Onderzoeksreview van Zhang et al. (2026)

When does GenAI work for L2 learning? A three-level meta-analysis

Review of Education | DOI: 10.1002/rev3.70252 | 14 november 2026

SAMENVATTING

Zhang et al. brachten 48 onderzoeken naar generatieve AI bij het leren van een tweede taal bijeen. Hun gemiddelde effect bedraagt g = 0,76. Zonder drie uitzonderlijk hoge effecten is dat g = 0,57. De onderzoeken verschillen sterk van elkaar. Bovendien tellen de deelnemersaantallen in tabel 2 op tot 4.378, terwijl de hoofdtekst 5.196 deelnemers vermeldt. Dat verschil van 818 is niet verklaard. Ook de selectie van enkele studies, de indeling naar onderzoeksopzet en twee statistische berekeningen roepen vragen op. De resultaten wijzen op betere prestaties in bepaalde lessituaties. Hoe groot het voordeel precies is, wat het veroorzaakt en wat leerlingen later zelfstandig kunnen, blijft minder duidelijk.



Figuur 1. Vier bevindingen die meewegen bij de beoordeling van Zhang et al. [1]

Wat onderzochten Zhang en collega's?

Zhang, Aubrey, Wu en Chiu onderzoeken of generatieve AI helpt bij het leren van een tweede of vreemde taal. Ze bekijken ook of de manier waarop leerlingen en docenten AI gebruiken verschil maakt. Hun meta-analyse bundelt onderzoeken naar schrijven, spreken, lezen en woordenschat. AI vervult daarin uiteenlopende rollen, van gesprekspartner tot feedbackgever. Die verschillen zijn relevant: een gesprek voeren met een chatbot is een andere leeractiviteit dan een tekst verbeteren op basis van automatische feedback. [1, pp. 1–6]

De onderzoekers doorzochten zes wetenschappelijke databanken voor publicaties uit de periode 2023 tot februari 2026. Ze beoordeelden 458 volledige artikelen en selecteerden er 48. Samen leverden die 67 effectgroottes op. Zo'n effectgrootte maakt resultaten van verschillende toetsen vergelijkbaar door het verschil tussen groepen in een gemeenschappelijke maat uit te drukken. Sommige studies leverden meerdere effecten. Het gebruikte model met drie niveaus houdt rekening met het feit dat uitkomsten uit hetzelfde onderzoek samenhangen. Dat is verstandig. Twee toetsen uit één klas vormen nu eenmaal geen twee onafhankelijke onderzoeken. [1, pp. 7–16]

Zhang en collega's kijken verder dan het gemiddelde. Ze onderzoeken verschillen tussen taalvaardigheden, toepassingen en lessituaties. Ook berekenen ze hoe het gemiddelde verandert zonder uitzonderlijk hoge effecten en controleren ze op

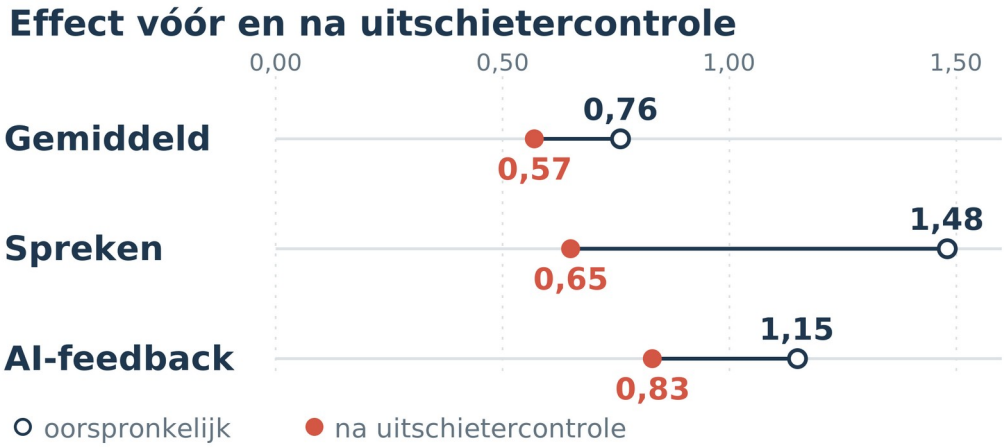
mogelijke publicatievertekening. Dat zijn passende controles bij een verzameling studies die zo sterk uiteenloopt. [1, pp. 14–20]

Wat laten de uitkomsten zien?

