Forschungsmonitor Schule
Rezensionen für die Praxis

Sonja Hensel

## Rezension zu

Zhang, S., Shan, C., Lee, J. S. Y., Che, S. & Kim, J. H. (2023). Effect of chatbot-assisted language learning: A meta-analysis. *Education and Information Technologies, 28*, 15223–15243.

## Kommentierter Kurzbefund

Chatbots eröffnen niedrigschwellige Interaktionsmöglichkeiten für das Sprachenlernen, allerdings hat sich ihr Einsatz in früheren Studien nicht durchgängig als vorteilhaft erwiesen. Daher untersuchen Zhang et al. in einer Metaanalyse den Nutzen von Chatbots für den Sprachenunterricht im Allgemeinen und beleuchten, unter welchen Bedingungen sich positive Effekte zeigen.

Hierfür recherchierten sie zunächst umfassend in einschlägigen Datenbanken und wählten nach vorab festgelegten Kriterien für ihre Metaanalyse solche Studien aus, die u. a. das Berechnen von Effektstärken und den Vergleich von Lerngruppen mit und ohne Chatboteinsatz erlaubten. Letztlich wurden 18 Studien mit insgesamt 61 Effektgrößen aus den Jahren 2014–2022 einbezogen.

Auf dieser Grundlage gehen die Forschenden 3 Forschungsfragen nach:

1. Wie hoch ist der durchschnittliche Effekt von Chatbot-unterstütztem Sprachenlernen (**c**hatbot-**a**ssisted **l**anguage **l**earning – CALL)?
2. Was sind entscheidende Faktoren (Moderatorvariablen) in CALL-basierten Experimentalstudien?
3. Inwieweit moderieren diese Faktoren die Ergebnisse des Sprachenlernens?

Moderatorvariablen bestimmen, unter welchen Bedingungen, wie stark und in welche Richtung ein Effekt auftritt, z. B. die Dauer der Intervention, der Bereich des Sprachenlernens oder die Stufe des Bildungssystems.

Als Ergebnis finden Zhang et al. einen mittleren positiven Effekt der Chatbotnutzung auf die Lernergebnisse ($g$ = .53). In Bezug auf die Moderatorvariablen zeigen sich signifikante Unterschiede bei der Zielsprache (Erst- oder Fremdsprache), dem Bereich des Sprachenlernens (u. a. höchste Effektivität beim Vokabellernen, keine nachweisbaren Effekte beim Lesen- und Schreibenlernen) und der Art des Lernergebnisses (u. a. bedeutsame Effekte auf Lernverhalten und kognitive Lernergebnisse, kein Effekt auf affektive Lernergebnisse).

Die Befunde sind für Lehrkräfte interessant, weil sie den Schluss nahelegen, dass die Nutzung von Chatbots im Sprachenunterricht unter verschiedenen Rahmenbedingungen im Allgemeinen positive Auswirkungen hat. Daneben lassen sich Ansatzpunkte dafür finden, wie Chatbots möglichst effektiv eingesetzt werden können. Zum Beispiel sollte bedacht werden, in welchen Bereichen des Sprachenlernens (Vokabellernen, Hören, Lesen, Grammatik etc.) der Einsatz von Chatbots vorteilhaft ist.

CC BY NC ND

## Hintergrund

Chatbots, also Computerprogramme, die Gespräche mit menschlichen Personen simulieren, werden zunehmend in Bildungskontexten eingesetzt. Besonders vielversprechend erscheint Zhang et al. der Einsatz in Sprachlernkontexten, da sich das Erlernen einer Sprache immer in sozialer Interaktion vollzieht (Lantolf, 2000) und ein Chatbot als Gesprächspartner neue Möglichkeiten z. B. in Bezug auf Verfügbarkeit, Angebot an Lerngelegenheiten, Abbau von Ängsten und immersive Lernerlebnisse (Wang, Petrina & Feng, 2017) bietet.

