Indizes verglichen (Dobson, 2002). Die Ergebnisse zeigten zusammengefasst, dass die Modellspezifikation mit dem negativen Binomialmodell ohne Berücksichtigung der ‚zero inflation' am besten zu den Daten passte.

Ergebnisse

Modellvergleiche. FF1 untersucht, inwieweit Lehrkräfte mathematikspezifische und -unspezifische SuS-Merkmale bei der Einschätzung von Schwierigkeiten in den mathematischen Fertigkeiten berücksichtigen. Dazu wurden zunächst Modelle mit konstantem Achsenabschnitt und konstanter Steigung mit mathematikspezifischen und -unspezifischen Level-1-Prädiktoren verglichen. Für das Kriterium FERMAT-RF verbesserte sich der Modellfit durch die Berücksichtigung mathematikunspezifischer SuS-Merkmale signifikant (Tab. 2). Für das Kriterium FERMAT-BF galt dies nicht.

In FF2a wurde getestet, ob die Urteilsstrenge bzw. -milde zwischen den Lehrkräften variiert. Dazu wurde ein klassen- bzw. lehrkraftbezogener Zufallseffekt bzgl. des Achsenabschnitts (random intercept) in die Zufallsstruktur des Modells aufgenommen. Der Modellvergleich für das Kriterium FERMAT-RF zeigte, dass sich die Modellanpassung erneut verbesserte (Tab. 2). Für das Kriterium FERMAT-BF verbesserte sich der Modellfit nicht. Ergänzend wurde die Jahrgangsstufe als fester Level-2-Prädiktor berücksichtigt, wodurch sich der Modellfit für beide Kriterien signifikant verbesserte (Tab. 2).

In FF2b wurde geprüft, ob das Ausmaß, in dem mathematikspezifische und -unspezifische SuS-Merkmale herangezogen werden, lehrkräfteabhängig ist. Aufgrund von Konvergenzproblemen mussten dazu alle kontinuierlichen Prädiktorvariablen nacheinander als zufällige Level-2-Prädiktoren modelliert werden. Infolgedessen wurden für das Kriterium FERMAT-RF fünf Modelle verglichen. Keiner der Modellvergleiche zeigte eine Verbesserung des Modellfits durch die Berücksichtig eines zufälligen Level-2-Prädiktors. Daher wurde das random-intercept-constant-slope Modell mit mathematikspezifischen und -unspezifischen Prädiktoren als finales Analysemodell für das Kriterium FERMAT-RF gewählt (Tab. 3). Für das Kriterium FERMAT-BF wurde das constant-intercept-constant-slope Modell mit mathematikspezifischen Prädiktoren herangezogen (Tab. 3), da dieses am besten zu den Daten passte, d.h. Lesefertigkeiten und Intelligenz zeigten keinen systematischen Zusammenhang mit dem FERMAT-BF (Tab. 2).

FF1: Für die Lehrkräfteeinschätzung im Teilbereich FERMAT-BF erwiesen sich die basalen und komplexen basisnumerischen Fertigkeiten sowie die Rechenfertigkeiten als prädiktiv. Die Intelligenz und Lesefertigkeit sowie das Geschlecht flossen nicht in das Urteil ein (Tab. 3).

Tabelle 2

Modellvergleiche mit festen und zufälligen Effekten

Modellspezifikation		Modellfit		Likelihood-Ratio-Test	
	AIC	BIC	LL	*df*	χ^2
FERMAT-BF					
cics MS	410.10	429.59	-200.05		
cics MS + MU [a]	410.32	441.49	-197.16	3	5.79
rics MS [a]	410.19	433.58	-199.10	1	1.91
cics MS + Jgst [a]	398.19	425.47	-192.10	2	15.91***
FERMAT-RF					
cics MS	819.60	838.93	-404.80		
cics MS + MU	812.39	843.32	-398.19	3	13.21**
rics MS + MU	802.92	837.72	-392.46	1	11.47***
rics MS + MU + Jgst	800.29	842.82	-389.15	2	6.63*

Anmerkungen. ***p < .001; **p < .01; *p < .05; constant-intercept-constant-slope (cics); random-intercept-constant-slope (rics); mathematikspezifische Prädiktoren (MS) auf Level-1 (SuS): basale und komplexe basisnumerische Fertigkeiten, Rechenfertigkeiten; mathematikunspezifische Prädiktoren (MU) auf Level-1 (SuS): Lesefertigkeit, Intelligenz, Geschlecht und Jahrgangsstufe (Jgst) als fester Level-2-Prädiktor; klassen- bzw. lehrkraftbezogener Zufallseffekt bzgl. des Achsenabschnitts (random intercept); [a] mit cics MS Modell verglichen.