Gemiddeld rapporteren de auteurs Hedges' g = 0,76, met een 95%-betrouwbaarheidsinterval van 0,50 tot 1,02. De groepen die AI gebruikten, scoorden op de gemeten taaluitkomsten doorgaans hoger dan de vergelijkingsgroepen. Het cijfer 0,76 betekent geen vooruitgang van 76 procent. Het is een gestandaardiseerd verschil tussen groepen. Daaruit valt ook niet af te leiden hoeveel iedere afzonderlijke leerling heeft geleerd. [1, p. 17]

Het gemiddelde verschilt per taalvaardigheid. Voor spreken vinden de auteurs g = 1,48, voor woordenschat 0,66 en voor schrijven 0,61. Bij lezen ligt de schatting op g = 0,076. Dat verschil is statistisch niet significant. De vijf leesstudies leveren samen dus geen overtuigend positief effect op. Daaruit volgt niet dat AI bij leesonderwijs nooit kan helpen. [1, p. 18]

Drie uitzonderlijk hoge effectgroottes drukken zwaar op het gemiddelde: 5,70 en 4,56 uit Elmaadaway et al. (2025), en 5,16 uit Wang (2025). Laat je ze weg, dan daalt de totale schatting van 0,76 naar 0,57. Bij spreken zakt het effect van 1,48 naar 0,65; bij AI-feedback van 1,15 naar 0,83. De uitkomst blijft in deze herberekening positief. Maar de grootte ervan verandert merkbaar door slechts drie resultaten weg te laten. [1, p. 20]



Figuur 2. De gepubliceerde schattingen op één schaal. Open cirkel: volledige analyse. Rode punt: na het weglaten van uitschieters. [1, p. 20]

Waarom lopen de resultaten zo uiteen?

De auteurs vinden een heterogeniteit van I^2 = 92,64%. De gemeten effecten lopen dus veel verder uiteen dan op basis van toeval tussen steekproeven te verwachten is. Dat cijfer betekent niet dat 92,64 procent van de studies slecht is. Het laat zien waarom één gemiddeld effect weinig zegt over wat er in een bepaalde klas zal gebeuren. [1, p. 17]

Er zijn ook aanwijzingen voor kleine-studie-effecten. De clusterrobuuste Egger-test geeft p = 0,006. Dat kan wijzen op publicatievertekening, bijvoorbeeld wanneer positieve resultaten vaker in tijdschriften terechtkomen. Andere verklaringen zijn mogelijk. De toets laat niet zien welke publicaties eventueel ontbreken. [1, pp. 17, 23]

Passen alle opgenomen studies bij de onderzoeksvraag?

Een meta-analyse staat of valt met de studies die erin terechtkomen. Zhang et al. eisen dat een onderzoek betrekking heeft op het leren van een tweede taal, een taaluitkomst meet en een vergelijking tussen groepen maakt. De gebruikte toepassing moet bovendien generatieve taal kunnen produceren. Toch is bij enkele opgenomen onderzoeken niet duidelijk hoe ze aan deze voorwaarden voldoen. [1, pp. 7–8]

Bašić et al. (2023) laten achttien masterstudenten forensische wetenschappen een essay over biometrische identificatie schrijven. Ze schrijven in het Kroatisch, hun moedertaal. De onderzoekers meten onder meer tekstkwaliteit, schrijftijd en authenticiteit met en zonder ChatGPT. Dat zegt iets over schrijfondersteuning. Uit de publicatie blijkt niet dat de studenten een tweede taal leren. Waarom deze studie toch is meegenomen, vraagt om uitleg. [1, p. 9; 2]

