The impact of automated writing evaluation on writing gains

Bart Deygers, Liisa Buelens, David Chan, Laura Schildt, Amaury Van Parys, and Marieke Vanbuel

This study examines the effectiveness of ChatGPT in improving English writing skills for English as a foreign language (EFL) students over a nine-week intervention. Specifically, it explores ChatGPT's impact on syntactic and lexical complexity—key dimensions of writing development. The study compares two groups: one receiving teacher feedback alone and another using ChatGPT alongside teacher feedback. Participants included 105 first-year university students in Belgium, divided into these two groups. Results indicate that while ChatGPT significantly affects text length and complexity, it does not necessarily lead to sustained writing improvement. This finding is crucial for researchers and educators integrating artificial intelligence (AI) tools with traditional teaching, as it refines current insights into generative AI's educational potential. The study underscores the need for balanced approaches in writing instruction, where AI complements rather than replaces human feedback, ensuring that short-term benefits translate into long-term writing proficiency for EFL learners.

Key words: feedback, automated writing evaluation, artificial intelligence, writing development

Introduction

On November 30, 2022, ChatGPT became accessible to the public, and it was greeted with a mix of acclaim, optimism, and skepticism, both in society at large and in the context of language teaching (Ahn, Lee, and Son 2024). Generative artificial intelligence (AI) such as ChatGPT has been described as a tool that enables hard-to-detect cheating or tempts both students and teachers to take shortcuts in learning and teaching (Ahn, Lee, and Son 2024). Other voices have pointed out the potential benefits of using generative AI in education, including generating individualized test prompts, reducing the burden of error correction for teachers, and offering

Advance Access publication 8 May 2025

real-time grammatical assistance (Ahn, Lee, and Son 2024). These last two points are manifestations of automated writing evaluation (AWE), which refers to computer-based evaluation of written prose. Automated writing evaluation can refer to automated scoring or to automated feedback. In this paper, we use AWE exclusively to refer to evaluation, not to scoring. Scoring refers to assigning a numerical or categorical value to a language learner's performance based on predefined criteria, while evaluation involves providing qualitative feedback on various aspects of writing, such as coherence, organization, and complexity (Weigle 2002).

To date, a considerable amount of research has focused on the potential of generative AI to increase learner autonomy and reduce teacher workload (Fleckenstein, Liebenow, and Meyer 2023). It has also been noted that computer-assisted text revision can lead to improved writing products (Berthele and Udry 2022), but little research has focused on the impact of ChatGPT on longitudinal writing gains as an AWE tool over time. Simply put, we do not know to what extent ChatGPT allows for sustainable writing gains that are maintained when access to ChatGPT is impossible. For that reason, in this paper, we focus on the longitudinal gains in students' writing performance using ChatGPT as a writing evaluation tool. We use improvements in syntactic and lexical complexity as key indicators of writing development over time (Deygers, Vanbuel, and Ute 2022).

AWE in educational settings

Numerous studies have shown that, individual and contextual variables notwithstanding, corrective feedback (CF) is associated with improved writing (Ferris 2012). Traditionally, CF is provided by teachers or peers, but more recently AWE has become widely accessible, as a stand-alone tool or as a supplement to human feedback. As the effectiveness of feedback is linked to timeliness and specificity, the potential benefits of incorporating AWE in the writing class are obvious. Language learners who have access to AWE-generated feedback might be able to benefit from quick and contextualized feedback, in addition to feedback from the teacher.

Research has shown that AWE can impact both the writing process by encouraging more frequent revisions, and the writing product by improving overall writing accuracy (Dikli and Bleyle 2014; Link, Mehrzad, and Rahimi 2022). Studies have also explored the accuracy of AI-generated feedback, with promising results. Depending on training modalities, advanced generative AI models such as ChatGPT that have been trained on a specific scoring model can offer reliable and consistent results (Escalante, Pack, and Barrett 2023). Different iterations of ChatGPT, starting with version 3.5, have been found to offer reliable and consistent scoring, and have shown potential for AWE (Mizumoto and Egushi 2023). Researchers have examined not only the accuracy of scores, but also the quality of the AWE feedback. Steiss *et al.* (2024) compared the feedback of sixteen teachers on 200 essays to feedback provided by ChatGPT (v.3.5). They found that while human feedback tended to be more accurate, clear, and reliable, especially for content and structure, AI-generated feedback was relatively close to human feedback. Interestingly, the authors noted that the quality of the AI-powered feedback seemed inversely proportionate to the quality of the text, with more proficient writers receiving comparatively less high-quality feedback.