Einzelstudien und systematische Literaturübersichten (Huang, Hew & Fryer, 2022) zum Effekt von Chatbots beim Sprachenlernen brachten nach Zhang et al. divergente Ergebnisse. Eine Metaanalyse zum Einsatz von Chatbots im Englischunterricht in Südkorea (Lee & Hwang, 2022) fand einen mittleren positiven Effekt ($g = 0.689$).

Die Forschenden erweitern diese Metaanalyse, indem sie Studien zum Sprachenlernen im Allgemeinen einbeziehen und sich nicht auf ein Land beschränken. Sie gehen folgenden Forschungsfragen nach:

1. Wie hoch ist der durchschnittliche Effekt von Chatbot-unterstütztem Sprachenlernen (**c**hatbot-**a**ssisted **l**anguage **l**earning – CALL)?
2. Was sind entscheidende Faktoren (Moderatorvariablen) in CALL-basierten Experimentalstudien?
3. Inwieweit moderieren diese Faktoren die Ergebnisse des Sprachenlernens?

## Design

Zhang et al. orientierten sich bei der Durchführung und Darstellung ihrer Metaanalyse an den PRISMA-Prinzipien (preferred reporting items for systematic reviews) (Moher, Liberati, Tetzlaff, Altman & Group, 2009).

Sie nutzten für die Suche nach einschlägigen Studien die drei Datenbanken Web of Science, Scopus und Wiley und gaben dort Kombinationen von Suchwörtern wie z. B. „chatbot“ oder „conversational agent“ und „language learning“ oder „language teaching“ ein. Zusätzlich suchten sie nach dem „Schneeballsystem“, also auf der Grundlage bereits gefundener Studien und berücksichtigen Tagungsberichte.

Aus den so gefundenen 201 Studien wählten sie diejenigen aus, die

- Lernsituationen mit und ohne Chatbot vergleichen,
- die notwendige deskriptive Statistik zur Verfügung stellen, um Effektstärken zu berechnen,
- die Ergebnisse des Spracherwerbs auswerten und
- die Lernenden randomisiert den Gruppen mit und ohne Chatbot zuteilen.

Schließlich blieben 18 Studien, die in die Metaanalyse aufgenommen wurden und aus denen 61 Effektgrößen in die Auswertungen eingingen.

Teilweise wurden aus den gefundenen Studien moderierende Variablen übernommen, teilweise wurden neue Moderatorvariablen hinzugefügt, mit denen vergleichbare Merkmale der 18 einbezogenen

Stichproben erfasst wurden. Zwei Forschende identifizierten dabei unabhängig voneinander Moderatorvariablen, aus denen sie in einem diskursiven Prozess ein Kodierschema mit neun Variablen erstellten:

1. Bildungsbereich (Primar- und Sekundarstufe oder Universität)
2. Zielsprache (Erst- oder Fremdsprache)
3. Bereich des Sprachenlernens (Vokabellernen, Grammatik, Schreiben, Hören etc.)
4. Art des Lernergebnisses (kognitiv, affektiv oder in Bezug auf Lernverhalten)
5. Dauer der Intervention (≤ 1 Monat, 1-3 Monate, ≥ 3 Monate)
6. Chatbot-Schnittstelle (App oder Browser)
7. Chatbot-Entwicklung (bereits vorhanden oder für die Studie geschaffen)
8. Aufgabendominanz (Chatbot oder die menschlichen Nutzenden)
9. Interaktionsart (Text oder Text und Stimme)

Für diese wurden jeweils Effektstärken (Hedges‘ *g*) berechnet, wobei ein Wert von 0.20 – 0.49 als kleiner, ein Wert von 0.50 – 0.79 als mittlerer und ein Wert ab 0.80 als großer Effekt galt.

## Ergebnisse

Der Gesamteffekt des Chatbot-Einsatzes im Sprachenunterricht über alle ausgewerteten Studien hinweg beträgt $g = 0.527$ ($SE = 0.080$, 95 % KI [0.371, 0.684], $p = 0.000$), was einer mittleren Effektstärke entspricht.