Avello et al. (2024) kijken of ChatGPT masterstudenten helpt om digitaal verhalen vertellen beter te begrijpen en of dat de cognitieve belasting verandert. Zhang et al. rekenen het onderzoek mee bij schrijven. Maar kennis over digitale verhalen vertellen is op zichzelf geen maat voor het leren van een tweede taal. Welke specifieke taalvaardigheid hier is gemeten, blijft onduidelijk. [1, pp. 8–9; 3]

Bij Aydın Yıldız (2023) is de situatie anders. Motivatie en leerstrategieën krijgen veel aandacht, maar er is ook een woordenschattoets afgenomen. De studie kan dus niet zomaar worden uitgesloten. De vraag is welke uitkomst Zhang et al. hebben gebruikt: de woordenschattoets of een vragenlijst die geen taalvaardigheid meet. [1, p. 9; 4]

Daarnaast moet vaststaan dat de onderzochte technologie werkelijk generatieve AI gebruikt. Elmaadaway et al. (2025) onderzochten hardop lezen met Amazon Alexa. Hwang et al. (2023) gebruikten Smart RoamLingo, dat herkenningstechnologie en feedback combineert. De naam ‘AI’ zegt nog niet of een taalmodel de relevante antwoorden genereerde. Dat moet per toepassing worden aangetoond. Bij Alexa is die vraag extra relevant: twee van de drie extreme effectgroottes komen uit dit onderzoek. [1, pp. 7–10, 20; 5, 6]



Figuur 3. Zes opgenomen onderzoeken met vragen over de taaluitkomst, de gebruikte techniek of de indeling naar onderzoeksopzet. [1–7]

Hoe stevig zijn de vergelijkingen?

In tabel 2 staan 25 gerandomiseerde en 23 quasi-experimentele studies. Alenezi en Alenezi (2025) is als gerandomiseerd experiment ingedeeld. De oorspronkelijke

auteurs noemen hun eigen onderzoek echter quasi-experimenteel en beschrijven een doelgerichte selectie van dertig scholieren. De geraadpleegde publicatie laat niet zien dat de leerlingen willekeurig over de groepen zijn verdeeld. [1, pp. 9, 16; 7]

Dat is niet alleen een kwestie van naamgeving. Zhang et al. gebruiken RoB 2 om gerandomiseerde studies op vertekening te beoordelen en ROBINS-I voor niet-gerandomiseerde studies. Twee beoordelaars voerden die beoordeling onafhankelijk uit; volgens de auteurs kwamen hun oordelen in 85 procent van de gevallen overeen. Voor Alenezi moet worden nagegaan welk instrument werkelijk is toegepast en of dat bij de onderzoeksopzet past. [1, p. 15; 7]

De indeling werkt ook door in de resultaten. Voor gerandomiseerde studies rapporteren Zhang et al. gemiddeld g = 1,02; voor quasi-experimentele studies g = 0,48. Een verkeerde classificatie kan die vergelijking beïnvloeden. Hoeveel, valt pas te zeggen nadat de studie opnieuw is ingedeeld en de analyse is herhaald. [1, p. 19]

Ook de vergelijking tussen AI-groepen en controlegroepen vraagt aandacht. Bij Meyer et al. (2024) werkten 459 leerlingen aan een Engelse schrijfopdracht. De AI-groep kreeg feedback van een taalmodel en mocht de tekst daarmee verbeteren. De andere groep moest zonder die feedback reviseren. Het voordeel voor de AI-groep was klein (d = 0,19). Dat is een geldig resultaat voor deze vergelijking. Het laat alleen niet zien wat AI toevoegt ten opzichte van evenveel feedback van een docent. [8]

In andere onderzoeken veranderen met de invoering van AI eveneens meerdere onderdelen van de les. Leerlingen krijgen bijvoorbeeld meer feedback, meer oefentijd of andere instructie. Een gemeten voordeel kan dan door het hele pakket ontstaan. Het effect van het taalmodel afzonderlijk is niet te bepalen. Zhang et al. erkennen die beperking. [1, p. 23; 9]

Waarom kloppen de deelnemersaantallen niet?

