

Evidenz. Besser. Kommunizieren.

Wie Wissenschaftskommunikation für Lehrpersonen weiterentwickelt werden kann

Samuel Merk, Sarah Bez und Kirstin Schmidt

Die bildungswissenschaftliche Literatur zu Schul- und Unterrichtsentwicklung bedient sich einer Vielzahl theoretischer Grundlegungen (Bohl, 2020) und blickt daher aus ganz verschiedenen Winkeln auf diesen Gegenstand: Neben eher systemtheoretischen Perspektiven (Bauer & Rolff, 1978) finden sich u. a. Ansätze mit Entlehnungen aus der Lehr-Lern- (Helmke, 2022) und Organisationspsychologie (Holtappels, 2007) oder mit dem Leitgedanken der Praxisorientierung (Brügelmann, 2018). Datenbasierte Schul- und Unterrichtsentwicklung hat im deutschsprachigen Raum erst in den vergangenen zwei Dekaden Verbreitung gefunden, wenngleich deren Grundidee des empirischen Einholens von Information über den Ist-Stand schon zuvor gefordert und auch umgesetzt wurde (Altrichter & Rolff, 2006). In jüngerer Zeit ist jedoch von inner- wie außerwissenschaftlichen Stakeholdern vermehrt die Forderung nach einer Entwicklung von Schule und Unterricht hörbar geworden, die ihre Entscheidungen durch Evidenz informiert (AERO, 2023; Bauer & Prenzel, 2012; Council of the European Union, 2024; Pellegrini & Vivanet, 2021; Slavin, 2020). Da jedoch einerseits die Genese und Interpretation von Evidenz nicht zu den ausgeprägtesten professionellen Kompetenzen von Lehrkräften gehört (Gussen et al., 2023) und andererseits Bildungswissenschaftler- und Fachdidaktiker:innen wohl oft keine Expert:innen für die praktische Gestaltung von Schule und Unterricht sein dürften, plädiert der vorliegende Beitrag dafür, Wissenschaftskommunikation erstens als wichtige Aufgabe von Forschenden aufzufassen, zweitens das Gelingen von Wissenschaftskommunikation zum Gegenstand empirischer Forschung zu machen und drittens die Entwicklung von neuen Wissenschaftskommunikationsformaten als dialogischen Prozess zwischen Forschenden und (angehenden) Lehrkräften aufzufassen.

Zur Begründung dieser Plädoyers führt der vorliegende Beitrag zunächst in Konzepte und Begriffe evidenzinformierter Praxis sowie in Wissenschaftskommunikation in der Lehrpersonenbildung ein. Anschließend wird die vorliegende Studie – ein kleines Experiment zur Kommunikation der Entwicklung der Lesekompetenz in den deutschen Kohorten des Programme of International Student Assessment (PISA) – als empirisches Beispiel angeführt.

Theoretischer Hintergrund

Evidenzinformiertes Handeln

Was kann unter ‚Evidenz‘ verstanden werden?

Etymologisch kann ‚Evidenz‘ als Substantivierung des Adjektivs ‚evident‘ gesehen werden (Kluge, 2011, S. 263), welches wiederum im 18. Jahrhundert dem lateinischen ‚evidens‘ (‚ersichtlich, augenscheinlich‘; Hau et al., 2012) entlehnt wurde (Stark, 2017). Allerdings meinen Forschende gerade nicht ‚das Augenscheinliche‘ oder ‚das direkt Ersichtliche‘, wenn sie von Evidenz sprechen. Vielmehr ist in Definitionsvorschlägen von ‚wissenschaftlichem Wissen‘ (Stark, 2017), von einer ‚Funktion‘ von Daten für die Bestätigung oder Widerlegung von Hypothesen und Theorien (Bromme et al., 2014) oder von ‚warrants for making assertions or knowledge claims‘ (Shavelson & Towne, 2002, S. 18) die Rede. In einer aktuellen Systematisierung verschiedener Verständnisse des Evidenz-Begriffs in den Bildungswissenschaften hebt Schmidt (2024) hervor, dass nur wenige Definitionen ausschließlich quantitativer Empirie die Möglichkeit zuschreiben, Evidenz zu generieren, sondern meistens auch qualitative Empirie, Theorien sowie mathematische und logische Analysen als potenziell evidenzgenerierend definiert werden. Insbesondere die Inklusion nicht-empirischer Entitäten wie Theorien oder logische Analysen mögen auf den ersten Blick widersprüchlich wirken, da der Begriff Evidenz insbesondere im deutschsprachigen Raum teils mit Ergebnissen explanativer quantitativer Studien assoziiert scheint. Dieser scheinbare Widerspruch wirkt jedoch weniger stark, berücksichtigt man, dass insbesondere in der Lehr-Lernforschung mit ‚Theorien‘ wohl eher sogenannte ‚tried-and-tested theories‘ (Renkl, 2022) gemeint sein dürften. Diese stellen eher Rahmenmodelle oder sogenannte ‚interventional models‘ (z. B. Cognitive Theory of Multi-Media Learning)

dar (Renkl, 2022). Da solche Theorien wiederum meist stark von empirischen Ergebnissen beeinflusst sind, ist es plausibel, ihnen die Funktion als ‚warrant‘ für ‚knowledge claims‘ zuzuschreiben und sie also auch als Evidenz zu bezeichnen.

Evidenzinformiert, evidenzorientiert, evidenzbasiert

Im vorherigen Abschnitt wurde deutlich, dass Evidenz ein uneinheitlich gebrauchter und gleichermaßen komplex wie unscharf definierter Begriff ist. Im Lichte dessen erscheint es nur konsequent, dass auch die Begriffe evidenzbasiert, evidenzinformiert, evidenzorientiert, datenbasiert, forschungsbasiert und forschungsinformiert als Jingle Jangle eingeordnet werden können (Kelley, 1927; Thorndike, 1904), dass also unterschiedliche Begriffe für das Gleiche und gleiche Begriffe für Unterschiedliches gebraucht werden. Dabei speisen sich die Differenzierungen von *evidenzbasiert* und *evidenzinformiert* sowie *evidenzorientiert* aus recht verschiedenen ontologischen, epistemologischen und wissenschaftskritischen (Schmid & Lutz, 2007) Überlegungen: Mit Evidenz*basierung* wird z. B. oft ‚the medical model‘ (Jones, 2024) im Sinne von Evidenz aus Kontrollgruppenexperimenten als notwendige Voraussetzung für praktische Entscheidungen assoziiert und damit sowohl epistemologische (hier Kontrollgruppenexperiment) wie wissenschaftskritische (hier notwendige Voraussetzung) Kriterien zur Abgrenzung herangezogen. Den Begriffen ‚*evidenzorientiert*‘ und ‚*evidenzinformiert*‘ wird zum einen ein weniger enger Evidenzbegriff zugeordnet (ontologischer und epistemologischer Aspekt; Stark, 2017) und zum anderen der Evidenz in der praktischen Verwendung eher eine heuristische Funktion (wissenschaftskritischer Aspekt).

