# Evidenz.Besser.Kommunizieren.: Wie Bildungswissenschaften und Fachdidaktiken ihre Wissenschaftskommunikation weiterentwickeln können.

Samuel Merk, Sarah Bez, and Kirstin Schmidt

Pädagogische Hochschule Karlsruhe

# Author Note

Samuel Merk  http://orcid.org/0000-0003-2594-5337

Die Daten dieses Artikels sind zu finden unter Die Autor:innen haben keine Interessenskonflikte zu berichten

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: *Samuel Merk***:** conceptualization, data curation, formal analysis, investigation, methodology, software, supervision, validation, visualization, writing inital draft, and writing – review & editing

Correspondence concerning this article should be addressed to Samuel Merk, Pädagogische Hochschule Karlsruhe, Bismarckstraße 10, Karlsruhe 76133, Germany, Email: merk@ph-karlsruhe.de

# Zusammenfassung

Lehrkräfte treffen tagtäglich unzählige Entscheidungen bzgl. ihrer Unterrichtsgestaltung und -entwicklung. Dabei rekurrieren Sie vornehmlich auf persönliche Erfahrung, Konzeptwissen oder Heuristiken. Evidenz aus Bildungswissenschaften und Fachdidaktiken wird das Potenzial zugeschrieben diese Entscheidungsprozesse komplementär zu informieren und zu objektivieren. Dazu ist es jedoch notwendig, dass die betroffenen Lehrkräfte diese Evidenz nicht fehlinterpretieren, was wiederum entsprechende Kompetenzen der Lehrkräfte oder besonders geschickte Wissenschaftkommunikation voraussetzt. Der vorliegende Beitrag untersucht daher die Möglichkeiten und Grenzen der Kommunikation von Effektstärken an Lehramtsstudierende am Beispiel des sog. zweiten PISA-Schocks. Im Ergebniss zeigt sich, dass Lehramtsstudierende Effektstärken sehr ungenau (Noise) ein- und im Mittel drastisch überschätzen (Practical Significance Bias). Dieser Bias konnte durch die Verwendung alternativer Visualisierungen deutlich eingedämmt werden

*Schlüsselwörter*: Lehrpersonenprofessionalisierung, Wissenschaftskommunikation, Practical Significance Bias

# Evidenz.Besser.Kommunizieren.: Wie Bildungswissenschaften und Fachdidaktiken ihre Wissenschaftskommunikation weiterentwickeln können.

Die bildungswissenschaftliche Literatur zu Schul- und Unterrichtsentwicklung bedient sich einer Vielzahl theoretischer Grundlegungen ([Bohl, 2020](#ref-bohl2020)) und blickt daher aus ganz verschiedenen Winkeln auf diesen Gegenstand: Neben eher systemtheoretischen Perspektiven ([K.-O. Bauer & Rolff, 1978](#ref-bauer1978)) finden sich u.a. Ansätze mit Entlehnungen aus der Lehr- Lern- ([Helmke, 2022](#ref-helmke2022)) und Organisationspsychologie ([Holtappels, 2007](#ref-holtappels2007)) oder mit Praxisorientierung als Leitgedanke ([Brügelmann, 2018](#ref-bruegelmann2018)). Datenbasierte Schulentwicklung hat im deutschsprachigen Raum erst in den vergangenen zwei Dekaden Momentum gefunden, wenngleich deren Grundidee des empirischen Einholens von Information über den Ist-Stand schon zuvor gefordert und auch umgesetzt wurde ([Altrichter & Rolff, 2006](#ref-altrichter2006)). 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](#ref-aero2023); [J. Bauer & Prenzel, 2012](#ref-bauer2012); [Council of the European Union, 2024](#ref-eurlex2024); [Pellegrini & Vivanet, 2021](#ref-pellegrini2021); [Slavin, 2020](#ref-slavin2020)). Da jedoch einerseits die Genese und Interpretation von Evidenz nicht zu den professionellen Kernkompetenzen von Lehrkäften gehört andererseits Bildungswissenschaftler- und Fachdidaktiker:innen keine Expert:innen für die Gestaltung von Unterricht sind, plädiert der vorliegende Beitrag dafür, Wissenschaftskommunikation erstens als dialogischen Prozess zwischen Bildungswissenschaften/Fachdidaktiken und Lehrkräften aufzufassen und zweitens diesen forschungsbasiert weiter zu entwickeln.