In de resultaten staat dat 5.196 mensen deelnamen, verdeeld over 2.584 deelnemers in de AI-groepen en 2.612 in de controlegroepen. Tabel 2 geeft voor elk van de 48 onderzoeken een steekproefgrootte. Tel je die aantallen op, dan kom je uit op 4.378. De tabel loopt van pagina 9 tot en met 12; op pagina 13 begint tabel 3. De vier paginatotalen zijn 1.041, 1.689, 1.261 en 387. Samen scheelt dat 818 deelnemers met het opgegeven totaal. [1, pp. 9–12, 16]

De publicatie verklaart het verschil van 818 niet. Misschien staat er een fout in de tabel, misschien zijn voor de totalen andere telregels gebruikt. De steekproefgrootte speelt mee bij het gewicht dat een studie in de berekening krijgt. De gebruikte groepsaantallen zijn nodig om na te gaan of alleen de verslaglegging verschilt of ook de analyse. [1]



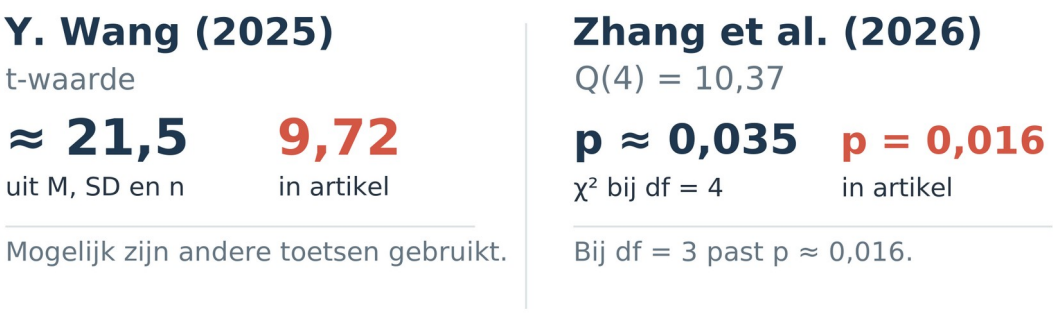
Figuur 4. Het rode deel geeft uitsluitend het verschil tussen twee gepubliceerde totalen aan. Het stelt geen afzonderlijke groep deelnemers voor. [1]

Wat laten de herberekeningen zien?

Een van de grootste effecten is afkomstig uit de afzonderlijke studie van Y. Wang (2025). De publicatie noemt groepen van 33 en 35 studenten, gemiddelden van 3,66 en 1,27 en standaarddeviaties van 0,38 en 0,52. Daaruit volgt Hedges' g ≈ 5,16, zoals Zhang rapporteert. Maar een gewone t-toets tussen deze onafhankelijke groepen geeft op basis van dezelfde cijfers t ≈ 21,5. In de bron staat t = 9,72. Mogelijk gaat die waarde over een andere toets of uitkomst. Zonder nadere uitleg zijn beide getallen niet als dezelfde vergelijking te reconstrueren. [1, p. 20; 10]

Ook de moderatoranalyse van Zhang bevat een verschil. De auteurs vermelden Q(4) = 10,37 en p = 0,016. Bij vier vrijheidsgraden hoort bij die Q-waarde ongeveer p = 0,035. Drie vrijheidsgraden leveren wel ongeveer 0,016 op en zouden bij vier categorieën gebruikelijk zijn. Het kan dus om een fout in de notatie gaan. De oorspronkelijke modeluitvoer is nodig om dat te controleren. [1, p. 19]

Het is niet vastgesteld of deze afwijkingen de resultaten veranderen. Daarvoor moeten we weten welke toetsen, uitkomsten en modelinstellingen daadwerkelijk zijn gebruikt. De gepubliceerde cijfers geven die informatie niet volledig.