Since AWE has the potential to provide effective feedback, a growing body of literature has examined the impact of AWE on writing. A meta-analysis by Fleckenstein, Liebenow, and Meyer (2023) showed medium effects ($g = 0.55$), indicating that AWE tools can positively influence student writing performance. However, the authors emphasized that the effectiveness of AWE feedback varies significantly depending on several factors. For instance, longer AWE interventions resulted in stronger effects than shorter ones, and university students made slightly larger AWE-induced gains than secondary-educated learners. Additionally, L2 learners tended to benefit more from AWE feedback than L1 learners. Despite these promising findings, the authors noted substantial heterogeneity in research contexts and operationalizations, suggesting that AWE feedback may not be universally effective and should be tailored to the specific needs of students and the educational context.

A well-known and widely used AWE tool is ChatGPT, which can provide feedback on multiple aspects of writing, including grammar, vocabulary, syntax, organization, and content development. Grammatical feedback in ChatGPT3 (the version used in the current study) focused on correctness and fluency rather than complexity and could provide an explanation for suggested corrections (Mizumoto and Eguchi 2023). Lexical feedback focused on word choice and lexical diversity, though perhaps with less context awareness than current versions (Escalante, Pack, and Barrett 2023). At the level of text organization, ChatGPT3 was less adept at providing text-level feedback.

Research indicates that when used as an AWE tool, ChatGPT3 can lead to short-term gains in grammatical accuracy and lexical variation (Steiss *et al.* 2024). However, long-term improvements, particularly in syntactic complexity and organizational structure, can be more challenging. Indeed, in a study among L1 learners, Escalante, Pack, and Barrett (2023) explored the efficacy of AI-generated feedback, showing that AI feedback from ChatGPT did not significantly differ from human tutor feedback in terms of improving writing gains. This study also found an almost even split in student preferences for AI versus human feedback, with some students valuing the engagement and interaction of human feedback, and others appreciating the clarity and availability of AI-generated feedback. These findings reinforce the notion that in an educational setting, AWE is likely most effective in tandem with human feedback (Wang 2015). Other studies have taken a more targeted approach to writing gains and focused on the impact of AWE on lexis and syntax. Real-time use of AWE has been shown to positively impact lexical complexity in the writing of adolescent and adult L2 learners (Berthele and Udry 2022; Link, Mehrzad, and Rahimi 2022). The findings related to syntactic complexity are less clear, with some studies reporting zero change following digital tool usage (Berthele 2025).

In sum, this literature review shows that research on the impact of generative AI-based AWE has yielded promising, yet inconclusive findings related to writing gains. Most of the existing research was based on short interventions and typically focused on redrafting processes rather than on tracing writing longitudinal gains in new pieces of writing. As such, we still lack vital information on the impact of AWE in naturalistic settings

(Link, Mehrzad, and Rahimi 2022), using repeated writing tasks instead of revisions.

Research questions

To address the research gaps and provide useful tools for teachers seeking to include AI in their feedback practice, this study focuses on the following research question: Does the use of ChatGPT during a nine-week intervention in EFL writing classes lead to measurable differences regarding syntactic and lexical complexity?

Method

Design

The data for this study were collected between February and May 2023; just months after the initial launch of ChatGPT3. In this quasi-experimental design, students were randomly assigned to one of two conditions. Participants in the experimental condition were able to use ChatGPT for AWE purposes while writing a text in class. Similar to the students in the control condition, the experimental group also had access to teacher feedback during writing sessions. Students in the control condition only had access to teacher feedback. Table 1 shows the design of the study by participant group.