Daher führt der folgende theoretische Hintergrund zunächst in Konzepte und Begriffe evidenzinformierter Praxis ein, bevor er auf Wissenschaftskommunikation in Bildungswissenschaften und Fachdidaktiken eingeht, um abschließend ein empirisches Beispiel zu skizzieren.

# Theoretischer Hintergrund

## Evidenzinformiertes Handeln

### Was kann unter »Evidenz« verstanden werden?

Ethymologisch kann »Evidenz« als Substantivierung des Adjektivs »evident« gesehen werden ([Kluge, 2011, S. 263](#ref-kluge2011)), welches wiederum im 18. Jahrhundert dem lateinischen »evidens« (»ersichtlich, augenscheinlich«, [Hau et al., 2012](#ref-hau2012)) entlehnt wurde ([Stark, 2017](#ref-stark2017)). Allerdings meinen Bildungswissenschaftler:innen und Fachdidaktiker:innen gerade nicht »das Augenscheinliche« oder »das direkt Ersichtliche« wenn sie von Evidenz sprechen - vielmehr ist in Definitionsvorschlägen von »wissenschaftlichem Wissen« ([Stark, 2017](#ref-stark2017)), von einer »Funktion« von Daten für die Bestätigung oder Widerlegung von Hypothesen und Theorien ([Bromme et al., 2014](#ref-bromme2014b)) oder von »warrants for making assertions or knowledge claims« ([Shavelson & Towne, 2002](#ref-shavelson2002)) die Rede. In einer aktuellen Systematisierung verschiedener Verständnisse des Evidenz-Begriffs in den Bildungswissenschaften hebt Schmidt ([2024](#ref-schmidt2024)) 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](#ref-renkl2022)) gemeint sein dürften. Diese stellen eher Rahmenmodelle oder sogenannte »interventional models« (z.B. Cognitive Theory of Multi-Media Learning) dar (ebd.), die wiederum meist stark von empirischen Ergebnissen beeinflusst sind. Daher ist es plausibel »Theorien« die Funktion als »warrant« für »knowledge claims« zuzuschreiben - sie also auch als »Evidenz« zu bezeichnen.

### Evidenzbasiert, evidenzinformiert, evidenzorientiert.

Im vorigen 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, e datenbasiert, forschungsbasiert und forschungsinformiert klassisches *Jingle and Jangle* ([Kelly & Farrie, 2023](#ref-kelly2023); [Thorndike, 1904](#ref-thorndike1904)) darstellen - hier also unterschiedliche Begriffe für das Gleiche und gleiche Begriffe für Unterschiedliches gebraucht werden. Die Differenzen zwischen evidenz**basiert** und evidenz**informiert** sowie evidenz**orientiert** können jedoch auch bedeutsam interpretiert werden: Da mit »Evidenz**basierung**« oft »the medical model« ([Jones, 2024](#ref-jones2024)) und damit Evidenz aus *Kontrollgruppenexperimenten* als *notwendige Voraussetzung* für eine Entscheidung assoziiert wird, zieht dieser Begriff die stärkste Kritik auf sich ([Bellmann & Müller, 2011](#ref-bellmann2011); [Biesta, 2007](#ref-biesta2007)). Den Begriffen »evidenz**orientiert**« und »evidenz**informiert**« wird mit weniger Fundamentalkritik begegnet, da diese schon rein sprachlich eher eine heuristische denn eine rechenschaftslegende Rolle impliziert.