In der deutschsprachigen bildungswissenschaftlichen Diskussion sind nach Bromme et al. (2014) zunächst zwei verschiedene Diskussionsstränge bzgl. evidenzinformierter Entscheidungen im Bildungskontext unterscheidbar: Ein Diskussionsstrang beschäftigt sich mit evidenzinformierten Entscheidungen in der Bildungspolitik und der andere mit evidenzinformierten Entscheidungen und Handlungen in der Bildungspraxis. In beiden Diskussionen werden der Evidenz verschiedene Funktionen zugeschrieben. Bromme et al. (2014) etwa sprechen davon, dass Evidenz über Zustände informieren, Mechanismen erklären oder Interventionen evaluieren kann. Groß Ophoff et al. (2023) wiederum unterscheiden konzeptuelle Nutzung („evidence allows focusing attention, provides new insights, chal-

lenges beliefs or reframes thinking“; S. 2), instrumentelle Nutzung („identify or develop concrete measures to be taken“; S. 2) und symbolische Nutzung („justif[y] or support of existing positions or established procedures“; S. 2).

Potenzielle Wege zu einer gelingenden Wissenschaftskommunikation

Unabhängig vom Kontext und der Funktion evidenzinformierter Entscheidungen ist es plausibel anzunehmen, dass eine erfolgreiche Kommunikation von Evidenz (im Sinne der Induktionen einer konsistenten Auffassung) eine notwendige Voraussetzung für das Gelingen evidenzinformierter Entscheidungen ist: Wird Evidenz fehlinterpretiert und erfolgt eine anschließende Entscheidung kohärent zu dieser Fehlinterpretation, wird die Wirkung dieser Entscheidung nicht die Erwünschte sein.

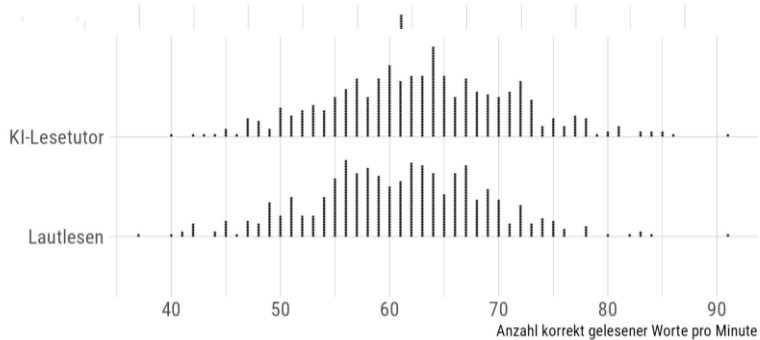
Liest eine Lehrkraft etwa die (fiktive) Pressemitteilung in Abbildung 1, stellt sich die Ergebnisse darauf basierend wie in Abbildung 1 Mitte vor (Schmidt et al., 2023) und überzeugt anschließend ihre Schulleitung, diesen KI-Lesetutor zu beschaffen und schulweit einzusetzen, liegt höchstwahrscheinlich dysfunktionales evidenzinformiertes Handeln vor. Während die Forscher:innen mit ‚signifikant bessere Leseflüssigkeit‘ zum Ausdruck bringen, dass ihre Daten unter der Annahme eines Nulleffekts unwahrscheinlich sind (signifikanter p -Wert), interpretiert die Lehrkraft diese Formulierung als ‚Unterschied bedeutsamer Größe‘. Folglich schlussfolgert sie, dass es Sinn macht, Geld und Zeit in Anschaffung und Implementation des KI-Lesetutors zu investieren, weil sie den KI-Lesetutor für deutlich lernwirksamer hält als lautes Lesen, obwohl evtl. die Implementation von Lautlesetandems lernwirksamer, kostengünstiger und weniger zeitaufwändig gewesen wäre.

Die Forschung zur Wissenschaftskommunikation hat eine Reihe solcher potenzieller Problematiken aufgezeigt: Z. B. das soeben beschriebene Verwechseln von Inferenzstatistik und Effektstärke (Schmidt et al., 2023), das automatische Annehmen starker Effekte, wenn keine Effektstärken berichtet wurden (Practical Significance Bias; Michal & Shah, 2024), Rückschaufehler (Masnick & Zimmerman, 2009) oder die verzerrte Einschätzung der Belastbarkeit von Befunden (z. B. das Ergebnis einer Laborstudie mit $N = 56$ mit großem Effekt und daher hoher statistischer Power) durch

irrelevante Zahlen (z. B. Stichprobengröße einer zuvor gelesenen Large-Scale-Studie; Bohrer et al., 2025).

Daten

der Studie



Ausschnitt

der Pressemitteilung

In einer randomisierten Studie mit $N = 1001$ Drittklässler:innen zeigten diejenigen, die täglich 15 Minuten mit dem KI-Vorlesetutor übten, **signifikant bessere Leseflüssigkeit**, als Drittklässler:innen, die täglich 15 Minuten (ohne KI-Tutor) laut lasen.

Interpretation

einer Lehrkraft

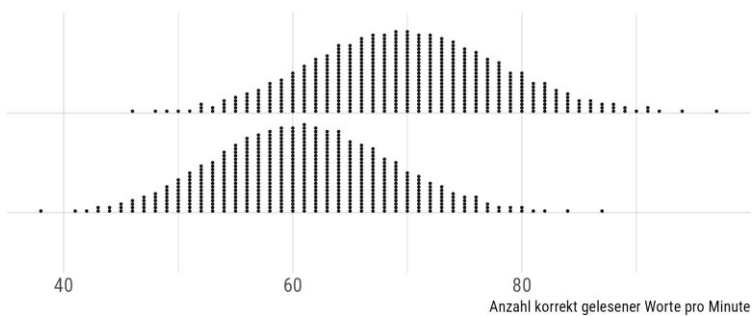


Abb. 1: Daten der (fiktiven) Studie, Pressemitteilung und Vorstellung der Lehrkraft.

