Forschungsmonitor Schule
Rezensionen für die Praxis

Sonja Hensel

Rezension zu

Liu, X., Guo, B., He, W. & Hu, X. (2025). Effects of Generative Artificial Intelligence on K-12 and Higher Education Students' Learn-ing Outcomes: A Meta-Analysis. *Journal of Educational Computing Research, 63*(5), 1249–1291.

Kommentierter Kurzbefund

Die Forschungsgruppe um Xiaohong Liu untersucht in einer Metaanalyse die Effekte des Einsatzes generativer künstlicher Intelligenz (GenAI) wie ChatGPT auf die Lernergebnisse von Schülerinnen und Schülern unterschiedlichen Alters sowie von Studierenden. Bei den Lernergebnissen unterscheiden sie zwischen Lernleistungen und Lernmotivation.

Zur Auswahl der Studien führten die Autorinnen und Autoren eine umfassende Recherche in einschlägigen Datenbanken durch und grenzten das Gefundene auf Basis von Kriterien ein. Dazu gehörten das Vorhandensein einer Experimental- und einer Kontrollgruppe sowie die Angabe von statistischen Daten, die das Berechnen von Effektstärken erlauben. Letztlich wurden 49 Studien in Bezug auf 2 Forschungsfragen ausgewertet:

1. Welche Effekte hat die Nutzung von generativer KI auf die Lernergebnisse von Schülerinnen und Schülern sowie Studierenden?
2. Welche Moderatorvariablen beeinflussen die Wirkung?

Als mögliche Moderatorvariablen berücksichtigten sie den Bildungsbereich (Schule vs. Hochschule), die Fächergruppe (Geisteswissenschaften / Künste, Informatik / Kommunikationstechnologie, Pädagogik etc.), die Art der genutzten KI-Benutzeroberfläche (u. a. App vs. Browser), den Interaktionsmodus mit der KI (nur Text oder zusätzlich Audio, Bilder etc.) und die Dauer der Intervention (5 Stufen von < *1 Woche* bis > *15 Wochen*).

Liu et al. finden insgesamt mittlere bis große Effektstärken für den Einsatz von generativer KI auf Lernleistungen (g = 0.86) und Lernmotivation (g = 0.80). Als einflussreiche Moderatorvariablen erweisen sich die Fächergruppe, die KI-Benutzeroberfläche und die Dauer der Interaktion: Bei einigen Fächergruppen, für die jedoch nur wenige Studien vorlagen, ist kein Effekt auf Lernleistungen sicher belegbar. Der Einsatz von Apps wirkt im Schnitt lernförderlicher als die Nutzung im Browser. Während Lernmotivation stärker durch kurze Interventionen gefördert wird (< 5 Wochen), profitieren Lernleistungen am meisten bei einer Dauer von 5–15 Wochen.

Die Ergebnisse sind für Lehrkräfte zum einen interessant, weil vom Einsatz generativer KI in vielen Fällen positive Effekte zu erwarten sind. Zum anderen wird die Notwendigkeit deutlich, die Einsatzmodalitäten für generative KI gezielt zu gestalten, da diese die Wirksamkeit beeinflussen können. Dies erfordert in erster Linie grundlegendes Know-how der Lehrkräfte, denn die Studie liefert nur wenige Hinweise auf Gelingensbedingungen. Zudem ist die Datenbasis zu den Moderatorvariablen teilweise klein und angesichts der fortschreitenden technologischen Entwicklungen bleibt offen, welche Befunde zukünftig Bestand haben werden.

Hintergrund

Nach Ansicht der Forschungsgruppe um Xiaohong Liu markierte der Start von ChatGPT im Jahr 2022 einen Meilenstein in der Entwicklung von künstlicher Intelligenz (KI). Dieser populäre KI-Chatbot steht stellvertretend für Anwendungen der generativen KI, d. h. für KI-Systeme, die neue Inhalte (Text, Bilder, Audio, Codes, Video usw.) erzeugen können, statt nur vorhandene Daten zu klassifizieren oder zu erkennen. Es folgte eine Fülle weiterer kostenfreier und datenschutzkonformer Angebote, sodass der Einsatz von generativer KI auch in Bildungskontexten möglich wurde.