Figuur 5. Berekende waarden vergeleken met de waarden die in de oorspronkelijke publicaties staan. [1, 10]

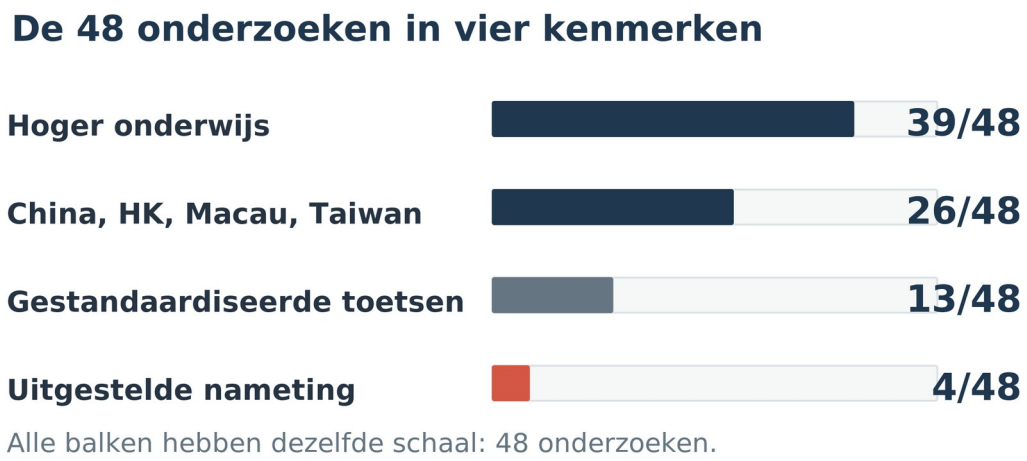
Voor welke leerlingen gelden de bevindingen?

De 48 studies vormen geen afspiegeling van al het taalonderwijs. Negenendertig zijn uitgevoerd in het hoger onderwijs en 26 in China, inclusief Hongkong, Macau en Taiwan. Slechts dertien studies gebruikten volgens Zhang een gestandaardiseerde taaltoets. Andere onderzoekers maakten eigen toetsen. Die hoeven niet slecht te zijn, maar verschillen in meetinstrument maken de studies lastiger vergelijkbaar. [1, p. 16]

Voor hun indeling naar taalniveau rekenen de auteurs onder meer Egyptische leerlingen uit het vierde leerjaar en Chinese leerlingen uit leerjaar 6 tot de beginners (CEFR A1–A2 of vergelijkbaar). Zij baseren dat op leeftijd en onderwijscontext, zonder voor deze groepen een gemeten CEFR-niveau te vermelden. Het zijn dus niet overal dezelfde soorten gegevens waarop de vergelijking tussen taalniveaus rust. [1, pp. 8, 13]

Over de invloed van de lesduur is weinig zekerheid. Voor die analyse hadden de auteurs gegevens over lesuren uit slechts 17 studies, samen 24 effectgroottes. Zij vonden geen overtuigend statistisch verband tussen het aantal uren en het effect. Die uitkomst laat niet zien dat lesduur er niet toe doet. Daarvoor is de onderzochte groep te beperkt. [1, p. 19]