Für die Lehrkräfteeinschätzung im Teilbereich FERMAT-RF waren die komplexen basisnumerischen Fertigkeiten, die Rechenfertigkeiten sowie die Intelligenz und Lesefertigkeiten der SuS prädiktiv. Das Geschlecht der SuS hatte keinen Effekt (Tab. 3).

Die negativen Koeffizienten der festen Effekte (*B*) zeigen, dass niedrigere objektive Testleistungen der SuS mit höheren FERMAT-Scores einhergingen (Tab. 3). Kinder die nach Einschätzung der Lehrkräfte mathematische Schwierigkeiten hatten, erreichten auch objektiv tendenziell niedrigere Testwerte. Dies galt für die mathematikspezifischen- und unspezifischen SuS-Merkmale.

Ergänzend wurde geprüft, ob die Jahrgangsstufe einen Effekt hat. Nach Einschätzung der Lehrkräfte hatten Kinder der dritten und vierten Jahrgangsstufe weniger Schwierigkeiten als Zweitklässler*innen. Ein tendenziell ähnlicher Effekt war auch zwischen den SuS der dritten und vierten Jahrgangsstufe zu beobachten (Tab. 3).

FF2a: Die geschätzte random intercept Varianz ($\tau 00_{\text{Lehrkräfte}}$) zeigte, dass sich die Lehrkräfte in der Urteilsstrenge bzw. -milde unterschieden. Für identisch leistungsfähige SuS wurden unterschiedliche FERMAT-Scores (FERMAT-RF) vergeben (Tab. 3). Für den Teilbereich FERMAT-BF gab es keinen Hinweis darauf, dass sich die Lehrkräfte in ihrer Urteilsstrenge bzw. -milde unterscheiden (Tab. 2).

FF2b: Keiner der Modellvergleiche zeigte eine Verbesserung des Modellfits durch die Berücksichtigung variierender Steigungen auf der Lehrkraftebene. Es gab keinen Hinweis darauf, dass die kontinuierlichen mathematikspezifischen- und unspezifischen Prädiktoren in ihrer Bedeutung für den FERMAT-Score zufällig zwischen den Lehrkräften variierten.

Tabelle 3
Ergebnisse der Analysemodelle

	FERMAT-BF			FERMAT-RF		
	B	*SE(B)*	*z*	*B*	*SE(B)*	*z*
Feste Effekte						
Intercept	-1.52	0.20	-7.63***	-0.54	0.20	-2.65**
Basale basisnumerische Fertigkeiten	-0.26	0.13	-2.03*	0.03	0.09	0.32
Komplexe basisnumerische Fertigkeiten	-0.52	0.16	-3.30***	-0.31	0.11	-2.83**
Rechenfertigkeiten	-0.91	0.18	-4.91***	-0.57	0.12	-4.85***
Lesefertigkeit	-	-	-	-0.23	0.11	-2.21*
Intelligenz	-	-	-	-0.35	0.10	-3.32***
Geschlecht[a]	-	-	-	0.27	0.18	1.48
Jgst 3[b]	-1.21	0.30	-4.03***	-0.46	0.25	-1.85
Jgst 4[b]	-0.76	0.39	-1.97*	-0.92	0.37	-2.47*
Zufällige Effekte						
$\tau 00_{\text{Lehrkräfte}}$	-			0.18		
$N_{\text{Lehrkräfte}}$	32			31		
N_{SuS}	364			353		

Anmerkungen. ***p < .001, **p < .01, *p < .05. Mathematikspezifische Prädiktoren auf Level-1 (SuS): basale basisnumerische und komplexe basisnumerische Fertigkeiten, Rechenfertigkeiten; mathematikunspezifische Prädiktoren auf Level-1 (SuS): Lesefertigkeit, Intelligenz, Geschlecht, [a] das männliche Geschlecht als Referenzkategorie; Jahrgangsstufe (Jgst) als fester Level-2-Prädiktor, [b] Jahrgangsstufe zwei als Referenzkategorie; zufällige Effekte: Lehrkräfte als zufällige Achsenabschnitte, $\tau 00_{\text{Lehrkräfte}}$ (geschätzte Varianz des random intercept). Alle kontinuierlichen Prädiktorvariablen wurden z-standardisiert. Für das Kriterium FERMAT-BF erwies sich das Modell mit den mathematikspezfischen Prädiktoren ohne random intercept Varianz als am besten passend.