Gleichzeitig liegt eine Reihe von Befunden vor, die implizieren, dass verbesserte Kommunikation von Evidenz an Lehrkräfte zu Zwecken evidenzinformatierten Handelns vergleichsweise einfach umsetzbar ist (z. B. Schneider et al., 2025). Grundsätzlich lassen sich die bisherigen Befunde in angebotsseitige und nutzendenseitige Ansätze unterscheiden, also in Interventionen, die die Auswahl und Darstellung der Evidenz optimieren möchten und Ansätze, die bei der Scientific, Data und Statistical Literacy der (angehenden) Lehrkräfte ansetzen (Brühwiler & Leutwyler, 2020).

Zu zweitem gehören Programme wie ‚Data Teams‘ (Schildkamp et al., 2018), welche durch ein umfangreiches Set an vordefinierten Leitlinien und Aktivitäten versuchen, konkrete schulische Probleme mit Hilfe von (oft eigens dafür generierten) Daten zu lösen, wobei meist vier bis sechs Lehrkräfte und Schulleiter:innen mit Forschenden kooperieren. Hierzu gehören auch ‚Brokering-Ansätze‘ (teilweise als ‚research practice partnerships‘ bezeichnet), in welchen Wissenschaftler:innen und Lehrpersonen (insbesondere Schulleitungen) gemeinsam versuchen, konkrete schulische Probleme unter Rückgriff auf wissenschaftliche Erkenntnisse zu lösen (z. B. Sharples & Sheard, 2015). Zudem können kurz- (Merk et al., 2020) oder längerfristig (Karst et al., 2024) angelegte Interventionen zur Anbahnung notwendiger Kompetenzen für evidenzinformatiertes Handeln wie die Interpretation von grafisch dargestellten Daten (Friel et al., 2001) oder Forschungskompetenz (Neuenschwander, 2005) sowie die konkrete Unterstützung für evidenzinformatiertes Handeln (Clearing House Unterricht Academy, 2025) diesem Ansatz zugerechnet werden.

Angebotsseitige Versuche, die Kommunikation von Evidenz zu verbessern, stammen aus verschiedensten Disziplinen: So wird z. B. in der Psychologie untersucht (Grice et al., 2020), welche algebraisch äquivalenten Formulierungen zu standardisierten Effektstärken bei der Rezeption durch Laien adäquatere Vorstellungen induzieren (siehe Tabelle 1). In der Human-Computer-Interaction-Forschung werden (teils dynamische) Visualisierungstechniken entwickelt, um Effektstärken und Inferential Uncertainty besser zu kommunizieren (z. B. Hullman et al., 2015; Zhang et al., 2023). Die bildungswissenschaftliche Lehrpersonenbildungsforschung sowie die Fachdidaktiken erproben innovative Formate für die Zielgruppe der (angehenden) Lehrkräfte (z. B. Rycroft-Smith & Stylianides, 2022; Schneider et al., 2025), was auch das Anliegen der vorliegenden Studie ist.

Tab. 1: Beispiele für angebotsseitige Versuche verbesserter Kommunikation von Evidenz.

	Unterschied	Zusammenhang
Standard-kommuni-kation	Die Leseleistung von Schüle-rinnen und Schülern sank in PISA 2022 um 28 Punkte und damit auf den Tiefststand.	Der sozioökonomische Status klärt 14% der Varianz der Mathematiklei-stung auf.
Verbesserte Kommunikation	Die Leseleistungen von Schü-lerinnen und Schülern in Deutschland aus PISA 2018 und aus PISA 2022 überlappen sich zu 88,9%, wobei der Mit-telwert um 28 Punkte sank.	Von 100 Schülerinnen und Schülern, die einen überdurchschnittlichen sozioökonomischen Status haben, zeigen 69 eine überdurchschnittliche Mathematikleistung.

Die vorliegende Studie

Die vorliegende Studie untersucht, inwiefern verbreitete Standardgrafiken zur Kommunikation der Entwicklung der Lesekompetenz in den deutschen Kohorten des Programme of International Student Assessment (PISA) ‚Practical Significance Bias‘, also das automatische Annehmen starker Ef-fekte, wenn keine Effektstärken berichtet werden, induzieren und ob dieser mit Grafiken verringert werden kann, bei deren Gestaltung theoretische und empirische Erkenntnisse der Wissenschaftskommunikation berück-sichtigt wurden.

Methode

Materialien

In der wissenschaftlichen wie journalistischen Berichterstattung zu den Er-gebnissen der PISA-2022-Kohorte wurden zahlreiche Darstellungsformate gewählt, insbesondere Liniendiagramme (siehe Abbildung 2 bis 4), was an-gesichts der Anlage von PISA als Trendstudie (Döring & Bortz, 2016) kon-sequent erscheint.

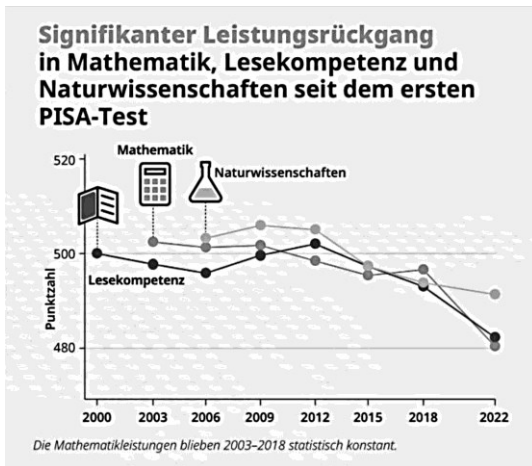


Abb. 2: Verwendetes Liniendiagramm in der Berichterstattung; Quelle: OECD (2023).

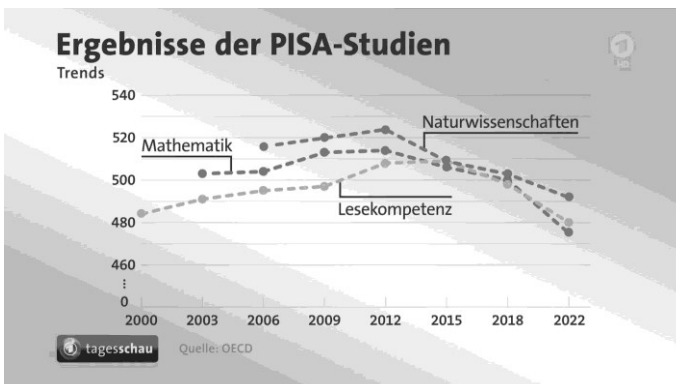


Abb. 3: Verwendetes Liniendiagramm in der Berichterstattung; Quelle: RBB (2023).