Inzwischen liegen zu den Wirkungen von KI auf Lernleistungen und Lernmotivation eine Fülle von Studien vor, die jeweils spezielle Einsatzgebiete beleuchten. Daneben erschienen bereits einige Metaanalysen, die vorhandene Studien aggregieren und systematisieren (z. B. Zheng, Niu, Zhong & Gyasi, 2023 oder Deng, Jiang, Yu, Lu & Liu, 2025).

Liu et al. erweitern die Befundlage, indem sie zum einen zwei Bildungsbereiche (Schule und Hochschule) und zum anderen Einflüsse auf verschiedene Lernergebnisse (kognitive Lernleistung, motivationale Merkmale, z. B. Einstellung zum Lernen, Selbstwirksamkeit) einschließen.

Konkret fragen die Forschenden:

1. Welche Effekte hat die Nutzung von generativer KI durch Schülerinnen und Schüler sowie Studierende auf ihre Lernergebnisse?
2. Welche Moderatorvariablen beeinflussen die Wirkung von generativer KI auf die Lernergebnisse von Schülerinnen und Schülern sowie Studierenden?

Design

Die Forschenden nutzten für ihre Studie die Datenbanken Web of Science (WOS) und Scopus und durchsuchten diese im August 2024 nach einer Kombination aus Suchwörtern aus den Bereichen „generative KI“, „Lernergebnisse“ und „experimentelle Studien im Bildungsbereich“ („educational experiments“). Dies führte zu 3 760 Treffern. Ergänzend wurde ein „Schneeball-Suchsystem“ angewendet, bei dem die Literaturverzeichnisse bereits gefundener Studien zum Ausgangspunkt neuer Recherchen wurden. Dies brachte weitere 122 Treffer.

Aus den so gefundenen 3 822 Artikeln in Fachzeitschriften wurden Studien ausgeschlossen, die nicht den im Vorhinein definierten Kriterien genügten – beispielsweise doppelte Texte, Studien ohne Kontrollgruppen und Studien, die nicht die nötigen statistischen Angaben enthielten, um Effektstärken zu berechnen. Schließlich konnten 49 Studien analysiert werden.

Für die Auswertung bezüglich möglicher Moderatorvariablen bedienten sich die Forschenden eines Kodierschemas, das aus 6 Variablen sowie deren möglichen Ausprägungen bestand: der Bildungsbereich (Primar- und Sekundarbereich vs. Hochschule), die Fächergruppe (angelehnt an die ISCED-F 2013-Klassifikation der UNESCO), die Schnittstelle mit der KI (App vs. Browser), die Art der genutzten KI-Anwendung (Nutzung eines bereits verfügbaren Tools vs. selbsterstellte Anwendung), der Interaktionsmodus mit der KI (nur Text oder zusätzlich Audio, Bilder etc.) und die Dauer der Intervention (5 Stufen von < *1 Woche* bis > *15 Wochen*).

Zur Kategorisierung der Lernergebnisse griffen die Forschenden auf die Bloom'sche Lernzieltaxonomie zurück, die grundsätzlich 3 Domänen umfasst: die kognitive, affektive und verhaltensbezogene (Bloom, 1956). Darauf bezogen identifizierten sie in der Forschung zu Effekten von generativen KI-Anwendungen im Bildungskontext sechs Ergebnisaspekte: kognitive Lernleistungen, Lernmotivation, Selbstwirksamkeit, Langzeitspeicherung von Wissen, kritisches Denken und Ängste im Lernprozess. Die drei letztgenannten Aspekte konnten letztlich nicht in die Auswertung einbezogen werden, da nicht genügend Studien vorlagen. In die Kategorie Lernmotivation wurde die Selbstwirksamkeit integriert, sodass für die Auswertung die Lernergebnisse in (kognitive) Lernleistung und Lernmotivation differenziert wurden.