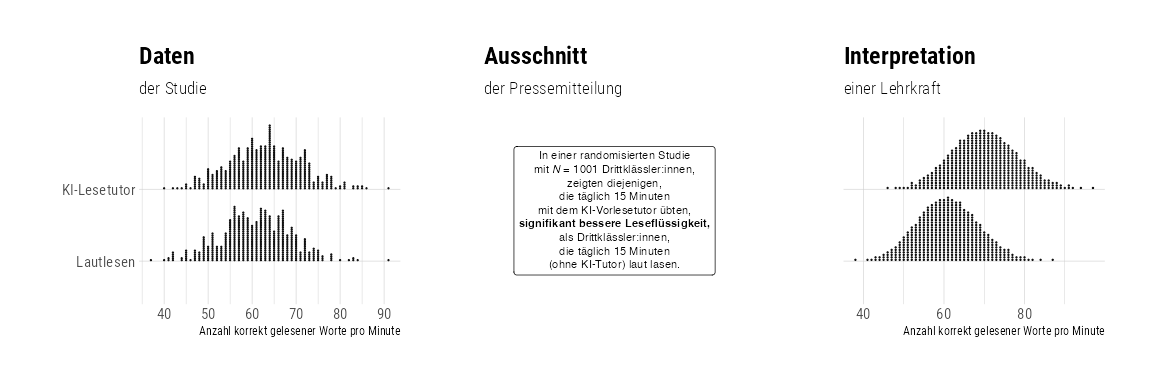
In der deutschsprachigen bildungswissenschaftlichen Diskussion sind nach Bromme et al.([2014](#ref-bromme2014b)) zunächst zwei verschiedene Diskussionsstränge bzgl. evidenzinformierter Entscheidungen im Bildungskontext unterscheidbar: Ein Diskussionsstrang beschäftigt sich mit evidenzinformierten Entscheidungen 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](#ref-bromme2014b)) etwa sprechen davon, dass Evidenz über Zustände informieren, Mechanismen erklären oder Interventionen evaluieren kann. Groß Ophoff et al. ([Groß Ophoff et al., 2023](#ref-großophoff2023)) wiederum unterscheiden konzeptuelle Nutzung (»*evidence allows focussing attention, provides new insights, challenges beliefs or reframes thinking*«, S. 02), instrumentelle Nutzung (»*identify or develop concrete measures to be taken*«, S. 02) und symbolische Nutzung (»*justif[y] or support of existing positions or established procedures*«, S. 02).

## Potenzielle Wege zu einer gelingenden Wissenschaftskommunikation

Unabhängig vom Kontext und der Funktion evidenzinformierter Entscheidungen ist es plausibel anzunehmen, dass eine erfolgreiche Kommunikation von Evidenz zwischen Bildungswissenschaftler:innen/Fachdidaktiker:innen und den Akteuren im Bildungssystem 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.

Abbildung 1

Daten einer (fiktiven) Studie, eine dazugehörige Pressemitteilung und die Vorstellung einer Lehkraft von den Daten



Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

Liest eine Lehrkraft etwa die (fiktive) Pressemitteilung in [Abbildung 1](#fig-AbbildungMoMa), stellt sich die Ergebnisse wie in [Abbildung 1](#fig-AbbildungMoMa) rechts vor ([Schmidt et al., 2023](#ref-schmidt2023)) 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 obwohl etwa die Implementation von Lesetandems kostengünstiger, weniger zeitaufwändig und lernwirksam gewesen wäre.

Die Forschung zur Wissenschaftskommunikation hat eine Reihe solcher potenziellen Problematiken aufgezeigt: Z.B. das soeben beschriebene Verwechseln von Inferenzstatistik und Effektstärke ([Schmidt et al., 2023](#ref-schmidt2023)), das automatische Annehmen starker Effekte, wenn keine Effektstärken berichtet wurden (Practical Significance Bias, [Michal & Shah, 2024](#ref-michal2024)), Rückschaufehler ([Masnick & Zimmerman, 2009](#ref-masnick2009)) oder die verzerrte Einschätzung der Belastbarkeit von Befunden (z.B. 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](#ref-bohrer2025)).

Gleichzeitig liegt eine Reihe von Befunden vor, die implizieren, dass verbesserte Kommunikation von Evidenz an Lehrkräfte zu Zwecken evidenzinformierten Handelns vergleichsweise einfach umsetzbar ist. 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 und Statistical Literacy der Lehrkräfte ansetzen.

|  |
| --- |
| Vorsicht |
| Gibt es diese Unterscheidung auch in der Literatur oder nur in unseren Gesprächen?  Antwort: Einen ersten Ansatz findest du bei Brühwiler et al (2020). Das Modell ist zwar sehr umfassen und m.E. auch nicht ideal, aber als Referenz ein erster guter Ansatz. 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. https://doi.org/10.36950/bzl.38.2020.9309  man könnte im weiteren Sinne auch auf Debiasing-Forschung verweisen, aber glaube die Referenz von Brühwiler passt am besten |