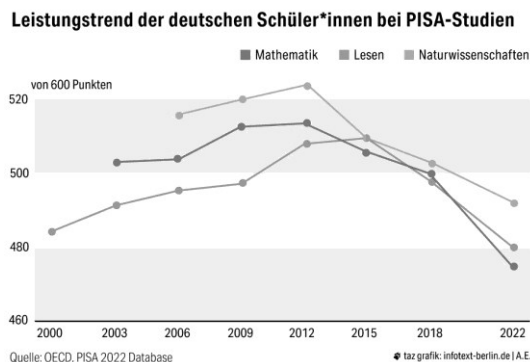


Abb. 4: Verwendetes Liniendiagramm in der Berichterstattung; Quelle: taz (2023).

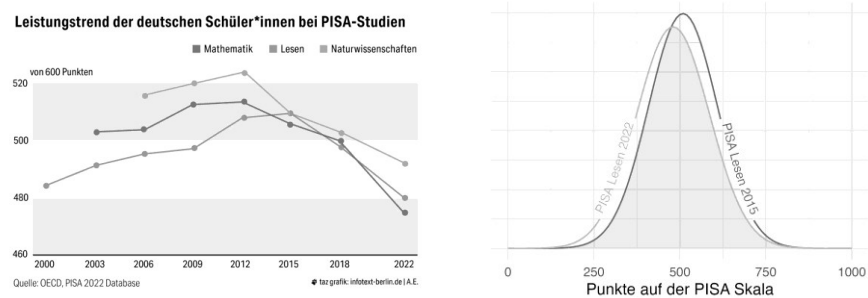
In solchen Grafiken ist jedoch die Bedeutsamkeit der Mittelwertsdifferenz nur bei bekannter Streuung interpretierbar: Abbildung 5 zeigt jeweils die gleichen Mittelwerte von 508 (PISA Lesen 2015) und 480 (PISA Lesen 2022) bei unterschiedlichen Streuungen.

Diese Liniendiagramme erlauben einen effizienten Vergleich der Mittelwerte sowohl über die Zeit als auch über Variablen (hier: Fächer) hinweg.

Das Ausmaß der Bedeutsamkeit dieses (gleichen) Mittelwertsunterschiedes entsteht aber erst durch die Streuung der Daten um diesen Mittelwert herum. Weil die Variablen im unteren Teil der Abbildung weniger streuen, ist die Überlappung der beiden Gruppen geringer (48%, großer Effekt), während die große Überlappung im oberen Teil (88%, kleiner Effekt) durch die große Streuung zustande kommt. Die Abbildungen 2 bis 4 sagen also nicht nur nichts über die Bedeutsamkeit der Mittelwertsunterschiede aus. Die nicht dargestellte Varianz induziert möglicherweise auch eine wahrgenommene große Bedeutsamkeit der Mittelwertsdifferenz (Kale et al., 2020).

Daher wurden vorliegend neben Liniendiagrammen auch überlappende Verteilungskurven verwendet. Um diese barriereärmer zu gestalten, wurde bei der Farbgebung auf hinreichenden Kontrast bei den prävalenten Sehbeeinträchtigungen geachtet (Garnier et al., 2023). Um unnötige Arbeitsgedächtnisbelastung zu vermeiden, wurde die Legende direkt in die Grafik integriert (Franconeri et al., 2021).

Tab. 2: Verwendete Stimuli.



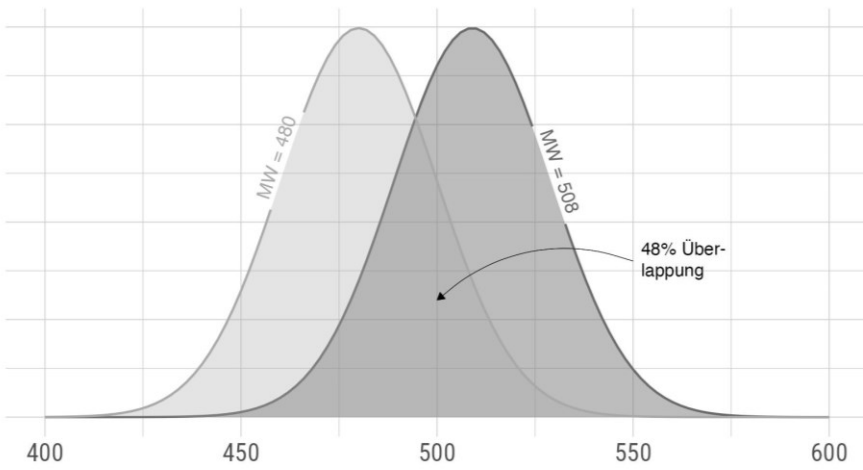
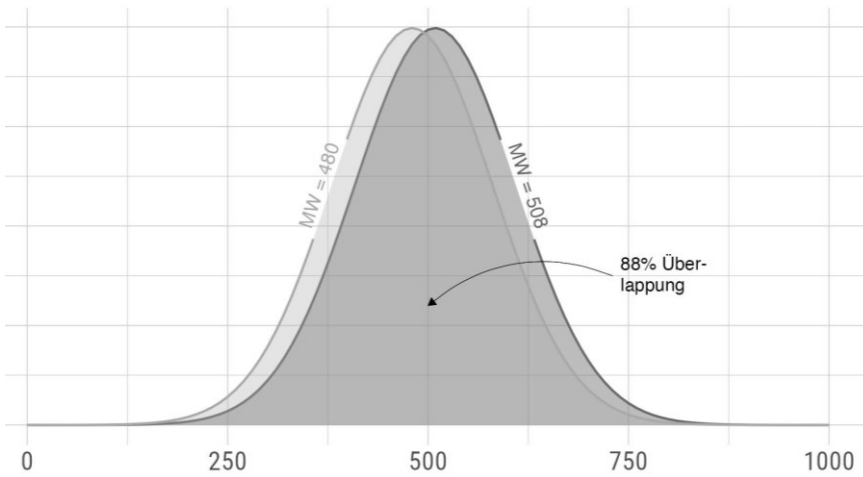


Abb. 5: Illustration der Unabhängigkeit von Mittelwertsdifferenz und Größe des Effekts.