Der Kodierungsprozess wurden von zwei Personen durchgeführt. Inkonsistenzen wurden diskutiert, bis Einvernehmen hergestellt war.

Es wurde eine umfassende Metaanalyse mit Hilfe des Programms Comprehensive Meta-Analysis 3.0 (CMA 3.0) durchgeführt. Die Effekte von GenAI auf die Lernergebnisse wurden anhand standardisierter Mittelwertunterschiede (SMD) mit einem 95 %-Konfidenzintervall (KI) berechnet. Die Effektstärken, d. h. die Größe des Effekts zwischen zwei Gruppen, wurden als Hedges' g angegeben. Gemäß den von Cohen (1988) festgelegten Schwellenwerten liegen kleine Effekte ab einem Wert von 0.2, mittlere Effekte ab 0.5 und große Effekte ab 0.8.

Für die Auswahl der Berechnungsmodelle wurde mithilfe von Cochran's Q-Test und der I^2-Statistik überprüft, ob die Ergebnisse der Studien ähnlich oder heterogen ausfielen. Die Analyse ergab, dass die Ergebnisse der Studien sowohl zu Lernleistungen als auch zu Lernmotivation stark variierten, weshalb Modelle mit zufälligen Effekten (random-effects models) verwendet wurden.

Um die Zuverlässigkeit der Ergebnisse der Metaanalyse sicherzustellen, wurden weitere statistische Analysen durchgeführt (Bewertung des Publikationsbias durch Methoden wie Trichterdiagramme, Egger's Test, Begg's Test, Trim- und Fill-Methode). Diese Analysen erfolgen, um zu prüfen, ob möglicherweise vor allem Studien genutzt wurden, in denen signifikante und positive Ergebnisse berichtet und nicht signifikante bzw. negative Ergebnisse außer Acht gelassen wurden.

Die Ergebnisse dieser Analysen zeigen in Bezug auf Lernergebnisse und Lernmotivation, dass es nur minimale Publikationsverzerrungen gibt, d. h., dass die Ergebnisse der Metaanalyse als gültig angesehen werden können. Sensitivitätsanalysen bestätigen die Robustheit der Ergebnisse, d. h., dass es nicht zu Verzerrungen durch einzelne Studien gekommen ist.

Ergebnisse

Die Auswertungen ergeben, dass der Einsatz von generativer KI positive Effekte auf Lernleistungen und Lernmotivation der Lernenden hat. Die Effekte sind sowohl in Bezug auf Lernleistungen (g = 0.857, 95 % KI [0.601, 1.113], $p < .05$) als auch auf Lernmotivation (g = 0.803, 95 % KI [0.445, 1.161], $p <. 05$) substanziell und liegen unter Berücksichtigung der Konfidenzintervalle im mittleren bis hohen Bereich.

Ob die Lernenden eine Schule oder eine Hochschule besuchen, erweist sich deskriptiv als einflussreiche Moderatorvariable, wobei die Unterschiede nicht signifikant sind. Auf Lernleistungen von Studierenden hat die Verwendung von KI einen großen Effekt (g = 0.959, $p < .05$, n = 39), während der Effekt auf Schülerinnen und Schüler im mittleren Bereich liegt (g = 0.495, $p < .05$, n = 10). Bei der Lernmotivation

ist es umgekehrt: ein sehr großer Effekt der Nutzung von KI zeigt sich bei Schülerinnen und Schülern ($g = 1.651$, $p < .05$, $n = 10$) gegenüber einem moderaten bei Studierenden ($g = 0.642$, $p < .05$, $n = 39$).