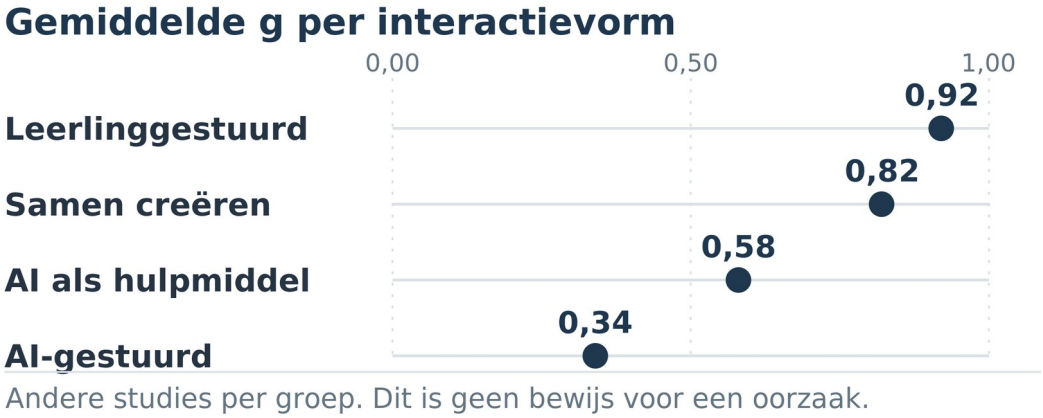
Slechts vier van de 48 studies bevatten een uitgestelde nameting. De meeste uitkomsten zijn tijdens of kort na de lessen gemeten. Dat kan zeker iets over leren zeggen, vooral wanneer leerlingen een vaardigheid zonder AI-hulp op een goede toets laten zien. Maar een tekst die tijdens een oefening met AI-feedback beter wordt, bewijst niet dat de leerling weken later zonder hulp even goed schrijft. Ook toepassing in een andere taak moet apart worden onderzocht. [1, pp. 16, 23; 8]



Figuur 6. Vier kenmerken van de 48 studies. De categorieën overlappen en vormen samen geen verdeling van alle studies. [1, p. 16]

Zhang et al. vergelijken ook de manier waarop het initiatief tussen leerling en AI is verdeeld. Voor leerlinggestuurd gebruik rapporteren ze g = 0,92, voor samen creëren 0,82, voor instrumenteel gebruik 0,58 en voor AI-gestuurd gebruik 0,34. Maar die schattingen komen uit verschillende groepen studies. Er is niet één experiment waarin alleen het initiatief veranderde. Een verband tussen de categorie en de uitkomst bewijst dus niet dat meer leerlingregie het effect veroorzaakt. [1, p. 19]

De onderzoekers bepaalden de categorie vooral op basis van gepubliceerde lesbeschrijvingen. Twee beoordelaars kwamen vaak tot hetzelfde oordeel (Cohen’s κ = 0,86). Dat wijst op een consistente codering. Het toont niet wie tijdens de les werkelijk initiatief nam, fouten opmerkte, feedback woog of een formulering koos. Die handelingen zijn niet rechtstreeks gevolgd. [1, pp. 8, 14, 21]



Figuur 7. Effectgroottes per interactievorm. Het zijn verschillende groepen studies, geen proef waarin alleen de taakverdeling veranderde. [1, p. 19]

Voor de rol van de docent vonden de auteurs in de hoofdanalyse geen overtuigend verschil. Pas na het verwijderen van de uitschieters kwam de categorie met een meer observerende docent gunstiger naar voren. De auteurs noemen dit verkennend. Er is geen grond om hieruit af te leiden dat docenten zich in het algemeen beter kunnen terugtrekken. [1, pp. 19–23]

Wat kunnen we uit dit onderzoek afleiden?

Ook zonder de drie extreme effecten blijft het gemiddelde resultaat positief. Dat verdient erkenning. De keuze voor een model met drie niveaus past bij de gegevens en de onderzoekers laten zien hoe groot de verschillen tussen studies zijn. De controles op uitschieters en mogelijke publicatievertekening maken de uitkomst inzichtelijker. In meerdere onderzochte lessituaties halen leerlingen met AI hogere scores op de gemeten taaluitkomsten. [1]

De precieze grootte van dat voordeel kunnen we uit de gepubliceerde tabellen niet onafhankelijk narekenen. Daarvoor zijn alle 67 effectgroottes nodig, met de groepsaantallen, spreidingsmaten, onderzoeksopzetten en gebruikte AI-toepassingen. Vervolgens kan worden nagegaan wat het effect is van een andere bronselectie of een gewijzigde classificatie van Alenezi. Die gevolgen zijn nu nog niet te berekenen.