Design, Stichprobe und Instrument

In einem Between-Person Design wurde $N = 195$ Studierenden in Bachelorstudiengängen des Primar- und Sekundarstufenlehramtes randomisiert eine der beiden in Tabelle 2 dargestellten Abbildungen gezeigt. Anschließend wurden sie mit folgendem Stimulus aufgefordert, die Effektstärke einzuschätzen: *„Basierend auf dieser Grafik: Wie hoch schätzen (exakte Antwort nicht möglich) Sie die Wahrscheinlichkeit ein, dass eine zufällig gezogene Schülerin oder ein zufällig gezogener Schüler aus dem Jahr 2022 im Lesen schlechter abscheidet als eine zufällig gezogene Schülerin oder Schüler aus dem Jahr 2015?“*. Beantwortet wurde diese Frage mit einem Schieberegler, dessen Enden mit *„50% (beide Gruppen gleich)“* und *„100% (maximaler Effekt)“* beschriftet waren. Diese Erfassung der wahrgenommenen Effektstärke als ‚Probability of Superiority‘ ist in der Human-Computer-Interaction-Forschung verbreitet und gilt als valide (Brooks et al., 2014; Kim et al., 2022), wenngleich die Operationalisierung als Schieberegler unklar lässt, inwiefern bei der Beantwortung tatsächlich eine Elaboration der Überlappung vorgenommen wird oder die Teilnehmenden eher intuitiv (etwa wie bei einem Likert-Item) vorgehen.

Statistische Analyse

Die abhängige Variable ‚Wahrgenommene Effektstärke‘ (operationalisiert als Probability of Superiority) ist per Design auf das geschlossene Intervall $[0,5; 1]$ beschränkt und zeigt empirisch Bimodalität (siehe Abbildung 6). Um diesen Umständen in der inferenzstatistischen Modellierung Rechnung zu tragen, wurden bayesianische Mixture Regressionsmodelle für zwei trunkierte Normalverteilungen (Frischkorn & Popov, 2023) in der probabilistischen Sprache Stan (Stan Development Team, 2024) mithilfe des R-Pakets {brms} (Bürkner, 2017) geschätzt.

Ergebnisse

Die Inspektion des Marcov-Chain-Monte-Carlo-Sampling-Prozesses zeigte eine zufriedenstellende Qualität bzgl. Konvergenz $\hat{R} < 1.01$ und effektiver Sampling Size ($ESS_{Bulk} > 1000 < ESS_{Tail}$; Vehtari et al., 2021).

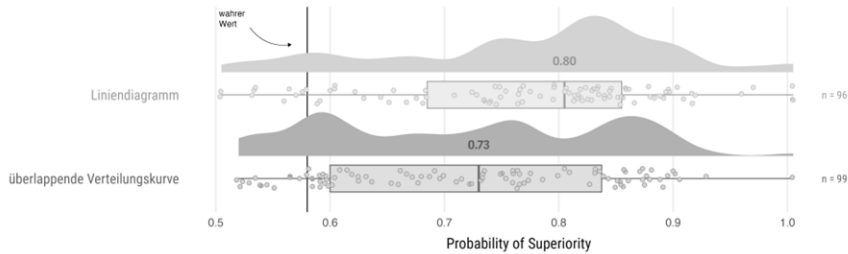


Abb. 6: Geschätzte Effektstärke (Probability of Superiority) nach Stimulus. Beide Gruppen zeigen einen sehr deutlichen Practical Significance Bias (Abstand von Median und wahrem Wert).

Die Medianeinschätzung der Probability of Superiority lag in beiden Gruppen deutlich über dem wahren Wert (Liniendiagramm .80, Überlappungsgrafik .73). Dieser Unterschied in der Einschätzung entspricht einer Überlappung von 81.71% (Cliff's $d = 0.23$) oder anders ausgedrückt: Legt man 100-mal einem:einer Studierenden die Originalgrafik und einem:einer Studierenden die Überlappungsgrafik vor, schätzt 61mal die:der Studierende mit der Überlappungsgrafik den Effekt weniger verzerrt ein. Die Inferenzstatistik für diesen Unterschied ist mit einer Evidence Ratio von 14.8 klar konklusiv: Die Alternativhypothese einer kleineren Probability of Superiority für die Überlappungsgrafik ist gegeben die Daten 14,8-fach wahrscheinlicher als die Nullhypothese einer größeren Probability of Superiority.

Limitationen

Die vorgestellte Studie stellt lediglich ein kleines Experiment an einer selektiven Gelegenheitsstichprobe ohne Präregistrierung dar. Interne und ökologische Validität konnten aufgrund des experimentellen Designs mit realweltlichen Materialien dennoch gestärkt werden. Die externe und Konstruktvalidität ist jedoch sicher aufgrund der großen Homogenität und Spezifität der Stichprobe, des sehr spezifisch gewählten Kontexts (PISA Lesekompetenz) sowie theoretisch etwas arbiträrer unabhängiger Variablen (viele weitere Gestaltungsprinzipien wären denkbar) eingeschränkt. Zudem

wäre die Untersuchung weiterer abhängiger Variablen wie Cognitive Load oder wahrgenommene Informativität wünschenswert.

Diskussion

Der vorliegende Beitrag zielt darauf ab, zu eruieren, inwiefern es nach dem Stand der Forschung gestaltete Wissenschaftskommunikation ermöglicht, Evidenz ‚besser‘ an Lehramtsstudierende und Lehrkräfte zu kommunizieren. Dabei wurde ‚besser‘ als ‚weniger gebiased‘ operationalisiert und gezeigt, dass die Wahl einer theoretisch fundierten grafischen Darstellung einen deutlich geringeren Bias induzierte als eine Standardgrafik. Allerdings war auch die Rezeption der verbesserten Darstellung immer noch erheblich verzerrt (siehe Abbildung 6).

Im Lichte dieser Ergebnisse werden im Folgenden drei Implikationen diskutiert: 1) Die Forderung, dass Lehrkräfte ihre professionelle Praxis evidenzinformiert gestalten sollen, setzt Anstrengungen in der Wissenschaftskommunikation seitens der Bildungswissenschaften und Fachdidaktiken voraus. 2) Inwiefern diese Anstrengungen zielführend sind, sollte empirisch überprüft werden. 3) Erfolgreiche Wissenschaftskommunikation in den Bildungswissenschaften und Fachdidaktiken impliziert eine Passung von Angebots- und Nutzendenmerkmalen und damit einen dialogischen Prozess für die Entwicklung einer solchen Passung.