The intervention took nine weeks in total. In week 1, all students wrote a 300-word text about an inspiring celebrity dinner companion. This text was hand-written with no writing assistance offered, and it served as a control variable for participants' writing ability. In weeks 2, 3, and 4, all students received the same writing instruction, but the experimental group also did an online course about using ChatGPT for AWE purposes and prompt engineering. Students in the experimental group were explicitly instructed to use ChatGPT for various forms of feedback, including structural, grammatical, and lexical support. The learning path had four sessions with a different focus: the nature and functionalities of ChatGPT, using ChatGPT for feedback on structure and flow, using ChatGPT for feedback on grammar and vocabulary, and using ChatGPT to generate arguments and ideas. Each session started with a five-minute video tutorial, followed by exercises. Students were instructed in prompt engineering to receive corrections and feedback to gain insights rather than simply implement improvements suggested by ChatGPT.

In weeks 6 and 8, all students had the benefit of teacher feedback while writing, but students in the experimental group also had access to ChatGPT for AWE. These students also filled out a survey on their use of ChatGPT

	Instruction		**Feedback while writing**	
	Experimental	**Control**	**Experimental**	**Control**
Week 1	Writing: Baseline		None	None
Weeks 2-4	Regular instruction ChatGPT module		Teacher + ChatGPT3	Teacher
Week 5	Writing practice		Teacher + ChatGPT3	Teacher
Week 6	Writing: Essay 1		Teacher + ChatGPT3	Teacher
Week 7	Feedback essay 1		Teacher + ChatGPT3	Teacher
Week 8	Writing: Essay 2		Teacher + ChatGPT3	Teacher
Week 9	Writing: Essay 3		None	None

TABLE 1
Study design

immediately after writing. The reported use of ChatGPT is not a research question, but it will be used in the discussion to help explain some of the results. These students completed a survey immediately after writing, providing insight into their AWE usage. In week 9 all students wrote another essay without access to ChatGPT. All writing was done under controlled conditions to ensure that students complied with the conditions of the study. After week 9, all students were given access to the ChatGPT learning path.

Participants

The participants in this study were first-year students at a major Belgian University. All students were enrolled in the English translation program, and none had taken an English writing composition course before. Of the 105 participants, sixty-six were assigned to the experimental condition. The students were divided into four groups, each with a different instructor, but instructors used the same feedback model and standardized teaching material. Most participants were female (n = 79) and one identified as non-binary. The average age of the respondents was nineteen years (SD = 1.5). Both groups considered themselves equally unfamiliar with ChatGPT at the start of the project (1.8 on a five-point scale in both groups). There were no significant differences between the experimental group and the control group for such variables as gender, age, writing enjoyment, or familiarity with ChatGPT.

Students in the experimental condition also completed a short survey on their use of ChatGPT immediately after writing in weeks 6 and 8. The survey contained multiple-choice and open-ended questions about the way in which students used ChatGPT feedback. Students were asked to indicate whether they used ChatGPT to: (1) improve text structure and coherence, (2) receive grammar and spelling corrections, (3) enhance vocabulary and lexical variety, (4) develop ideas and content, or (5) check general fluency and style. This survey was not the primary research instrument in the study but provided supplementary insights that were later used to interpret the results.

Data and analysis

To date, AWE research has often been based on short interventions with a focus on redrafting processes rather than on tracing writing longitudinal gains in new pieces of writing. As such, we still lack vital information on the impact of AWE in naturalistic settings. Therefore, to avoid the experiment becoming a rewriting endeavor, each writing assignment had a different focus (Essay 1: UK national newspapers, Essay 2: mental health, Essay 3: privacy online). To minimize the impact of topic familiarity on the writing product, we gave all students a factsheet on the essay topic. This factsheet provided background information and information presented in different modalities (text, quotations, tables, charts). The expository essays were written during regular classes, and all students had a maximum of ninety minutes to complete the writing task.

The combined corpus contained 373 essays (161,532 words). Given that the aim of the study was to trace lexical and syntactic gains, we used Natural Language Processing tools designed to conduct refined linguistic analyses in these domains. For an analysis of lexical sophistication, we used TAALES (Tool for the Automatic analysis of Lexical Sophistication; Kyle and Crossley 2015; Kyle, Crossley, and Berger 2018)—a software package that quantifies over 400 indices of lexical complexity. For syntactic analysis, we relied on TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity; Kyle 2016), which evaluates various indices associated with syntactic development. It includes traditional measures of syntactic complexity, such as the average length of T-units, as well as detailed indices for phrasal (*e.g.*, the number of adjectives per noun phrase) and clausal complexity (*e.g.*, the number of adverbials per clause). Prior to processing, all texts were cleaned and formatted in line with the requirements of TAASSC and TAALES.