Zu zweiterem gehören Programme wie »Data Teams« ([Schildkamp & Poortman, 2015](#ref-schildkamp2015)), welche durch ein umfängliches Set an vordefinierten Leitlinien und Aktivitäten versucht, konkrete schulische Probleme mit Hilfe von (oft eigens dafür genierten) Daten zu lösen, wobei meist 4-6 Lehrkräfte und Schulleiter:innen mit Bildungswissenschaftler:innen und Fachdidaktiker:innen kooperieren. Auch kurz- ([Merk et al., 2020](#ref-merk2020)) oder längerfristig angelegte ([Karst et al., 2024](#ref-karst2024)) Interventionen zur Anbahnung notwendiger Kompetenzen für evidenzinformiertes Handeln wie Graph Literacy ([Friel et al., 2001](#ref-friel2001)) oder Forschungskompetenz ([Neuenschwander, 2005](#ref-neuenschwander2005)) sowie die konkrete Unterstützung für solches ([Academy, 2025](#ref-zotero-8935)), können diesem Ansatz zugerechnet werden.

Angebotsseitige Versuche die Kommunikation von Evidenz zu verbessern, stammen aus verschiedensten sozialwissenschaftlichen Disziplinen: So wird z.B. in der Psychologie untersucht, welche algebraisch äquivalenten Formulierungen zu standardisierten Effektstärken bei Rezeption durch Laien adäquatere Vorstellungen induzieren [siehe 1](#tbl-wisskommbsp). 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](#ref-hullman2015); [Zhang et al., 2023](#ref-zhang2023)) und die bildungswissenschaftliche Lehrerbildungsforschung sowie die Fachdidaktiken erproben innovative Formate für die Zielgruppe der Lehrkräfte (z.B. [Schneider et al., 2024](#ref-schneider2024)), was auch das Anliegen der vorliegenden Studie ist.

Tabelle 1

Beispiele für angebotsseitige Versuche verbesserter Kommunikation von Evidenz

|  | Unterschied | Zusammenhang |
| --- | --- | --- |
| Standard-kommunikation | Die Leseleistung von Schülerinnen und Schülern in (PISA2022) sank um 28 Punkte und damit auf den Tiefststand. | Der sozioökonomische Status klärt 14% der Varianz der Mathematikleistung auf. |
| Verbesserte Kommunikation | Die Leseleistungen von Schülerinnen und Schülern in Deutschland aus PISA2018 und aus PISA2022 überlappen sich zu 88,9% wobei der Mittelwert um 28 Punkte sank. | Von 100 Schülerinnen und Schülern, die einen überdurchschnittlichen sozioökonomischen Status haben, zeigen 69 eine überdurchschnittliche Mathematikleistung. |

Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

# Die vorliegende Studie

In diesem Kontext untersucht die vorliegende Studie, inwiefern verbreitete Standardgrafiken zur Kommunikation der Entwicklung der Lesekompetenz in den deutschen Kohorten des Programme of International Student Assessment (PISA) Practical Significance Bias induzieren und ob dieser mit Grafiken eingedämmt werden kann, bei deren Gestaltung theoretische und empirische Erkenntnisse der Wissenschaftkommunikation [siehe Abschnitt 2.1.1](#sec-materialien) berücksichtigt wurden.

## Methode

### Materialien

In der Berichterstattung zu den Ergebnissen der PISA2022-Kohorte wurden durch journalistische Medien zahlreiche Darstellungsformate gewählt, insbesondere Liniendiagramme [siehe Tabelle 2](#tbl-pisalinegraphs), was angesichts der Anlage des PISA als Trendstudie ([Döring & Bortz, 2016](#ref-döring2016)) konsequent erscheint.

Tabelle 1

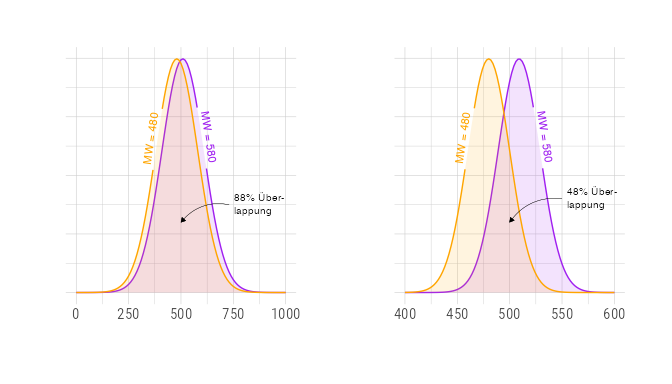
Verwendete Liniendiagramme in der Berichterstattung.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Süddeutsche Zeitung ([2023](#ref-volkert2023)) | RBB ([2023](#ref-rbb2023)) | taz ([2023](#ref-taz.de2023)) |