Diskussion

Untersucht wurde, ob die Einschätzung von mathematischen Schwierigkeiten durch Lehrkräfte von der tatsächlichen mathematikspezifischen Leistung der SuS abhängt oder ob auch mathematikunspezifische SuS-Merkmale eine Rolle spielen. Außerdem wurde geprüft, ob die Urteilsstrenge bzw. -milde zwischen den Lehrkräften variiert und ob das Ausmaß, in dem SuS-Merkmale herangezogen werden, lehrkraftabhängig ist. Die Lehrkräfteeinschätzungen wurden vorrangig von der tatsächlichen mathematischen Leistung, aber auch durch mathematikunspezifische SuS-Merkmale beeinflusst (FF1). Unterschiede zwischen den Lehrkräften deuten auf Variationen in der Urteilsstrenge hin, da objektiv gleiche SuS-Leistungen unterschiedlich bewertet wurden (FF2a), die Gewichtung verschiedener SuS-Merkmale variierte jedoch nicht (FF2b).

Der Einfluss mathematikspezifischer und -unspezifischer SuS-Merkmale auf die Lehrkräfteeinschätzung (FF1)

Die Lehrkräfteeinschätzungen über die Schwierigkeiten ihrer SuS in den basisnumerischen Fertigkeiten wurden ausschließlich durch mathematikspezifische SuS-Merkmale bestimmt, wobei die Testleistung in den Rechenfertigkeiten, wie erwartet, stärker in das Urteil einfloss als die basale und komplexe basisnumerische Leistung. Unspezifische SuS-Merkmale beeinflussten das Urteil nicht, was auf eine unverzerrte Einschätzung hindeutet. Dieses Ergebnis ist für die Früherkennung von RES praktisch bedeutsam, da Defizite in der Basisnumerik in engem Zusammenhang mit Beeinträchtigungen in höheren Rechenfertigkeiten stehen (Braeuning et al., 2021; Lyons et al., 2014) und charakteristisch für RES sind (Landerl et al., 2022). Ob das standardisierte Bewertungssystem das der FERMAT bietet, die Einschätzung präzisierte, sollte in Folgestudien überprüft werden, etwa durch einen Vergleich der Lehrkräfteeinschätzung mit und ohne FERMAT.

Erwartungsgemäß flossen die Testergebnisse der SuS in den Rechenfertigkeiten und die komplexen basisnumerischen Fertigkeiten in die Lehrkräfteeinschätzung über die Schwierigkeiten ihrer SuS in den Rechenfertigkeiten ein. Auch die Lesefertigkeit und Intelligenz der SuS beeinflussten das Lehrkräfteurteil. Die Leistung in der basalen Basisnumerik wurde nicht berücksichtigt. Auf den ersten Blick deutet dieses Ergebnis auf einen logischen Fehlschluss hin (z.B. Mack et al., 2023). Da die Lesefertigkeit und Intelligenz der SuS keinen Einfluss auf die Lehrkräfteeinschätzung von Schwierigkeiten in der Basisnumerik hatte, widerspricht dieses Ergebnis einer Urteilsverzerrung. Plausibel wäre, dass die Lehrkräfte, die im FERMAT zu beurteilenden basisnumerischen Fertigkeiten (z.B. Zahlen vergleichen) im Vergleich zu den Rechenfertigkeiten (z.B. Textaufgaben) als sprachunabhängiger eingeschätzt haben. So könnten sie, wie teils empirisch gezeigt (z.B. Cowan & Powell, 2014), davon ausgegangen sein, dass die Intelligenz für die Beherrschung von Rechenfertigkeiten bedeutsamer ist, als für das Lösen basalerer Aufgaben. Im Umkehrschluss könnte der Einfluss der Lesefertigkeit und Intelligenz auf das Lehrkräfteurteil über die Schwierigkeiten in den Rechenfertigkeiten durch einen Zusammenhang zwischen diesen Fertigkeiten begründet sein (z.B. Akin, 2022; Peng et al., 2019). Auch, wenn domänengenerelle Fertigkeiten förderlich oder hemmend auf die mathematische Kompetenzentwicklung wirken (z.B. von Aster & Shalev, 2007) und mit der mathematischen Leistung korreliert sind, sind mathematikspezifische Fertigkeiten für die Einschätzung von Schwierigkeiten in den Rechenfertigkeiten insgesamt bedeutsamer (z. B. Chen & Li, 2014; Duncan et al., 2007; Kuhn et al., 2019). Daher sollten domänenspezifische Merkmale bzw. mathematikspezifische Fertigkeiten die Lehrkräfteeinschätzung von Schwierigkeiten in den Rechenfertigkeiten bestimmen.