We measured four variables of syntactic complexity and of lexical complexity (Table 2). The syntactic complexity measures provide insight into the structural sophistication of sentences, whereas the lexical complexity measures focus on vocabulary sophistication, diversity, and frequency. By analyzing these measures, we can trace linguistic development over time and across groups. As an indicator of fluency, we measured the number of words written in the available ninety-minute time span.

After exploring data plots and descriptive statistics, we used a multilevel linear regression model to determine the impact of the condition and time on these outcome variables. The model specified was: *outcome variable ~ 1 + week + (1 + week|ID), allowing the intercept (*i.e.*, the baseline score)

Syntactic complexity

Variable	Description	Explanation
Mean length of sentence (MLS)	Overall sentence complexity	Longer sentences often signal higher complexity.
Clauses/sentence (Cl/S)	Clausal subordination and coordination	More clauses signal higher complexity
Mean length of T-unit (MLTU)	Overall T-unit complexity	A T-unit is defined as an independent clause plus any attached dependent clauses or phrases. A higher MLTU means longer, more complex structures.
Mean length of clause (MLC)	Elaboration at clause level	Longer clauses signal higher complexity

Lexical complexity

Variable	Description	Explanation
Age of acquisition (AoA)	Age of acquisition for all words	Measures how old people are when they typically learn certain words. Lower AoA suggests simpler, more familiar vocabulary.
Concreteness	Concreteness ratings for all words	Measures how concrete the main content words are. More concrete content words suggest a simpler text.
Frequency	Word frequency based on SUBTLEXus corpus.	Measures how often function words appear in a text. More frequent function words suggest less complex language use.
Academic words	Prevalence of academic words	More academic words suggest increased use of academic lexis.

TABLE 2
Complexity variables used

and the slope (*i.e.*, the effect of time/week) to vary between students. In the analysis, we controlled for initial ability in week 1 to ensure that any observed effects are attributable to the intervention itself rather than initial differences in student abilities. To be clear, we did not directly compare the baseline essay with the essay written in W9, but the complexity measures from the baseline essay were used as control variables. The analysis was conducted in R (4.1.1) using the *lme4* (Bates *et al.* 2015), *psych* (Revelle 2018), *ggplot2* (Wickham 2016), and *sjPlot* (Lüdecke 2024) packages.

Results

A first glance at the word trends during the intervention (Figure 1) shows that the experimental group consistently wrote longer texts throughout the nine-week period. During the ChatGPT experiment, the difference between the groups is visible, with students who have access to ChatGPT writing significantly longer texts, with large effect sizes. When the experimental group loses access to ChatGPT, the difference in word count evens out and loses significance (week 6: $W = 480.5$, $p = .001$, $R_{rb} = -.5$; week 8: $W = 681$, $p = .01$, $R_{rb} = -.31$; week 9: $W = 944.5$, $p = .9$, $R_{rb} = -.01$).

Figure 2 shows rather parallel trendlines for the two groups, but some significant differences can be discerned during the intervention. Students in the experimental group wrote significantly longer sentences (W8: $t(92) = -2.25$, $p = .03$, $d = .22$) and clauses (W6: $t(91) = -2.3$, $p = .02$, $d = -.5$; W8: $t(92) = -3.04$, $p = .003$, $d = -.65$), their vocabulary showed a higher age of acquisition (W6: $t(90) = -3.99$, $p = .001$, $d = -.8$), their texts contained a higher concentration of academic lexis (W6: $t(90) = -2.7$, $p = .001$, $d = 0.2$), and low-frequency words (W6: $t(90) = -2.2$, $p = .03$, $d = .48$). However, as foreseen in the design of the study, for the final writing task in week 9, students in the experimental group completed their texts without access to ChatGPT.