Sowohl Wissenschaftstheorie (z. B. Mitchell & Jolley, 2010) als auch bildungswissenschaftliche Literatur (z. B. Bohl et al., 2015; Dewe et al., 1992) haben die Möglichkeiten und Limitationen der Abgrenzbarkeit von ‚Wissenschaft‘ und ‚Nicht-Wissenschaft‘ (bzw. in den Bildungswissenschaften von ‚Theorie‘ und ‚Praxis‘) diskutiert und heben u. a. hervor, dass Entitäten und Aussagen in ihrer Bedeutung primär an den Herkunftskontext (z. B. ‚Wissenschaft‘ oder ‚Praxis‘) gebunden sind. Also ist auch z. B. die ‚Evidenz‘ einer explanativen bildungswissenschaftlichen Studie per se zunächst bildungswissenschaftlich und muss für eine evidenzinformierte Handlung in der Praxis reinterpretiert werden (z. B. Groß Ophoff et al., 2023). Damit liegt es auf der Hand, dass sich Forschende fragen sollten, welche ‚wissenschaftlichen/theoretischen‘ Entitäten (z. B. Effektstärken oder inferenzstatistischen Maße) und Aussagen (z. B. kausale Effekte) sie wie in die Kommunikation ihrer Ergebnisse gegenüber der Praxis aufnehmen.

Dass diese Forderung selbst rein innerwissenschaftlich betrachtet nicht trivial ist, zeigt z. B. die Tatsache, dass Guidelines von Fachgesellschaften wie z.B. der American Psychological Association (2019) die Verwendung von Effektstärken fordern, diese aber in Pressemitteilungen (etwa der American Educational Research Association) äußerst selten und selbst in Fachzeitschriften nicht unbedingt erwartbar sind (McMillan & Jennifer, 2011). Für eine Wissenschaftskommunikation, die sich an die Praxis richtet, scheint es also plausibel zu schlussfolgern, dass es unter Forschenden noch nicht verbreitet scheint, sich literaturbasiert darüber Gedanken zu machen, inwiefern die eigene Wissenschaftskommunikation z. B. für Lehrkräfte günstig rezipierbar ist.

Doch selbst ein Bewusstsein für die Fallstricke der Kommunikation wissenschaftlicher Ergebnisse schützt nicht zwangsläufig vor der Induktion von Fehlvorstellungen: So fanden Schneider et al. (2025) etwa, dass selbst eine als leicht verständlich geltende Effektstärke für Mittelwertsvergleiche (wie etwa Cohen's U_3) bei einem erheblichen Anteil ($\geq 29\%$) der Rezipient:innen zu Fehlvorstellungen führte. Die erste Implikation scheint also nicht hinreichend für eine gelingende Kommunikation von Evidenz an Lehrkräfte.

Dies führt zur zweiten Implikation: Forschende sollten nicht nur den Stand der Forschung bei der Kommunikation von Evidenz berücksichtigen, sondern auch in intern und extern validen Studien untersuchen, inwieweit diese Berücksichtigung erfolgreich war. Denn statistische Informationen werden nicht nur von unterschiedlichen Berufsgruppen (McDowell & Jacobs, 2017), sondern auch in unterschiedlichen geografischen Regionen differentiell interpretiert (Gigerenzer et al., 2005). Inwiefern sich also generische Determinanten erfolgreicher Wissenschaftskommunikation auf die Kommunikation von Evidenz an Lehrkräfte etwa in einem bestimmten Teil eines Bildungssystems übertragen lassen, scheint nur schwer a priori bestimmbar.

Was aber, wenn Forschende ihre Wissenschaftskommunikation literaturbasiert verbessern, aber in empirischen Experimenten sehen, dass sie dennoch deutlich verzerrt, verrauscht oder konzeptuell falsch rezipiert wird? Der vorliegende Beitrag macht als dritte Implikation den Vorschlag, die Kommunikation von Evidenz an Lehrkräfte dialogisch weiterzuentwi-

ckeln und zu berücksichtigen, dass bei der Rezeption vermutlich eine komplexe Interaktion von Angebots- und Nutzendenmerkmalen (Brühwiler & Leutwyler, 2020) sowie Bottom-Up- bzw. Top-Down-Prozessen (Schmidt, 2024) vorliegt: Man stelle sich eine Lehrkraft vor, die auf der Suche nach einer Entscheidungsgrundlage für oder gegen eine unterrichtsgestalterische Maßnahme A auf der Seite eines Clearing Houses landet. Dort liest sie, dass über viele Studien gemittelt Maßnahme A dazu geführt hat, dass 63% der Schüler:innen bessere Leistungen zeigen als der Mittelwert der Schüler:innen mit Maßnahme B. Dann können daraus manche Lehrkräfte möglicherweise anhand ihres Vorwissens unmittelbar eine korrekte/konsistente Vorstellung der Effektstärke dieses Unterschieds *schlussfolgern* (z. B. zwei Normalverteilungen mit 87% Überlappung). Hier läge ein Top-Down-Prozess vor, da die Merkmale der Kommunikation mit den im Langzeitgedächtnis der Rezipient:in vorhandenen Dispositionen wie Graph, Data oder Statistical Literacy dazu führen, dass in einem Schlussfolgerungsprozess ein korrektes mentales Modell erstellt wird. Umgekehrt kann es passieren, dass eine Lehrkraft auf diese Formulierung stößt und eben kein auf Wissen basierendes mentales Modell abrufen kann, aber sich Stück für Stück mithilfe der gegebenen Informationen ein konsistentes mentales Modell erarbeitet. Dabei *lernt* sie, d. h. erwirbt Graph, Data, oder Statistical Literacy, was einem Bottom-Up-Prozess entspricht. Da Lehrkräfte über sehr unterschiedliche Dispositionen zu Top-Down-Prozessen verfügen, aber auch Bottom-Up-Prozesse sehr individuell verlaufen dürften, liegt die dritte Implikation nahe: Die Kommunikation von Evidenz an Lehrkräfte sollte als dialogischer und differenzieller Prozess aufgefasst werden. Demnach würden zum einen Forschende Kenntnis über Top-Down- und Bottom-Up-Prozesse ihrer Rezipienten:innen erwerben und deren Ausprägung und Entwicklung z. B. anhand von Think-Aloud-Studien wie z. B. Bez et al. (2021) beobachten und daraufhin ihre Angebote entsprechend differenzieren und anpassen. Zum anderen könnten Lehrkräfte in die Entwicklung von Kommunikationsprodukten anhand kokonstruktiver Verfahren eingebunden werden, in der Hoffnung, dass eine solche Kooperation von Akteur:innen aus den Systemen ‚Wissenschaft/Theorie‘ und ‚Nicht-Wissenschaft/Praxis‘ dazu führt, dass innerhalb dieser Systeme Ausdrucksweisen verfügbar werden, die zu verlustfreieren und damit erfolgreicherem Kommunikationsprozessen führen können (Leitz et al., 2024).