Im Gegensatz zu einigen früheren Studien (z.B. Mack et al., 2023), hatte das Geschlecht der SuS in dieser Studie keinen Einfluss auf die Lehrkräfteeinschätzung, konform zur Übersichtsarbeit von Urhahne und Wijnia (2021). Sollte dieses Ergebnis dauerhaft repliziert werden, könnten Lehrkräfte durch ihren Einfluss auf geschlechtsspezifische Erwartungen langfristig zur Reduktion von Geschlechterunterschieden (z.B. Schütky, 2022) beitragen.

Nach Einschätzung der Lehrkräfte hatten die Zweitklässler*innen mehr basisnumerische Schwierigkeiten, obwohl ihre objektive Testleistung nicht substanziell schlechter war (Tab. 1), was durch eine Depriorisierung basisnumerischer Inhalte in der dritten und vierten Jahrgangsstufe erklärt werden könnte. Diese Interpretation muss in zukünftigen Studien überprüft werden. Da RES aus Defiziten in den Kernsystemen der Zahlenverarbeitung, die für die Verarbeitung der basisnumerischen Inhalte zuständig sind, resultieren können (z.B. Lamb et al., 2024 a), sollten basisnumerische Schwierigkeiten bei der Identifikation von RES auch in höheren Jahrgangsstufen berücksichtigt werden.

Unterschiede zwischen Lehrkräften (FF2)

Konform zum Strenge- bzw. Mildeeffekt (Hesse & Latzko, 2017) erhielten identisch leistungsfähige SuS unterschiedliche FERMAT-Scores. Bei gleicher objektiver Testleistung in den mathematikspezifischen und -unspezifischen Fertigkeiten gaben einige Lehrkräfte Schwierigkeiten in den Rechenfertigkeiten an, während andere Lehrkräfte keine Schwierigkeiten sahen. Dies deutet darauf hin, dass die gleiche Leistung von verschiedenen Lehrkräften unterschiedlich bewertet wird. Es gab jedoch keinen Hinweis darauf, dass das Ausmaß in dem mathematikspezifische und -unspezifische SuS-Merkmale herangezogen werden, lehrkräfteabhängig ist. Für den einzuschätzenden Bereich der Basisnumerik wurden keine Lehrkrafteffekte festgestellt.

Limitationen

Die Ergebnisse sind von den eingesetzten Testverfahren abhängig. Daher sollten die Analysen mit anderen standardisierten Tests, die u.a. Textaufgaben beinhalten, repliziert werden, um alle im FERMAT zu beurteilenden Fertigkeiten abzudecken. Auch fehlende Informationen zu den Lehrkräften limitieren die Ergebnisse, da eine Vielzahl weiterer Einflussfaktoren nicht kontrolliert wurde. Zukünftige Forschungen sollten weitere Merkmale der SuS (z. B. Verhalten im Unterricht) und Lehrkräfte (z. B. stereotype Vorstellungen in Bezug auf den Migrationshintergrund) untersuchen, die die Lehrkräfteeinschätzung beeinflussen können (z.B. Urhahne & Wijnia, 2021; Wagner, 2024). Das gleiche gilt für Effekte auf SuS- und Klassenebene.

Fazit

Lehrkräfte entscheiden darüber, ob SuS gezielte Förderungen erhalten oder nicht (Fischer et al., 2015). Daher sollten Lehrkräfte die mathematischen Schwierigkeiten ihrer SuS basierend auf deren tatsächlichen mathematischen Leistung einschätzen. Urteilsirrelevante Informationen sollten die Einschätzung nicht beeinflussen. Die Studie zeigte, dass die Testleistung in den basalen und komplexen basisnumerischen Fertigkeiten ebenso wie die Leistung in den Rechenfertigkeiten, die bei Kindern mit RES beeinträchtigt sind (z.B. Lamb et al., 2024 a; Landerl, 2022), die Lehrkräfteeinschätzung über *Schwierigkeiten in der Basisnumerik* bestimmten. Das Urteil wurde nicht durch mathematikunspezifische SuS-Merkmale verzerrt. Bei der Einschätzung von *Schwierigkeiten in den Rechenfertigkeiten* identifizierten einige Lehrkräfte bei gleichen objektiven Testleistungen Schwierigkeiten, andere jedoch nicht. Neben den tatsächlichen komplexen basisnumerischen Fertigkeiten und der Leistung in den Rechenfertigkeiten beeinflussten auch die Intelligenz und Lesefertigkeit der SuS die Lehrkräfteein-