Table 3 summarizes the multilevel analysis results, providing insights into the effects of time, condition, and their interaction on syntactic and lexical complexity variables. We found a significant main effect of time across all syntactic and lexical complexity measures, indicating that both groups showed improvement that can be attributed to instruction. However, the condition (*i.e.*, access to ChatGPT) did not have a statistically

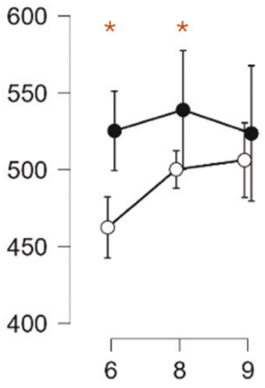

FIGURE 1
Word count changes over time.
Note. ○ control | ● experimental. Note. * indicates significant difference.

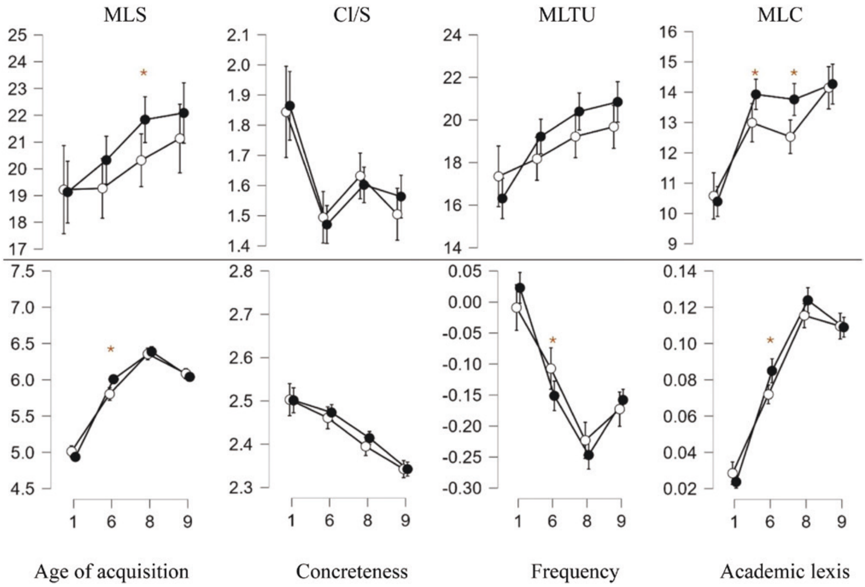

FIGURE 2
Syntactic and lexical changes over time.
Note. ○ control | ●
experimental.

	MLS		Cl/S		MLTU		MLC	
	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*
Intercept	18.69	.001	1.806	.001	16.77	001	12.84	.001
Week	0.670	.002	–0.07	.001	0.770	.001	–.46	.001
Condition	–0.04	ns	–0.03	ns	–1.04	ns	–.26	ns
Week*Cond.	.313	ns	0.012	ns	.640	.009	–.06	.04

	AoA		Concreteness		Frequency		Academic words	
	Est.	p	Est.	p	Est.	p	Est.	p
Intercept	4.78	.000	2.51	.000	0.061	.074	–0.005	0.481
Week	0.16	.000	–0.02	.000	–0.024	.000	0.011	.000
Condition	0.048	.55	0.018	.44	–0.003	.901	0.010	.109
Week*Cond.	0.008	.44	0.001	.722	–0.004	.135	0.0010	.16

TABLE 3
Impact of time and condition on syntactic and lexical variables

significant main effect on any of the variables, suggesting that the use of ChatGPT alone did not contribute substantially to the gains observed. Two statistically significant interaction effects were found for syntactic complexity measures (mean length of T-unit [MLTU] and mean length of clause [MLC]). These findings suggest that the combination of traditional instruction and ChatGPT-supported feedback may have provided additional support for syntactic complexity development specifically, as seen in the gains in MLTU and MLC.

In contrast, for lexical complexity variables, no significant interaction effects were observed, indicating that improvements in vocabulary sophistication and diversity were likely due to instructional time rather than the use of

ChatGPT. These results suggest that while ChatGPT might bolster certain aspects of syntactic complexity when used alongside instruction, its impact on lexical development appears limited within the scope of this study.