Im Hinblick auf folgende Unterscheidungsmerkmale beim Sprachenlernen ergeben sich bei einer Chatbot-Nutzung signifikant bessere Ergebnisse im Vergleich zur Kontrollgruppe (in Klammern die Zahl der zugrundeliegenden Effektgrößen):

- Bildungsbereich
  - Primar- und Sekundarstufe: $g = 0.786$ (n = 20)
  - Hochschulbildung: $g = 0.422$ (n = 41)
- Zielsprache
  - Fremdsprache: $g = 0.543$ (n = 59)
- Bereich des Sprachenlernens
  - Vokabellernen: $g = 1.827$ (n = 8)
  - Hören: $g = 0.752$ (n = 1)
  - Sprechen: $g = 0.647$ (n = 18)
  - Grammatik: $g = 0.343$ (n = 15)
- Art des Lernergebnisses
  - Lernverhalten: $g = 1.250$ (n = 1)
  - Kognitiv: $g = 0.578$ (n = 46)
- Dauer der Intervention
  - 1–3 Monate: $g = 0.681$ (n = 31)
  - ≥ 3 Monate: $g = 0.444$ (n = 23)
- Chatbot-Schnittstelle
  - Browser: $g = 0.619$ (n = 37)
  - App: $g = 0.388$ (n = 24)
- Chatbot-Entwicklung

    - Für die Studie geschaffen: $g = 0.554$ (n = 39)
    - Existierende Systeme: $g = 0.481$ (n = 22)
- Aufgabendominanz
    - Nutzergesteuert: $g = 0.643$ (n = 29)
    - Chatbot-gesteuert: $g = 0.400$ (n = 32)
- Interaktionsart
    - Text + Stimme: $g = 0.706$ (n = 25)
    - Text: $g = 0.413$ (n = 36)

Signifikante Unterschiede zwischen den Ausprägungen der Moderatorvariablen werden vor allem für die Zielsprache (kein Effekt auf das Lernen der Erstsprache) und die Art des Lernergebnisses (kein Effekt auf affektive Lernergebnisse) sichtbar, wobei zum Lernen der Erstsprache nur zwei Effektgrößen vorlagen und zum Lernverhalten nur eine Effektgröße.

Deutlich sind die Unterschiede zudem in Bezug darauf, in welchem Bereich des Sprachenlernens der Chatbot zum Einsatz kommt: Der Effekt beim Vokabellernen ($g = 1.827$) ist sehr groß, beim Hörverstehen ($g = 0.752$) und Sprechen ($g = 0.647$) mittelgroß und bei Grammatik klein ($g = 0.343$). Bei der Förderung der Kompetenzen im Bereich Lesen und Schreiben sind keine Effekte nachweisbar.

Auch bei den anderen Moderatorvariablen zeigen sich deskriptiv teilweise Unterschiede $g > 0.4$ zwischen den Ausprägungen, allerdings werden diese aufgrund der geringen Anzahl an zugrundeliegenden Effektgrößen nicht signifikant.

## Diskussion und Einschätzung

### Hintergrund

Anlage und Fragestellung der Studie werden nachvollziehbar in den aktuellen Forschungsdiskurs eingebettet. Durch den Entstehungszeitraum der Studie sind allerdings nur Forschungsergebnisse bis Oktober 2022 berücksichtigt worden.

### Design

Die Metaanalyse wird transparent nach den PRISMA-Prinzipien (**p**referred **r**eporting **i**tems for **s** ystematic reviews and **m**eta-**a**nalysis) (Moher et al., 2009) durchgeführt und dargestellt.