schätzung über die *Schwierigkeiten in den Rechenfertigkeiten*. Geschlechterbezogene Urteilsfehler wurden nicht gefunden. Insgesamt sprechen die Ergebnisse dafür, dass die Lehrkräfteeinschätzung über die mathematischen Schwierigkeiten ihrer SuS vorrangig von deren tatsächlicher mathematischer Leistung bestimmt wird. Die Ergebnisse bieten positive Erkenntnisse zur Früherkennung von RES, die für inklusive Schul- und Unterrichtsprozesse relevant sind.

Literatur

Akin, A. (2022). Is reading comprehension associated with mathematics skills: A meta-analysis research. *International Online Journal of Primary Education, 11*(1), 47-61. https://doi.org/10.55020/iojpe.1052559

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders* (5. Aufl.). American Psychiatric Association.

Von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental medicine & child neurology*, *49*(11), 868–873. https://doi.org/10.1111/j.1469-8749.2007.00868.x

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2

Braeuning, D., Hornung, C., Hoffmann, D., Lambert, K., Ugen, S., Fischbach, A., ... & Moeller, K. (2021). Long-term relevance and interrelation of symbolic and non-symbolic abilities in mathematical-numerical development: Evidence from large-scale assessment data. *Cognitive Development, 58*, 101008. https://doi.org/10.1016/j.cogdev.2021.101008

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Waxmann.

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: a meta-analysis. *Acta psychologica*, *148*, 163–172. https://doi.org/10.1016/j.actpsy.2014.01.016

Cowan, R., & Powell, D. (2014). The contributions of domain-general and numerical factors to third-grade arithmetic skills and mathematical learning disability. *Journal of educational psychology, 106*(1), 214–229. https://doi.org/10.1037/a0034097

D'Agostino, R. B. (1970). Transformation to Normality of the Null Distribution of g_1. *Biometrika*, *57*(3), 679–681. https://doi.org/10.2307/2334794

Dobson, A. J. (2002). *An introduction to generalized linear models*. Chapman & Hall/CRC.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Fahrmeir, L., Kneib, T., & Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer.

Fischer, U., Rösch, S., Nuerk, H.C., & Moeller, K. (2015). Erkennen von Rechenschwäche durch LehrerInnen und Testungen im Klassenverband. *Lernen und Lernstörungen*, *4*(4), 269–285. https://doi.org/10.1024/2235-0977/a000116

Fischer, U., Rösch, S., & Moeller, K. (2017). Diagnostik und Förderung bei Rechenschwäche. *Lernen und Lernstörungen, 6*(1), 25–38. https://doi.org/10.1024/2235-0977/a000160

Fischbach, A., Schuchardt, K., Brandenburg, J., Klesczewski, J., Balke-Melcher, C., Schmidt, C., … & Hasselhorn, M. (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

Haffner, J., Baro, K., Parzer, P., & Resch, F. (2005). *Heidelberger Rechentest (HRT 1 – 4)*. Hogrefe.

Helmke, A. (2017). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbes-*

serung des Unterrichts (7. Aufl.). Klett-Kallmeyer.

Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3., vollst. Überarb. und erw. Auflage). Budrich.

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social psychology of education*, *20*, 471–490. https://doi.org/10.1007/s11218-017-9384-z

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaft*, *18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern*. Waxmann.

Karst, K., Schoreit, E., & Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern. *Zeitschrift für pädagogische Psychologie*, *28*, 237–248. https://doi.org/10.1024/1010-0652/a000133

Kaufmann, E. (2020). How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*(4), 103298. https://doi.org/10.1016/j.tate.2021.103298

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

Kuhn, J.-T., Schwenk, C., Souvignier, E., & Holling, H. (2019). Arithmetische Kompetenz und Rechenschwäche am Ende der Grundschulzeit. Die Rolle statusdiagnostischer und lernverlaufsbezogener Prädiktoren. *Empirische Sonderpädagogik*, *11*(2), 95–117. https://doi.org/10.25656/01:17773

Lamb, S., Krieger, F., & Kuhn, J.-T. (2024 a). Delayed development of basic numerical skills in children with developmental dyscalculia. *Frontiers in Psychology*, *14*, 1187785. https://doi.org/10.3389/fpsyg.2023.1187785

Lamb, S., Schulz, A.-K., & Kuhn, J.-T. (2024 b). Entwicklung eines Lehrkräftefragebogens zur Früherkennung von Rechenstörungen in der Grundschule. *Lernen und Lernstörungen, 13*(4), 165–177. https://doi.org/10.1024/2235-0977/a000456

Landerl, K., Vogel, S., & Kaufmann, L. (2022). Dyskalkulie: Modelle, Diagnostik, Intervention. Ernst Reinhardt.