The post-writing survey provided insights which showed a small difference in ChatGPT use among the experimental group. The mean usage—as measured by a 5-point Likert scale—in week 6 was slightly higher (4.86, SD = 1.72) than in week 8 (4.31, SD = 1.89). Seventy-two students increased their ChatGPT use over time and eighty-eight students maintained a consistent level of engagement, but thirty-six students reported a reduced usage. The reported use data also show that students primarily used ChatGPT for feedback rather than for generating full sentences or paragraphs. Students requested primarily grammatical and lexical feedback and looked for idiomatic phrasing. The reported use suggests that students looked to ChatGPT as a scaffolding tool, rather than as a text generator.

Discussion & conclusion

This study aimed to investigate whether integrating ChatGPT as an AWE tool in English as a Foreign Language (EFL) writing classes could enhance students' writing development in terms of syntactic and lexical complexity. The findings reveal that ChatGPT usage had a notable impact on certain aspects of writing quality during the intervention. Access to ChatGPT significantly contributed to increased text length, sentence and clause length, as well as select indicators of lexical complexity. These findings align with earlier studies that highlighted the impact of the real-time use of ChatGPT and other AWE tools on lexical complexity (Berthele and Udry 2022; Link, Mehrzad, and Rahimi 2022).

To assess the sustained impact of ChatGPT on writing skills, students completed a final writing task in week 9 without access to the tool. This allowed us to determine whether gains observed during the intervention persisted when students wrote independently. Analyses of these performances showed overall syntactic and lexical gains in both the experimental and control groups. However, we did not find a statistically significant effect of ChatGPT on writing gains. Instead, the primary factor associated with gains was instructional time. We did find significant interaction effects between instructional time and the use of real-time AWE in MLTU and MLC. This seems to confirm that the combination of traditional instruction and ChatGPT-supported feedback may have provided additional support for syntactic complexity development (see Escalante, Pack, and Barrett 2023).

An explanation for why ChatGPT did not produce significant, sustained lexical complexity gains over time might be that students made limited use of ChatGPT for vocabulary support. A brief survey that students completed after each ChatGPT session on how ChatGPT was used revealed a relatively low usage of ChatGPT for lexical feedback specifically. Most students saw ChatGPT as a tool for content, structural, and syntactic feedback. Perhaps students needed clearer guidance on using ChatGPT more effectively for vocabulary development. These findings resonate with previous research, which has shown that while AWE tools can aid writing development, their impact may not be sustained and may be limited to specific linguistic dimensions. For instance, studies have shown that AWE feedback can positively affect aspects of syntactic complexity (Dikli and Bleyle 2014) but that lexical complexity gains are less likely unless targeted interventions are in place (Berthele and Udry 2022). Additionally, recent research sheds doubt on the effectiveness of ChatGPT as a vocabulary-learning tool (Berthele 2025).

Our findings, together with what we know from earlier work, seem to suggest that ChatGPT can benefit EFL learning when used as a supplement to teacher instruction and feedback (Escalante, Pack, and Barrett 2023). Teachers who invest time in training students on effective prompt engineering and specific AWE functions may find ChatGPT a valuable resource in the writing classroom. However, it is essential to not overestimate its impact on lexical gains without targeted, structured use, and to monitor how ChatGPT is used.

There are limitations to consider when interpreting these results. First, the intervention period—while longer than most studies in this area of research—might not be sufficient to capture long-term writing gains. Additionally, our focus on syntactic and lexical complexity measures excludes other critical aspects of writing, such as content and organization, meaning that we did not cover the full construct of writing in this study. Finally, while we did our utmost to provide a controlled classroom setting, this study was conducted in real-world conditions, meaning that we could not fully control how students used ChatGPT.

Final version received February 2025

References

Ahn, J., **J. Lee**, and **M. Son**. 2024. 'ChatGPT in ELT: Disruptor? Or Well-Trained Teaching Assistant?' *ELT Journal* 78(3):345–55.

Bates, D., **M. Mächler**, **B. Bolker**, and **S. Walker**. 2015. 'Fitting linear mixed-effects models using lme4'. *Journal of Statistical Software* 67(1): 1–48.