In welcher Fächergruppe KI zum Einsatz kommt, wirkt als Moderator auf die gefundenen Effektstärken im Bereich der Lernleistungen ($Q = 17.918$, $p < .05$). Signifikant sind die Effektstärken in den Bereichen Geisteswissenschaften/Künste ($g = 1.271$, $p < .05$, $n = 12$), Informatik/Kommunikationstechnologien ($g = 0.814$, $p < .05$, $n = 15$) und Pädagogik ($g = 0.926$, $p < .05$, $n = 8$). Auch bei der Lernmotivation unterscheiden sich die Effektstärken nach Fächergruppen ($Q = 75.342$, $p < .05$). Für Informatik/Kommunikationstechnologien ergibt sich kein signifikanter Effekt auf Lernmotivation, in den anderen Gruppen hingegen schon: Gesundheit/Sozialwesen ($g = 0.740$, $p < .05$, $n = 7$), Geisteswissenschaften/Künste ($g = 0.747$, $p < .05$, $n = 7$) und Pädagogik ($g = 0.442$, $p < .05$, $n = 9$).

Die Schnittstelle, über die die Lernenden mit der KI interagieren, moderiert ebenfalls den Effekt in Bezug auf Lernleistungen ($Q = 17.139$, $p < .05$), aber nicht bei der Motivation ($Q = 0.277$, $p > .05$). Wenn Lernende z. B. eine App (application-based interface) verwenden, gibt es einen größeren positiven Effekt sowohl auf Lernleistungen ($g = 1.107$, $p < .05$, $n = 37$) als auch auf Lernmotivation ($g = 0.869$, $p < .05$, $n = 20$), als wenn die KI webbasiert genutzt wird. Hier zeigt sich dann ein (nicht signifikanter) mäßiger Effekt für Lernmotivation ($g = 0.595$, $p > .05$, $n = 10$) und ein kleiner Effekt für Lernleistungen ($g = 0.259$, $p < .05$, $n = 12$).

Des Weiteren zeigen die statistischen Auswertungen, dass eigens kreierte KI-Umgebungen einen großen positiven Effekt auf die Lernmotivation ($g = 1.029$, $p < .05$, $n = 10$) und einen moderaten auf die Lernleistung ($g = 0.614$, $p < .05$, $n = 19$) haben. Bei der Nutzung bereits vorhandener KI-Tools sind die Effekte tendenziell umgekehrt, d. h. moderat in Bezug auf die Lernmotivation ($g = 0.678$, $p < .05$, $n = 20$) und stark in Bezug auf die Lernleistung ($g = 1.028$, $p < .05$, $n = 30$), allerdings sind die Unterschiede nicht signifikant.

Beide untersuchten Interaktionsarten mit der KI (nur Text vs. gemischt – also Text, Bild, Audio etc.) beeinflussen die Lernergebnisse positiv, aber die Interaktion nur über Texte zeigt sowohl bei der Lernleistung ($g = 1.033$, $p < .05$, $n = 11$ gegenüber $g = 0.816$, $p < .05$, $n = 38$) als auch bei der Motivation ($g = 2.086$, $p < .05$, $n = 4$ vs. $g = 0.621$, $p < .05$, $n = 26$) tendenziell stärkere Effekte als die mit unterschiedlichen Medien, allerdings sind die Unterschiede nicht signifikant.

Schließlich moderiert die Länge des KI-Einsatzes die Effektstärke in signifikantem Maße (Lernleistung: $Q = 11.936$, $p < .05$; Lernmotivation: $Q = 79.398$, $p < .05$): Lernleistungen verbessern sich deutlich, wenn die Intervention länger als 5 Wochen dauerte ($g = 1.892$, $p < .05$, $n = 9$), wobei die Effektstärke von Interventionen, die länger als 15 Wochen dauerten, abnimmt ($g = 0.413$, $p < .05$, $n = 3$). In Bezug auf Lernmotivation ist der Einfluss signifikant positiv, wenn die Intervention länger als 1 Woche dauerte ($g = 0.821$, $p < .05$, $n = 7$), aber nimmt ab, wenn sie über 5 Wochen hinausging ($g = 0.697$, $p < .05$, $n = 9$).