Beschränkungen liegen unter anderem darin, dass bei einigen Variablen wie zum Beispiel dem Lernverhalten nur wenige Studien und Effektgrößen in die Analyse eingegangen sind, beim Lernbereich Lesen sogar nur eine Studie. Zudem sind die Ausprägungen der Variablen zum Teil ungenau beschrieben, wie zum Beispiel, dass mit Lernergebnissen im Bereich Verhalten ausschließlich die „Teilnahme“ (participation) gemeint zu sein scheint? Hier schafft erst ein Blick in die Originalstudie Aufklärung (s. nächster Abschnitt). Das schränkt die Aussagekraft und Nutzbarkeit der Ergebnisse ein. Außerdem liegt der Schwerpunkt der verarbeiteten Studien in asiatischen Staaten, was durch die Unterschiede in den Bildungssystemen Fragen nach der Übertragbarkeit aufwirft.

Zudem sind handwerkliche Fehler zu konstatieren. In der tabellierten Ergebnisdarstellung zu den Moderatorvariablen werden die Z- und P-Werte der Effektstärken mehrheitlich nicht korrekt angegeben. Entsprechend sind die im Text dokumentierten Signifikanzniveaus der Effektstärken häufig falsch und inkonsistent zu den inhaltlichen Ausführungen.

Daneben ähnelt der Diskussionsbeitrag zur Wirksamkeit beim Erlernen der Erst- und Fremdsprache dem einer anderen Studie (Lee & Lee, 2022), obwohl der eigene Befund im Gegensatz dazu steht. Schließlich ist zumindest unkonventionell, dass deskriptive Unterschiede trotz fehlender Signifikanz interpretiert werden.

Wenngleich demnach die wissenschaftliche Sorgfalt an einigen Stellen zu wünschen übrig lässt, wirken die Ergebnisse ansonsten stimmig.

**Ergebnisse**

Zusammenfassend lässt sich festhalten, dass ein Chatbot-Einsatz im Sprachenunterricht im Allgemeinen einen positiven Effekt mittlerer Stärke nach sich zieht. Nur für wenige Settings ist kein Vorteil gegenüber den Kontrollgruppen ohne Chatbot-Einsatz nachweisbar, was teilweise der geringen Anzahl an auswertbaren Effektgrößen geschuldet sein dürfte und nicht dahingehend interpretiert werden kann, dass Chatbots, die auf neueren (z. B. KI-)Sprachverarbeitungstechnologien basieren, unter der jeweiligen Bedingung ebenso wirkungslos bleiben.

Dies gilt beispielsweise für den ausbleibenden Effekt beim Lernen der Erstsprache. Dieses Ergebnis beruht auf nur zwei Effektgrößen und demgegenüber zeigt sich in der Metaanalyse von Lee und Lee (2022) zu interaktiven Sprachlernrobotern ein Effekt auf das Lernen der Erstsprache ($d = 0.72$, $n = 2$), der sich dort nicht signifikant vom Einsatz im Fremdsprachenunterricht ($d = 0.55$, $n = 9$) unterschied.

Signifikante Unterschiede ergeben sich für die verschiedenen Bereiche des Sprachenlernens. Hier weist das Vokabeltraining einen sehr großen Effekt auf, gefolgt von Interventionen zum Hörverstehen und Sprechen mit mittleren Effekten, während Maßnahmen zum Grammatik-, Lesen- und Schreibenlernen nur geringe oder gar keine Vorteile gegenüber den Kontrollgruppen haben.

In einer neueren Metaanalyse von Saarela, Gunasekara und Kumarage (2026) wird davon abweichend für das Schreibenlernen mithilfe von Tools wie ChatGPT eine mittlere Effektstärke ($g = .545$, $n = 68$) belegt, was erneut verdeutlicht, dass Effekte von Chatbot-unterstütztem Sprachenlernen abhängig von Technologie und Einsatzszenario unterschiedlich ausfallen können.

Weitere signifikante Unterschiede bestehen je nach Art des Lernergebnisses. Der Effekt auf das Lernverhalten sticht hervor, die Wirkung auf kognitive Lernergebnisse liegt im mittleren Bereich, wohingegen sich affektive Lernergebnisse nicht von den Kontrollgruppen unterscheiden. Hierbei muss jedoch beachtet werden, dass die Befunde zum Lernverhalten nur aus einer Studie (Goda, Yamada, Matsukawa, Hata & Yasunami, 2014) stammen und sich darauf beziehen, ob Lernende sich nach einer Konversation mit einem Chatbot aktiver an einer Gruppendiskussion beteiligen. Bei den anderen Variablen zeigen sich keine signifikanten Unterschiede zwischen den Ausprägungen der Moderatorvariablen, wobei auch hierfür die geringe Zahl an Effektgrößen eine Rolle gespielt haben mag.