Berthele, R. 2025. 'Apprendre des Mots en FLE avec ChatGPT'. *Babylonia Journal of Language Education* 1(1): 12–17. https://doi.org/10.55393/babylonia.v1i.504

Berthele, R., and **I. Udry**. 2022. '*Investigating the impact of digital tools on written second language output' (Conference presentation)*. *EuroSLA Conference*, Fribourg, Switzerland.

Deygers, B., **M. Vanbuel**, and **K. Ute**. 2022. 'Can L2 Course Duration Compensate for the Impact of Demographic and Educational Background Variables on Second Language Writing Development?' *System* 109: 102864.

Dikli, S., and **S. Bleyle**. 2014. 'Automated Essay Scoring Feedback for Second Language Writers: How does it Compare to Instructor Feedback?' *Assessing Writing* 22: 1–17.

Escalante, J., **A. Pack**, and **A. Barrett**. 2023. 'AI-Generated Feedback on Writing: Insights into Efficacy and ENL Student Preference'. *International Journal of Educational Technology in Higher Education* 20(1): 57–77.

Ferris, D. 2012. 'Written Corrective Feedback in Second Language Acquisition and Writing Studies'. *Language Teaching* 45(4): 446–59.

Fleckenstein, J., **L. Liebenow**, and **J. Meyer**. 2023. 'Automated Feedback and Writing: a Multi-level Meta-analysis of Effects on Students' Performance'. *Frontiers in Artificial Intelligence* 6.

Kyle, K. 2016. '*Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication'* (Doctoral Dissertation, Georgia State University).

Kyle, K., and **S. A. Crossley**. 2015. 'Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application'. *TESOL Quarterly* 49(4): 757–86.

Kyle, K., **S. A. Crossley**, and **C. Berger**. 2018. 'The Tool for the Analysis of Lexical Sophistication (TAALES): Version 2.0'. *Behavior Research Methods* 50(3): 1030–46.

Link, S., **M. Mehrzad**, and **M. Rahimi**. 2022. 'Impact of Automated Writing Evaluation on Teacher Feedback, Student Revision, and Writing Improvement'. *Computer Assisted Language Learning* 35(4): 605–34.

Lüdecke, D. 2024. *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.17, https://CRAN.R-project.org/package=sjPlot

Mizumoto, A., and **M. Eguchi**. 2023. 'Exploring the Potential of Using an AI Language Model for

Automated Essay Scoring'. *Research Methods in Applied Linguistics* 2(2): 100050.

Revelle, W. 2018. *psych: Procedures for Personality and Psychological Research*. Northwestern University.

Steiss, J., **T. Tate**, **S. Graham**, **J. Cruz**, **M. Hebert**, **J. Wang**, **Y. Moon**, **W. Tseng**, **M. Warschauer**, and **C. B. Olson**. 2024. 'Comparing the Quality of Human and ChatGPT Feedback of Students' Writing'. *Learning and Instruction* 91: 101894.

Wang, P. -l. 2015. 'Effects of an automated writing evaluation program: Student experiences and perceptions'. *Electronic Journal of Foreign Language Teaching* 12(1): 79–100.

Weigle, S. C. 2002. *Assessing Writing*. Cambridge, Cambridge University Press.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

The authors

Bart Deygers is an associate professor at Ghent University (Belgium), where he specializes in second language didactics, language testing and assessment, and language policy.

Email: Bart.Deygers@ugent.be

Liisa Buelens is a lecturer at Ghent University (Belgium), where they teach English language proficiency (ESL/EFL) with a specialization in grammar to literature and (applied) linguistics students.

David Chan is a lecturer at Ghent University (Belgium) who teaches academic writing, translation, and English literature.

Laura Schildt is a doctoral researcher at Ghent University (Belgium). Her areas of interest include migration policy, language testing, and the role of language experts in policymaking.

Amaury Van Parys is a PhD candidate in Second Language Acquisition at Ghent University (Belgium). His research interests include vocabulary, linguistic complexity, and reading comprehension.

Marieke Vanbuel works as a postdoctoral researcher at Ghent University (Belgium). She focuses on instructed second language acquisition and language policy.