De onderwijsbetekenis vraagt een tweede controle. Welke vaardigheid is gemeten en hoeveel hulp kregen leerlingen tijdens de toets? Een hogere score is een relevant resultaat. Wat een leerling zelfstandig kan, blijft onbekend zodra de benodigde handeling door AI werd uitgevoerd en daarover geen onafhankelijke toets beschikbaar is. Voor onthouden over tijd en toepassen in nieuwe situaties biedt deze meta-analyse weinig rechtstreeks bewijs. Een algemene conclusie over blijvende taalvaardigheid gaat daarom verder dan de gegevens dragen.

Wat is nodig om de uitkomsten te controleren?

Zhang et al. schrijven dat alle geanalyseerde gegevens in het artikel zijn opgenomen en dat aanvullende gegevens op verzoek beschikbaar zijn. De 67 effectgroottes en hun standaardfouten staan echter niet met alle groepsaantallen, kwaliteitsbeoordelingen en modelvariabelen in één volledig narekenbare dataset. De auteurs noemen de software, maar leveren in de geraadpleegde publicatie geen uitvoerbare analysecode. Met de tabellen alleen kan de meta-analyse niet zelfstandig worden herhaald. [1, pp. 15, 24]

Met de ontbrekende gegevens kan de analyse opnieuw worden uitgevoerd: eerst met dezelfde studies als de auteurs gebruikten, daarna met twijfelachtige studies afzonderlijk weggelaten en met de onderzoeksopzetten opnieuw beoordeeld. Zo wordt duidelijk of de gevonden afwijkingen beperkt blijven tot de publicatie of ook het gepoolde resultaat raken. Tot die tijd staat er een positief gerapporteerd effect, maar is de precieze omvang niet onafhankelijk bevestigd.

- Alle 67 effectgroottes, inclusief de standaardfouten, onderlinge samenhang en werkelijk gebruikte aantallen per groep.
- De lijst van opgenomen en uitgesloten studies, met de reden van de keuze en de gebruikte taaluitkomst per onderzoek.
- Technische informatie over de gebruikte Alexa-toepassing en Smart RoamLingo, zodat is vast te stellen of zij aan de GenAI-criteria voldoen.
- De codering van onderzoeksopzet en interactievormen, met onderbouwing van de RCT-indeling bij Alenezi.
- De analysecode en de uitkomsten van nieuwe berekeningen, ook zonder de drie uitschieters en met afzonderlijke gevoeligheidsanalyses voor betwiste studies.

Bronnen

[1] Zhang, Z., Aubrey, S., Wu, X., & Chiu, T. K. F. (2026). When does GenAI work for L2 learning? A three-level meta-analysis. Review of Education, 14, e70252. DOI

[2] Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. Humanities & Social Sciences Communications. DOI

[3] Avello, R., Gajderowicz, T., & Gómez-Rodríguez, V. G. (2024). Is ChatGPT helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? Revista de Educación a Distancia. DOI

[4] Aydın Yıldız, T. (2023). The impact of ChatGPT on language learners' motivation. Journal of Teacher Education and Lifelong Learning, 5(2), 582–597. DOI

[5] Elmaadaway, M. A. N., et al. (2025). Improving primary school students' oral reading fluency through voice chatbot-based AI. Journal of Computer Assisted Learning. DOI

[6] Hwang, W.-Y., et al. (2023). AI and recognition technologies to facilitate English as foreign language writing. Journal of Educational Computing Research. DOI

[7] Alenezi, A., & Alenezi, A. (2025). Evaluating the effectiveness of chatbot-assisted learning in enhancing English conversational skills. Education Sciences, 15(9), 1136. DOI

[8] Meyer, J., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom. Computers and Education: Artificial Intelligence, 6, 100199. DOI

[9] Lawson, A. P., et al. (2026). Color me confounded: A critical analysis of media comparisons on ChatGPT in education. Computers & Education, 254, 105701. DOI

[10] Wang, Y. (2025). A study on the efficacy of ChatGPT-4 in enhancing students' English communication skills. SAGE Open, 15(1). DOI