Literatur

- AERO. (2023).** *Evidence-based teaching practices*. Australian Education Research Organisation. Abgerufen am 31.01.2025 von <https://www.education.gov.au/quality-initial-teacher-education-review/resources/aero-evidence-based-teaching-practices>.
- ALTRICHTER, H., & ROLFF, H.-G. (2006).** Datenbasierte Schulentwicklung. Editorial. *Journal für Schulentwicklung*, 10(4), 4–6.
- AMERICAN PSYCHOLOGICAL ASSOCIATION. (2020).** *Publication manual of the American Psychological Association* (7. Aufl.). American Psychological Association.
- BAUER, J., & PRENZEL, M. (2012).** European teacher training reforms. *Science*, 336(6089), 1642–1643.
- BAUER, K.-O., & ROLFF, H.-G. (1978).** Vorarbeiten zu einer Theorie der Schulentwicklung. In K.-O. Bauer & H.-G. Rolff (Hrsg.), *Innovation und Schulentwicklung* (S. 219–263). Beltz.
- BEZ, S., POINDL, S., BOHL, T., & MERK, S. (2021).** Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien. *Zeitschrift für Pädagogik*, 67(4), 551–572.
- BOHL, T. (2020).** Theorien der Schulentwicklung. In M. Harant, P. Thomas, & U. Küchler (Hrsg.), *Theorien! Horizonte für die Lehrerinnen- und Lehrerbildung* (S. 97–109). Tübingen University Press.
- BOHL, T., WACKER, A., & HARANT, M. (2015).** *Schulpädagogik und Schultheorie*. utb.
- BOHRER, K., SCHMIDT, K., & MERK, S. (2025).** Zwei Studien, ein Ergebnis: Lehramtsstudierende unterliegen im Umgang mit Evidenz dem Ankereffekt. *Zeitschrift für Erziehungswissenschaft*.
- BROMME, R., PRENZEL, M., & JÄGER, M. (2014).** Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft*, 17(4), 3–54.
- BROOKS, M. E., DALAL, D. K., & NOLAN, K. P. (2014).** Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, 99(2), 332–340.
- BRÜGELMANN, H. (2018).** Unterrichts- und Schulentwicklung in Communities of Practice. In H. Barz (Hrsg.), *Handbuch Bildungsreform und Reformpädagogik* (S. 479–484). Springer Fachmedien.
- BRÜHWILER, C., & LEUTWYLER, B. (2020).** Praxisrelevanz von Forschung als gemeinsame Aufgabe von Wissenschaft und Praxis: Entwurf eines Angebots-Nutzungs-Modells. *BzL - Beiträge zur Lehrerinnen- und Lehrerbildung*, 38(1), 21–36.
- BÜRKNER, P.-C. (2017).** brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28.
- CLEARING HOUSE UNTERRICHT ACADEMY. (2025).** *Clearing House Unterricht Academy*. Abgerufen am 23.01.2025 von <https://clearinghouse-academy.de/>
- COUNCIL OF THE EUROPEAN UNION. (2024).** *Council conclusions on promoting evidence-informed policy and practice in education and training to achieve the European Education Area*. Abgerufen am

- 31.01.2025 von https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C_202403642
- DEWE, B., FERCHHOFF, W., & RADTKE, F.-O. (1992). Das „Professionswissen“ von Pädagogen. In B. Dewe, W. Ferchhoff, & F. Olaf-Radtke (Hrsg.), *Erziehen als Profession* (S. 70–91). Springer VS.
- DÖRING, N., & BORTZ, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer.
- FRANCONERI, S. L., PADILLA, L. M., SHAH, P., ZACKS, J. M., & HULLMAN, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3), 110–161.
- FRIEL, S. N., CURCIO, F. R., & BRIGHT, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2), 124–158.
- FRISCHKORN, G. T., & POPOV, V. (2023). *A tutorial for estimating Bayesian hierarchical mixture models for visual working memory tasks: Introducing the Bayesian measurement modeling (bmm) package for R*. *PsyArXiv*. https://doi.org/10.1007/978-3-663-09988-8_5
- GARNIER, S., ROSS, N., RUDIS, B., SCIAINI, M., CAMARGO, A. P., & SCHERER, C. (2024). *viridis: Colorblind-Friendly Color Maps for R (Version 0.6.5)* [Software]. <https://doi.org/10.32614/CRAN.package.viridis>
- GIGERENZER, G., HERTWIG, R., VAN DEN BROEK, E., FASOLO, B., & KATSIKOPOULOS, K. V. (2005). “A 30% Chance of Rain Tomorrow”: How does the public understand probabilistic weather forecasts? *Risk Analysis*, 25(3), 623–629.
- GRICE, J. W., MEDELLIN, E., JONES, I., HORVATH, S., MCDANIEL, H., O’LANSEN, C., & BAKER, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, 3(4), 443–455.
- GROB OPHOFF, J., BROWN, C., & HELM, C. (2023). Do pupils at research-informed schools actually perform better? Findings from a study at English schools. *Frontiers in Education*, 7, 1011241.
- GUSSEN, L., SCHUMACHER, F., GROßMANN, N., FERREIRA GONZÁLEZ, L., SCHLÜTER, K., & GROBSCHEDL, J. (2023). Supporting pre-service teachers in developing research competence. *Frontiers in Education*, 8, 1197938.
- HAU, R., MARTINI, U., & DRALLE, A. (2012). *PONS Wörterbuch für Schule und Studium Latein-Deutsch*. PONS.
- HELMKE, A. (2022). *Unterrichtsqualität und Professionalisierung: Diagnostik von Lehr-Lern-Prozessen und evidenzbasierte Unterrichtsentwicklung*. Klett Kallmeyer.