Für den Unterricht kann aus dieser Studie mitgenommen werden, dass vom Einsatz eines Chatbots im Sprachenunterricht positive Effekte zu erwarten sind, dieser gemäß der bisherigen Befundlage insbesondere das Vokabellernen und Interventionen zum Hörverstehen und Sprechen positiv beeinflussen kann. In diesem Zusammenhang ist auch zu überlegen, ob der Aufwand lohnt, einen Chatbot selbst zu kreieren oder ob es ausreicht, vorhandene, weniger spezifisch auf die Lernenden zugeschnittene Angebote zu nutzen, da diesbezüglich kein signifikanter Unterschied nachgewiesen wurde.

## 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

- Inwiefern nutze ich in meinem Unterricht Chatbots als Lernpartner für meine Schülerinnen und Schüler?
- In welchem Verhältnis stehen die Studienergebnisse zu meinen persönlichen Erfahrungen?
- Welche Faktoren beziehe ich ein, wenn ich die Rahmenbedingungen ihres Einsatzes reflektiere?
- Was brauche ich, um Chatbots regelmäßig im Unterricht einzusetzen?
- Welche Möglichkeiten gibt es für mich, um mich darüber mit anderen auszutauschen und gemeinsame Lösungen zu entwickeln?

### Reflexionsfragen für Schulleitungen

- Inwiefern nutzen die Lehrkräfte meiner Schule im Unterricht Chatbots als Lernpartner für ihre Schülerinnen und Schüler?
- Wie wird der Einsatz reflektiert?
- Welche Unterstützung erhalten die Lehrkräfte, um Chatbots nutzen zu können?
- Welchen Entwicklungsbedarf sehe ich in diesem Bereich?

## Literatur

Goda, Y., Yamada, M., Matsukawa, H., Hata, K. & Yasunami, S. (2014). Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of information and systems in education*, *13*, 1–7.

Huang, W., Hew, K. F. & Fryer, L. K. (2022). Chatbots for language learning — Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38* (1), 237–257.

Lantolf, J. P. (2000). Introducing sociocultural theory. In J. P. Lantolf (Ed.), *Sociocultural Theory and Second Language Learning* (1–26). Oxford: Oxford University Press.

Lee, J.-Y. & Hwang, Y. (2022). A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language & Literature*, *48*(1), 213–243.

Lee, H. & Lee, J. H. (2022). The effects of robot-assisted language learning: A meta-analysis. *Educational Research Review*, *35*, 100425.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264–269.

Saarela, M., Gunasekara, S. & Kumarage, P. (2026). A meta-analysis of generative AI effects on language proficiency and affective-cognitive outcomes in language learning. *Discover Computing*, *29*, 116.

Wang, Y. F., Petrina, S. & Feng, F. (2017). VILLAGE—Virtual Immersive Language Learning and Gaming Environment: Immersion and presence. *British Journal of Educational Technology*, *48*(2), 431–450.

## 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 Zhang, S., Shan, C., Lee, J. S. Y., Che, S. & Kim, J. H. (2023). Effect of chatbot-assisted language learning: A meta-analysis. Education and Information Technologies, 28, 15223–15243. *Forschungsmonitor Schule, 198*. Abgerufen von https://www.forschungsmonitor-schule.de/print.php?id=197

## Urheberrecht

Dieser Text steht unter der CC BY-NC-ND 4.0 Lizenz. Der Name des Urhebers / der Urheberin soll bei einer Weiterverwendung wie folgt genannt werden: Sonja Hensel (2026) für den Forschungsmonitor Schule.