Insgesamt betrachtet kann der Einsatz von KI auf Lernergebnisse als positiv angesehen werden.

Diskussion und Einschätzung

Zum Hintergrund

Die Studie bezieht sich auf die zum Datum ihrer Entstehung vorhandenen Metaanalysen zum Einsatz von KI im Bildungsbereich. Liu et al. leiten daraus Desiderate in Bezug auf Effekte des Einsatzes von generativer KI auf Lernleistungen und Lernmotivation ab, wobei sie sowohl Schülerinnen und Schüler als auch Studierende in den Blick nehmen.

Zum Design

Die Studie stützt sich auf eine intensive Recherche vorhandener Forschungsarbeiten und wählt die Arbeiten für ihre Metaanalyse nach nachvollziehbaren Kriterien aus. Die Kodierung anhand von Moderatorvariablen und deren Ausprägungen werden transparent gemacht. Es werden Studien bis einschließlich 2024 berücksichtigt. Viele der Studien kommen nicht aus dem deutschsprachigen Raum, was die Frage nach der Übertragbarkeit der Ergebnisse auf deutsche Bildungseinrichtungen aufwirft. Zudem liegen zu den verschiedenen Ausprägungen der Moderatorvariablen teilweise nur (sehr) wenige Studienergebnisse vor, so dass zahlreiche deskriptive Unterschiede nicht signifikant werden und fraglich bleibt, inwiefern die Befunde in diesen Fällen verallgemeinerbar sind.

Zu den Ergebnissen

Der Einsatz von generativer KI hat einen mittleren bis großen positiven Effekt auf Lernleistungen und Lernmotivation von Schülerinnen und Schülern. Der Effekt auf Lernleistungen ist bei Studierenden deskriptiv größer als bei Schülerinnen und Schülern, während es bei der Lernmotivation umgekehrt ist, wobei die Unterschiede nicht statistisch signifikant ausfallen.

Die Effektstärken variieren je nach Fächergruppe, wobei sich in Geisteswissenschaften/Künsten, Informatik/Kommunikationstechnologien, Gesundheit/Sozialwesen sowie Pädagogik die größten Effekte zeigen und in den anderen Bereichen jeweils nur wenige Studien vorlagen.

Die Schnittstelle, über die die Lernenden mit der KI interagieren, moderiert den Effekt auf Lernleistungen, aber nicht auf Lernmotivation. Eine App-basierte Schnittstelle zeigt größere positive Effekte als eine webbasierte Schnittstelle. Die Länge des KI-Einsatzes moderiert ebenfalls die Effektstärke, wobei Lernmotivation stärker durch kurze Interventionen gefördert wird (< 5 Wochen) während Lernleistungen am meisten bei einer Dauer von 5–15 Wochen profitieren.

Relevant ist die Metaanalyse zum einen, weil sie zeigt, dass der Einsatz von generativer KI im Bildungsbereich große Chancen bietet, Lernergebnisse zu verbessern und Lernmotivation zu steigern. Ihre Nutzung sollte also grundsätzlich als Möglichkeit in vielen schulischen Kontexten mitgedacht werden. Interessant ist in diesem Zusammenhang, dass es eine Art „Neuigkeitseffekt“ zu geben scheint, dass also positive Effekte ab einer gewissen Dauer des KI-Einsatzes geringer werden.

Zum anderen zeigt die Metaanalyse, dass die genutzten Technologien und die Rahmenbedingungen eines KI-Einsatzes einen großen Einfluss auf deren Effektivität haben. Es ist also grundsätzlich nötig, diese Faktoren genau zu betrachten und den KI-Einsatz immer wieder zu hinterfragen und anzupassen. Dazu bedarf es eines entsprechenden Know-hows auf Seiten der Lehrkräfte und bei der Umsetzung von Lehrkonzepten.