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen*. University of Bamberg Press.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science, 17*(5), 714–726. https://doi.org/10.1111/desc.12152

Mack, E., Gnas, J., Vock, M., & Preckel, F. (2023). The Domain-Specificity of Elementary School Teachers' Judgment Accuracy. *Contemporary Educational Psychology, 72*, 102142. https://doi.org/10.1016/j.cedpsych.2022.102142

Mayringer, H. & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufen 1 – 4 (SLS 1 – 4)*. Hans Huber.

Karing, C., & Artelt, C. (2014). Urteilsgenauigkeit von Lehrer(inne)n im emotionalmotivationalen Bereich und im Leistungsbereich. In M. Mudiappa & C. Artelt (Hrsg.) *BiKS – Ergebnisse aus den Längsschnittstudien. Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich* (S. 111–118). University of Bamberg Press.

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

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Salaschek, M., Zeuch, N., & Souvignier, E. (2014). Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense. *Learning and Individual Differences*, *35*, 103–112. https://doi.org/10.1016/j.lindif.2014.06.009

Schäfer, H., & Rittmeyer, C. (2015). Inklusive Diagnostik. In H. Schäfer & C. Rittmeyer (Hrsg.), *Handbuch Inklusive Diagnostik* (S. 103–134). Beltz.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Peter Lang.

Schrader, F.W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerbildung*, *31*(2), 154–165. https://doi.org/10.25656/01:13843

Schrader, F.W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik*, *1*(1), 27–52.

Schütky, R. (2022). Der Einfluss nicht-kognitiver Merkmale (Selbstkonzept, Stereotypen und Stereotype Threat) auf Mathematikleistungen im Bereich der Größen im Verlauf der Grundschulzeit. *Mathematik im Unterricht*, *13*, 31–46. https://doi.org/10.25598/miu/2022-13-3

Selter, C., Walter, D., Heinze, A., Brandt, J., & Jentsch, A. (2020). Mathematische Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnissse. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky, & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 25–56). Waxmann. https://doi.org/10.31244/9783830993193

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Wagner, L. (2024). Einflussfaktoren auf die Diagnosekompetenz (angehender) Lehrkräfte – ein systematisches Literaturreview. *Unterrichtswissenschaft,* 1–30. https://doi.org/10.1007/s42010-024-00215-3

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 – Revision (CFT 20-R)*. Hogrefe.

Weiß, R. H. & Osterland, J. (2012). *Grundintelligenztest: Skala 1-Revision (CFT 1-R)*. Hogrefe.

Autorinnen- und Autorenhinweis

Sarah Lamb
https://orcid.org/0000-0003-4619-3269

Jörg-Tobias Kuhn
https://orcid.org/0000-0002-4399-9569

Ann-Katrin Schulz

Korrespondenzadresse
Sarah Lamb
Fakultät für Rehabilitationswissenschaften, Technische Universtität Dortmund
Emil-Figge-Straße 50, D-44227 Dortmund,
sarah.lamb@tu-dortmund.de

Erstmals eingereicht: 22.12.2023
Überarbeitung eingereicht: 29.07.2024
Angenommen: 25.02.2025

Offene Daten	Bisher wurden die Daten nicht zur Verfügung gestellt, da weitere Publikationen geplant sind.
Offener Code	keine Angabe
Offene Materialien	Der in der Studie verwendete Fragebogen (FERMAT) ist öffentlich zugänglich.
Präregistrierung	Bisher nicht.
Votum Ethikkommission	Es liegt ein positives Ethikvotum der Ethikkommission des Fachbereichs 7 der Universität Münster vor.
Finanzielle und weitere sachliche Unterstützung	Diese Arbeit wurde finanziell durch das Bundesministerium für Bildung und Forschung unterstützt (Projektnummer: 01GJ1302).
Autorenschaft	SL hat den ersten Entwurf des Manuskripts geschrieben. A-KS und J-TK haben den Entwurf überarbeitet, J-TK hat die Studie geplant und die Daten erhoben, SL und J-TK haben die Daten analysiert.