Diese Abbildungen erlauben einen effizienten Vergleich der Mittelwerte sowohl über die Zeit als auch Variablen (hier: Fächer) hinweg. In solchen Grafiken ist jedoch die Bedeutsamkeit der Mittelwertsdifferenz nur bei bekannter Streuung interpretierbar: [Abbildung 2](#fig-mwdiffstreuung) zeigt jeweils die gleichen Mittelwertsdifferenzen von 508 (PISA Lesen 2015) und 480 (PISA Lesen 2022).

Abbildung 2

Illustration der Unabhängigkeit von Mittelwertsdifferenz und Größe des Effekts



Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

Das Ausmaß der Bedeutsamkeit dieses (gleichen) Mittelwertsunterschiedes entsteht aber erst durch die Streuung der Daten um diesen Mittelwert herum. Dadurch dass die Variablen im rechten Teil der Abbildung kaum streuen, ist die Überlappung der beiden Gruppen gering (48%, großer Effekt), während die große Überlappung im linken Teil (88%, kleiner Effekt) durch die große Streuung zustande kommt. Die Abbildungen in [Tabelle 2](#tbl-pisalinegraphs) 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](#ref-kale2020)).

Tabelle 2

Verwendete Stimuli

|  |  |
| --- | --- |
|  |  |

Daher wurden vorliegend neben Liniendiagrammen auch überlappende Verteilungskurven verwendet. Um diese barriereärmer zu gestalten wurde bei der Farbgebung auf hinreichenden Kontrast bei den prävalenten Sehbehinderungen geachtet ([Garnier et al., 2023](#ref-garnier2023)). Um unnötige Arbeitsgedächtnisbelastung zu vermeiden wurde die Legende direkt in die Grafik integriert ([Franconeri et al., 2021](#ref-franconeri2021)).

### Design, Stichprobe und Instrument

Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

In einem Between-Person Design wurde *N* = 216 Studierenden in Bachelorstudiengängen des Primar- und Sekundarstufenlehramtes randomisiert entweder der Line-Graph oder die überlappende Verteilungskurve (vgl. [Tabelle 3](#tbl-materials) ) gezeigt. Anschließend wurden sie mit folgenden 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 abschneidet 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](#ref-brooks2014); [Kim et al., 2022](#ref-kim2022)), wenngleich die Operationalisierung als Schieberegeler unklar lässt, inwiefern bei der Beantwortung tatsächlich eine Elaboration der Überlappung vorgenommen wird oder die Teilnehmenden eher 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 3](#fig-plotresults)). Um diesen Umständen in der inferenzstatistischen Modellierung Rechnung zu tragen, wurden Bayesianische Mixture Regressionsmodelle für zwei trunkierte Normalverteilungen ([Frischkorn & Popov, o. J.](#ref-frischkorn2023)) in der probabilistischen Sprache Stan ([Stan Development Team, 2024](#ref-standevelopmentteam2024)) mithilfe des R-Pakets {brms} ([Bürkner, 2017](#ref-bürkner2017)) geschätzt.

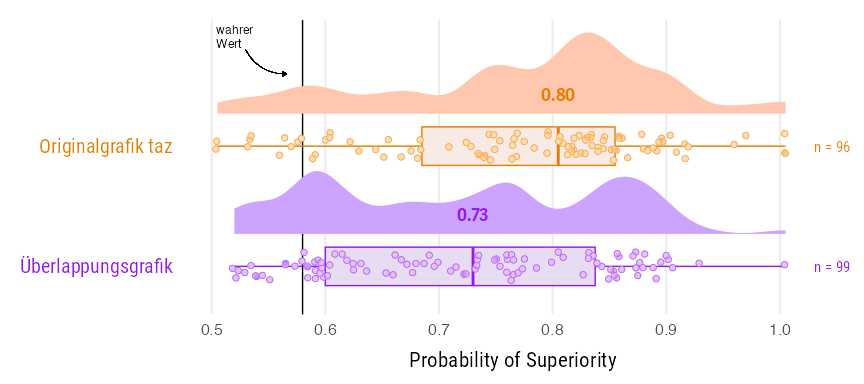
Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

### Ergebnisse

Die Inspektion des Marcov-Chain-Monte-Carlo-Sampling-Prozesses zeigte eine zufriedenstellende Qualität bzgl. der Konvergenz und effektiver Sampling Size (, [Vehtari et al., 2021](#ref-vehtari2021prefix)).