- HOLTAPPELS, H. G. (2007).** Schulentwicklungsprozesse und Change Management. Innovations- theoretische Reflexionen und Forschungsbefunde über Steuergruppen. In N. Berkemeyer (Hrsg.), *Schulische Steuergruppen und Change Management. Theoretische Ansätze und empirische Befunde zur schulinternen Schulentwicklung* (S. 11–39). Juventa.
- HULLMAN, J., RESNICK, P., & ADAR, E. (2015).** Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, 10(11), e0142444.
- JONES, A. (2024).** Rethinking evidence-based practice in education: A critical literature review of the ‘What Works’ approach. *International Journal of Educational Researchers*, 15(2), 37–51.
- KALE, A., KAY, M., & HULLMAN, J. (2021).** Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 272–282.
- KARST, K., YENDELL, O., MARX, A., LETTAU, W.-D., & HAWLITSCHKE, P. (2024).** Die Etablierung von Evidenzteams in SchuMaS – Eine Strategie zur systematischen Nutzung von Daten für die Schul- und Unterrichtsentwicklung. In K. Maaz & A. Marx (Hrsg.), *SchuMaS – Schule macht stark. Sozialraumorientierte Schul- und Unterrichtsentwicklung an Schulen in schwierigen Lagen. Aufbau und erste Arbeitsergebnisse des Forschungsverbunds*. (S. 225–240). Waxmann.
- KELLEY, T. L. (1927).** *Interpretation of educational measurements*. World Book Company.
- KIM, Y.-S., HOFMAN, J. M., & GOLDSTEIN, D. G. (2022, 30. APRIL-05. MAI).** *Putting scientific results in perspective: Improving the communication of standardized effect sizes* [Konferenzbeitrag]. CHI Conference on Human Factors in Computing Systems (CHI 22), New Orleans, USA.
- KLUGE, F. (2011).** *Etymologisches Wörterbuch der deutschen Sprache* (25. Aufl.). De Gruyter.
- LEITZ, A., KLEEN, H., HARTMANN, U., & KUNTER, M. (2024, 18.-20. MÄRZ).** *Was können wir aus der Grundschullehrkraft-Perspektive auf Transferaktivitäten lernen? Die Vorstellung eines kokonstruktiven Clearinghouse-Ansatzes* [Konferenzbeitrag]. 11. Jahrestagung der Gesellschaft für empirische Bildungsforschung (GEBF), Potsdam, Deutschland.
- MASNICK, A. M., & ZIMMERMAN, C. (2009).** Evaluating scientific research in the context of prior belief: Hindsight bias or confirmation bias? *Journal of Psychology of Science and Technology*, 2(1), 29–36.
- MCDOWELL, M., & JACOBS, P. (2017).** Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12), 1273–1312.
- MCMILLAN, J. H., & JENNIFER, F. (2011).** Reporting and discussing effect size: Still the road less traveled? *Practical Assessment, Research, and Evaluation*, 16(1), 1–12.
- MERK, S., POINDL, S., WURSTER, S., & BOHL, T. (2020).** Fostering aspects of pre-service teachers’ data literacy: Results of a randomized controlled trial. *Teaching and Teacher Education*, 91, Artikel 103043.

- MICHAL, A. L., & SHAH, P. (2024). A practical significance bias in laypeople's evaluation of scientific findings. *Psychological Science* 35(4), 315–327.
- MITCHELL, M. L., & JOLLEY, J. M. (2010). *Research design explained* (7. Aufl.). Wadsworth.
- NEUENSCHWANDER, M. P. (2005). Forschungskompetenzen in der Lehrerinnen- und Lehrerbildung erweitern: Ein Weiterbildungskonzept. *BzL - Beiträge zur Lehrerinnen- und Lehrerbildung*, 23(2), 270–280.
- OECD (HRSG.). (2023). *PISA 2022 Ergebnisse (Band I): Lernstände und Bildungsgerechtigkeit*. wbv Media.
- PELLEGRINI, M., & VIVANET, G. (2021). Evidence-based policies in education: Initiatives and challenges in europe. *ECNU Review of Education*, 4(1), 25–45.
- RBB (2023). *Deutsche Schülerinnen und Schüler schneiden bei neuer PISA-Studie so schlecht ab wie nie zuvor*. Abgerufen am 31.01.2025 von <https://www.tagesschau.de/multimedia/video/video-1280422.html>
- RENKL, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice? A plea for theories as primus inter pares. *Zeitschrift für Pädagogische Psychologie*, 36(4), 217–231.
- RYCROFT-SMITH, L., & STYLIANIDES, A. J. (2022). What makes a good educational research summary? A comparative judgement study of mathematics teachers' and mathematics education researchers' views. *Review of Education*, 10(1), Artikel e3338.
- SCHILDKAMP, K., HANDELZALTS, A., POORTMAN, C. L., LEUSINK, H., MEERDINK, M., SMIT, M., EBBELER, J., & HUBERS, M. D. (2018). *The data team™ procedure: A systematic approach to school improvement*. Springer.
- SCHMID, S., & LUTZ, A. (2007). Epistemologische Überzeugungen als kohärente Laientheorien. *Zeitschrift für pädagogische Psychologie*, 21(1), 29–40.
- SCHMIDT, K. (2024). *Teachers' engagement with educational science. How to communicate findings from educational science in a user-friendly way to teachers* [Dissertation, Pädagogische Hochschule Karlsruhe].
- SCHMIDT, K., EDELSBRUNNER, P. A., ROSMAN, T., CRAMER, C., & MERK, S. (2023). When perceived informativity is not enough. How teachers perceive and interpret statistical results of educational research. *Teaching and Teacher Education*, 130, Artikel 104134.
- SCHNEIDER, J., SCHMIDT, K., BOHRER, K., & MERK, S. (2025). Communicating Effect Sizes to Teachers: Exploring Different Visualizations and Their Enrichment Options. *Zeitschrift für Psychologie*, 233(1), 52–63.
- SHARPLES, J., & SHEARD, M. (2015). Developing an evidence-informed support service for schools – reflections on a UK model. *Evidence & Policy*, 11(4), 577–587.
- SHAVELSON, R. J., & TOWNE, L. (2002). *Scientific research in education*. National Academies Press.

- SLAVIN, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, 55(1), 21–31.
- STAN DEVELOPMENT TEAM. (2024). *Stan modeling language users guide and reference manual*. Abgerufen am 31.01.2025 von <https://mc-stan.org>
- STARK, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer Praxis. *Zeitschrift für Pädagogische Psychologie*, 31(2), 99–110.
- TAZ (2023). *Pisa-Schock für deutsche Schüler:innen: Im freien Fall*. taz.de. Abgerufen am 31.01.2025 von <https://taz.de/Pisa-Schock-fuer-deutsche-Schuelerinnen/!5974146/>
- THORNDIKE, E. L. (1904). *Theory of mental and social measurements*. The Science Press.
- VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B., & BÜRKNER, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2), 667–718.
- ZHANG, S., HECK, P. R., MEYER, M. N., CHABRIS, C. F., GOLDSTEIN, D. G., & HOFMAN, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, 120(33), Artikel e2302491120.