Weiterführende Hinweise im Kontext

In einer weiteren Rezension können die Ergebnisse einer früheren Metaanalyse zu Effekten des Einsatzes von KI-Anwendungen vor der Einführung von ChatGPT nachgelesen werden (Zheng, Niu, Zhong & Gyasi, 2023). Diese findet einen ähnlich großen positiven Effekt des KI-Einsatzes auf Lernleistungen (g = 0.812, n = 24) und nur einen kleinen (g = 0.208, n = 6) auf die Wahrnehmung des Lernens (u. a. Einstellung zum Lernen, Lernmotivation). Als einflussreiche Moderatorvariablen erweisen sich dort u. a. der Bildungsbereich (geringerer Effekt in der Primarstufe, kein Unterschied zwischen Sekundar- und Tertiärbereich), die Fächergruppe (größere Effekte in Ingenieur-/Technikwissenschaften als in Sozialwissenschaften) und die Sozialform (größere Effekte beim Einsatz in Gruppen als bei individueller Nutzung).

Reflexionsfragen für die Praxis

Nachfolgende Reflexionsfragen sind ein Angebot, die Befunde der rezensierten Studie auf das eigene Handeln als Lehrkraft oder Schulleitungsmitglied zu beziehen und zu überlegen, inwiefern sich Anregungen für die eigene Handlungspraxis ergeben. Die Befunde der rezensierten Studien sind nicht immer generalisierbar, was z. B. in einer begrenzten Stichprobe begründet ist. Aber auch in diesen Fällen können die Ergebnisse interessante Hinweise liefern, um über die eigene pädagogische und schulentwicklerische Praxis zu reflektieren.

Reflexionsfragen für Lehrkräfte

- Welchen Stellenwert hat die Arbeit mit KI in meinem Unterricht?
- In welchen Kontexten (Fächer, Interaktionsformen) setze ich KI ein?
- Inwieweit reflektiere ich die Rahmenbedingungen des KI-Einsatzes?
- Welche weiteren Möglichkeiten gibt es, die positiven Effekte von KI zur Förderung meiner Schülerinnen und Schüler zu nutzen?
- Wie kann ich meine Kompetenzen und die meiner Schülerinnen und Schüler in diesem Bereich ausbauen?

Reflexionsfragen für Schulleitungen

- Welchen Stellenwert hat die Arbeit mit KI an meiner Schule?
- Wo gibt es Potenziale, verstärkt die positiven Effekte von KI zur Förderung von Schülerinnen und Schülern zu nutzen?
- Was kann ich tun, um die Kompetenzen meiner Lehrkräfte in diesem Bereich zu stärken?

Literatur

Bloom, B. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1, cognitive domain.* New York: Longman.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9780203771587

Deng, R., Jiang, M., Yu, X., Lu, Y. & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227,* 105224. https://doi.org/10.1016/j.compedu.2024.105224

Zheng, L., Niu, J., Zhong, L. & Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments, 31*(9), 5650–5664. https://doi.org/10.1080/10494820.2021.2015693

Rezensent/-in

Dr. Sonja Hensel, Lehrerin am Berufskolleg in Siegburg sowie Lehrbeauftragte an der Universität Siegen. Arbeitsschwerpunkte: Rechtschreib-, Schreib- und Lesedidaktik, selbstreguliertes und kooperatives Lernen.

Zitiervorschlag

Hensel, S. (2026). Rezension zu Liu, X., Guo, B., He, W. & Hu, X. (2025). Effects of Generative Artificial Intelligence on K-12 and Higher Education Students' Learn-ing Outcomes: A Meta-Analysis. Journal of Educational Computing Research, 63(5), 1249–1291. *Forschungsmonitor Schule, 201*. Abgerufen von https://www.forschungsmonitor-schule.de/print.php?id=201

Urheberrecht