Abbildung 3

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



Quelle: [Artikel-Notizbuch](https://sammerk.github.io/StudienergebnisseBesserKommunizieren/index.qmd.html)

Die Medianeinschätzung der Probability of Superiority lag in beiden nGruppe deutlich über dem wahren Wert (Originalgrafik .80, Überlappungsgrafik .73). Dieser Unterschied in der Einschätzung entspricht einer Überlappung von 81.71% (Cliff’s *d* = 0) 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 vorliegenden Daten sind 14,8-fach wahrscheinlicher unter der Annahme, dass der Mittelwert in der geschätzten Effektstärke in der Gruppe mit der Überlappungsgrafik niedriger ist als unter der Annahme, dass beide gleich sind.

# Diskussion

## Literatur

Academy, C. H. U. (2025). *Clearing House Unterricht Academy*. <https://clearinghouse-academy.de/>

AERO. (2023). *Evidence-based teaching practices*.

Altrichter, H., & Rolff, H.-G. (2006). Datenbasierte Schulentwicklung. Editorial. *Journal für Schulentwicklung*, *10*(4), 4–6.

Bauer, J., & Prenzel, M. (2012). European teacher training reforms. *Science*, *336*(6089), 1642–1643. <https://doi.org/10.1126/science.1218387>

Bauer, K.-O., & Rolff, H.-G. (1978). *Vorarbeiten zu einer Theorie der Schulentwicklung* (K.-O. Bauer & H.-G. Rolff, Hrsg.; S. 219–263). Beltz.

Bellmann, J., & Müller, T. (2011). *Evidenzbasierte Pädagogik ein Déjà-vu?* (J. Bellmann & T. Müller, Hrsg.; S. 9–32). VS Verlag für Sozialwissenschaften. <https://doi.org/10.1007/978-3-531-93296-5_1>

Biesta, G. (2007). Why “What Works” Won’t Work: Evidence-Based Practice and the Democratic Deficit in Educational Research. *Educational Theory*, *57*(1), 1–22. <https://doi.org/10.1111/j.1741-5446.2006.00241.x>

Bohl, T. (2020). *Theorien der Schulentwicklung* (M. Harant, P. Thomas, & U. Küchler, Hrsg.; S. 97–109). Tübingen University Press. <https://doi.org/10.15496/publikation-45627>

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. In *Zeitschrift für Erziehungswissenschaft* (Bd. 17). <https://doi.org/10.1007/s11618-014-0514-5>

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. <https://doi.org/10.1037/a0034745>

Brügelmann, H. (2018). *Unterrichts- und Schulentwicklung in Communities of Practice* (H. Barz, Hrsg.; S. 479–484). Springer Fachmedien. <https://doi.org/10.1007/978-3-658-07491-3_44>

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). <https://doi.org/10.18637/jss.v080.i01>

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*. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C_202403642>

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5., vollst). Springer. <http://dx.doi.org/10.1007/978-3-642-41089-5>

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. <https://doi.org/10.1177/15291006211051956>

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. <https://doi.org/10.2307/749671>

Frischkorn, G. T., & Popov, V. (o. J.). *A Tutorial for Estimating Bayesian hierarchical mixture models for visual working memory tasks: Introducing the Bayesian Measurement Modeling (bmm) package for R*. <https://doi.org/10.31234/osf.io/umt57>

Garnier, S., Ross, N., BoB Rudis, Filipovic-Pierucci, A., Galili, T., Timelyportfolio, O’Callaghan, A., Greenwell, B., Sievert, C., Harris, D. J., Sciaini, M., & JJ Chen. (2023). *sjmgarnier/viridis: CRAN release v0.6.3*. Zenodo. <https://doi.org/10.5281/ZENODO.4679423>

Groß 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. <https://doi.org/10.3389/feduc.2022.1011241>

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. Innovationstheoretische Reflexionen und Forschungsbefunde über Steuergruppen.* (N. Berkemeyer, Hrsg.; 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. <https://doi.org/10.1371/journal.pone.0142444>