Rekennotities en controlepunten

Deelnemers: p. 9 = 1.041 (14 studies); p. 10 = 1.689 (13); p. 11 = 1.261 (15); p. 12 = 387 (6). 48 studies, 4.378 deelnemers. In de hoofdtekst staat N = 5.196; verschil 818. Tabel 3 begint op p. 13. [1]

Wang (2025): $n_1 = 33$, $n_2 = 35$, $M_1 = 3{,}66$, $M_2 = 1{,}27$, $SD_1 = 0{,}38$, $SD_2 = 0{,}52$. Gepoolde SD ≈ 0,4575; d ≈ 5,224; Hedges' g ≈ 5,164; t bij onafhankelijke groepsvergelijking ≈ 21,5. De bron vermeldt ook t = 9,72; de toetsvergelijking moet worden opgehelderd. [10]

Interactie-moderator Zhang (p. 19): $Q(4) = 10{,}37$, $p = 0{,}016$ volgens artikel. Voor χ^2 met df = 4 volgt p ≈ 0,0346; bij df = 3 volgt p ≈ 0,0157. [1]

Dit is een gerichte controle van gepubliceerde cijfers en geselecteerde bronnen, geen volledige reconstructie van alle 67 effectgroottes. EAI Core 0.5.4 wordt alleen gebruikt om prestatie, zelfstandigheid, behoud en transfer niet als gelijkwaardige uitkomsten te behandelen.

EAI-verantwoording: totstandkoming van deze review

Hans Visser en ChatGPT werkten samen aan deze onderzoeksreview. De analyse, de tekst en de vormgeving kwamen gezamenlijk tot stand. De werkwijze hieronder is beschreven op basis van Vissers opgave, niet op een volledig logboek van alle gebruikte tools.

Activiteit	Hans Visser	ChatGPT
Onderzoek	• bronnen selecteren	• zoeken en ordenen
Analyse	• onderzoek en duiding	• narekenen, vergelijken
Tekst en figuren	• veel tekst herschreven	• concepttekst, beeld
Eindcontrole en inhoudelijke verantwoordelijkheid: Hans Visser		

Figuur 8. De analyse en het schrijfwerk werden gedeeld. Hans Visser voerde de inhoudelijke eindcontrole uit.

Brononderzoek. Hans Visser bepaalde de onderzoeksvragen, selecteerde bronnen en beoordeelde hun bruikbaarheid. Volgens de door hem beschreven werkwijze ondersteunde de Exa-agent het zoeken naar oorspronkelijke publicaties en werd Firecrawl gebruikt om openbare webpagina’s uit te lezen. ChatGPT hielp bij het ordenen en vergelijken van het gevonden materiaal.

Analyse. Hans Visser en ChatGPT onderzochten samen de methode, de bronselectie, de berekeningen en de conclusies. Visser controleerde de onderbouwing en bracht eigen analyses in. ChatGPT hielp onder meer bij het opnieuw berekenen van gepubliceerde statistieken en het naast elkaar leggen van bronnen. De inhoudelijke afwegingen zijn in samenwerking gemaakt.

Schrijven en redactie. De opbouw en het schrijfwerk ontstonden gezamenlijk. Hans Visser heeft een groot deel van de tekst zelf aangepast en stuurde de argumentatie, stijl en toon. ChatGPT leverde conceptteksten, redactievoorstellen, de verklarende illustraties en de documentopmaak. Visser voerde de uiteindelijke inhoudelijke controle uit en draagt de redactionele eindverantwoordelijkheid.

Verantwoording. AI-hulpmiddelen zijn geen zelfstandig bewijs voor de onderzoeksclaims. Die moeten te controleren zijn aan de hand van de aangehaalde publicaties en berekeningen. De genoemde werkwijze is gebaseerd op de opgave van de auteur, niet op een volledig overzicht van toolaanroepen. De 67 effectgroottes uit Zhang et al. zijn niet allemaal onafhankelijk opnieuw geschat.