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. <https://doi.org/10.29329/ijer.2024.1041.3>

Kale, A., Kay, M., & Hullman, J. (2020). *Visual reasoning strategies for effect size judgments and decisions*. <https://doi.org/10.48550/ARXIV.2007.14516>

Karst, K., Yendell, O., Marx, A., Lettau, W.-D., & Hawlitschek, P. (2024). *Die Etablierung von Evidenzteams in SchuMaS Eine Strategie zur systematischen Nutzung von Daten für die Schul- und Unterrichtsentwicklung* (K. Maaz & A. Marx, Hrsg.; S. 225–240). Waxmann. <https://madoc.bib.uni-mannheim.de/67727/>

Kelly, M. G., & Farrie, D. (2023). Misrepresented Funding Gaps in Data for Some States. *Educational Researcher*, *52*(4), 244–247. <https://doi.org/10.3102/0013189X221133396>

Kim, Y.-S., Hofman, J. M., & Goldstein, D. G. (2022). *CHI ’22: CHI Conference on Human Factors in Computing Systems*. 1–14. <https://doi.org/10.1145/3491102.3502053>

Kluge, F. (2011). *Etymologisches Wörterbuch der deutschen Sprache* (25. Aufl.). De Gruyter.

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. <https://doi.org/10.1891/1939-7054.2.1.29>

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*, 103043. <https://doi.org/10.1016/j.tate.2020.103043>

Michal, A. L., & Shah, P. (2024). A Practical Significance Bias in Laypeople’s Evaluation of Scientific Findings. *Psychological Science*, 09567976241231506. <https://doi.org/10.1177/09567976241231506>

Neuenschwander, M. P. (2005). Forschungskompetenzen in der Lehrerinnen- und Lehrerbildung erweitern: Ein Weiterbildungskonzept. *BzL - Beiträge zur Lehrerinnen- und Lehrerbildung*, *23*(2), 270–280. <https://doi.org/10.36950/bzl.23.2.2005.10132>

Pellegrini, M., & Vivanet, G. (2021). Evidence-Based Policies in Education: Initiatives and Challenges in Europe. *ECNU Review of Education*, *4*(1), 2545. <https://doi.org/10.1177/2096531120924670>

RBB, M. K. (2023). *Deutsche Schülerinnen und Schüler schneiden bei neuer PISA-Studie so schlecht ab wie nie zuvor*. <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. <https://doi.org/10.1024/1010-0652/a000345>

Schildkamp, K., & Poortman, C. L. (2015). Factors influencing the functioning of data teams. *Teachers College record*, *117*(4), 1–42. <https://doi.org/10.1080/09243453.2016.1256901>

Schmidt, K. (2024). *Teachers’ Engagement With Educational Science How to Communicate Findings From Educational Science in a User-Friendly Way to Teachers* [Phdthesis].

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*, 104134. <https://doi.org/10.1016/j.tate.2023.104134>

Schneider, J., Schmidt, K., Bohrer, K., & Merk, S. (2024). Communicating Effect Sizes to Teachers. *Zeitschrift Für Psychologie*. <https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000573>

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. <https://doi.org/10.1080/00461520.2019.1611432>

Stan Development Team. (2024). *Stan Modeling Language Users Guide and Reference Manual*. <https://mc-stan.org>

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer Praxis. *Zeitschrift für Pädagogische Psychologie*, *31*(2), 99–110. <https://doi.org/10.1024/1010-0652/a000201>

taz.de. (2023). *Pisa-Schock für deutsche Schü­le­r:in­nen: Im freien Fall | taz.de*. <https://taz.de/Pisa-Schock-fuer-deutsche-Schuelerinnen/!5974146/>

Thorndike, E. L. (1904). *Theory of mental and social measurements.* The Science Press. <https://doi.org/10.1037/13283-000>

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). <https://doi.org/10.1214/20-BA1221>

Volkert, L., & Germany, S. de G., Munich. (2023). *Jeder dritte 15-Jährige scheitert an leichten Mathe-Aufgaben*. <https://www.sueddeutsche.de/projekte/artikel/politik/pisa-studie-schulen-mathe-aufgaben-jugendliche-scheitern-gruende-e206264/>

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), e2302491120. <https://doi.org/10.1073/pnas.2302491120>