Effects of a ChatGPT-Assisted Writing Intervention on the Productive Writing Skills of Students With Learning Disabilities: A Single-Case Experimental Design Study

Susanne Hoff, Matthias Grünke, Janine Bracht, and Anne Barwasser
University of Cologne, Germany

Abstract

This single-case design study explored the impact of AI-supported writing instruction for students with learning disabilities (LD). We examined how ChatGPT-generated feedback embedded within a structured Self-Regulated Strategy Development (SRSD)-based writing intervention affected students' writing productivity, as measured by Total Words Written (TWW), and text quality, as measured by the Teacher Evaluation of Story Elements rubric (TESE). Drawing on the Hayes and Flower (1980) writing process model and the SRSD framework, the intervention provided scaffolded task-level feedback using standardized prompts and iterative revision cycles. The effects were evaluated across alternating baseline, intervention, and follow-up phases using visual inspection and non-parametric overlap indices. Results indicated positive effects on both TWW and TESE, with consistent level effects across intervention phases. Moreover, students gradually internalized feedback routines, indicating the development of metacognitive strategies. The findings highlight the pedagogical potential and limitations of AI-mediated scaffolding in inclusive education and call for critical reflection on its implementation.

Keywords: ChatGPT, self-regulated strategy development, learning disabilities, AI-generated feedback, inclusive writing instruction, single-case design, textuality

Introduction

Writing is widely recognized as a cornerstone of academic achievement, social engagement, and vocational integration (Graham & Harris, 2018; Kellogg & Whiteford, 2009). High levels of writing proficiency allow individuals to articulate ideas, engage in critical thinking, and communicate effectively across varied contexts (Graham & Perin, 2007). Writing proficiency is a complex, multicomponent skill involving the interplay of subprocesses, including transcription, planning, text generation, and revision, regulated by metacognitive and motivational processes (Berninger & Winn, 2006; Hayes, 2012). Developing this multifaceted competence poses considerable challenges, particularly for students with learning disabilities (LD), who often exhibit persistent difficulties across both lower-order (e.g., spelling, handwriting) and higher-order (e.g., text organization, idea generation, revision) processes.

Educational Challenges in Writing for Students With LD

Students with LD – broadly defined as learners who experience specific, persistent difficulties in acquiring conventional reading, writing, or mathematics skills that cannot be attributed to intellectual disability, sensory impairment, or inadequate instruction (UNESCO, 2020) – face enduring academic challenges that typically manifest across the school years, often as difficulties in working memory, self-regulation, and linguistic planning, all of which are functions central to writing development (Swanson & Sachse-Lee, 2001;

Hoff, S., Grünke, M., Bracht, J., & Barwasser, A. (2026). Effects of a ChatGPT-Assisted Writing Intervention on the Productive Writing Skills of Students With Learning Disabilities: A Single-Case Experimental Design Study. *International Journal for Research in Learning Disabilities,* XX. https://doi.org/10.28987/ijrld.9.1.XX

Troia, 2011). In Germany, this population represents approximately 3% of all school-aged students (European Agency for Special Needs and Inclusive Education, 2020), frequently categorized under broader special education frameworks. Notably, these learners tend to produce shorter, less cohesive, and structurally weaker texts than their peers (Graham & Perin, 2007), and they struggle to revise and plan their writing independently (Graham et al., 2014).

The importance of addressing these deficits is underscored by large-scale assessments. For example, according to the 2022 PISA report, 25.5% of German adolescents are classified as poor readers, and similar patterns are observable in writing outcomes (Organisation for Economic Co-operation and Development [OECD], 2022). Given the reciprocal relationship between reading and writing (Graham et al., 2025), difficulties in one domain frequently reinforce problems in the other. Furthermore, data from the IQB Education Trends show that growing numbers of students fail to meet minimal literacy standards, reinforcing the urgency of evidence-based, targeted interventions, particularly for writing, which has received comparatively less attention than reading (Stanat et al., 2022).

Theoretical Foundations of Writing Instruction

The present study was grounded in two complementary theoretical models. First, the Hayes and Flower (1980) cognitive process model conceptualizes writing as a recursive, goal-directed activity comprising planning, translating, and reviewing. Second, the Self-Regulated Strategy Development (SRSD) model (Graham & Harris, 2005) provides a pedagogical framework that emphasizes explicit instruction, modeling, self-monitoring, and motivation. Both frameworks underscore the need for structured, scaffolded, and individualized feedback –conditions that are often difficult to realize in conventional classroom settings for learners with LD.

Beyond instructional frameworks, formative feedback – defined as information used to improve learning rather than to evaluate it – plays a central role in writing development. Hattie and Timperley (2007) conceptualized feedback along three levels: task level (information about correctness), process level (strategies for improvement), and self-regulation level (supporting independent monitoring and control). These differentiated feedback types contribute uniquely to the development of writing competence. Particularly for students with LD, formative and specific feedback has been shown to enhance metacognitive engagement, foster self-efficacy, and improve text quality (Shute, 2008).

Given the central role of feedback in writing development, the question arises whether recent advances in artificial intelligence (AI), particularly large language models, can provide feedback that meets these pedagogical requirements in a scalable and individualized manner by simulating core features of both teacher and peer feedback – especially when scaffolded and goal-directed – and contribute to students' internal feedback loops. This potential is particularly relevant in SRSD-based instruction, which explicitly incorporates feedback as a mechanism for fostering strategic behavior, gradual autonomy, and reflective thinking.

AI and Its Pedagogical Promise

Recent advances in educational technology, especially in the field of generative AI, offer new possibilities for tailoring instruction to diverse learner needs. Thus, large language models such as ChatGPT can generate coherent, context-sensitive text and simulate dialogic interactions. A recent systematic review found that ChatGPT is increasingly adopted in K–12 educational contexts, particularly for lesson planning, material creation, and differentiated instruction (Zhang & Tur, 2024). While the review highlights both opportunities and risks of classroom use of ChatGPT, such as concerns about academic integrity and the need for co-designed implementation, it does not address the specific challenges faced by students with diagnosed learning disabilities.

When purposefully embedded into instruction, tools like ChatGPT may support writing development of this population of students by offering real-time feedback, prompting elaboration, and suggesting revisions – functions that correspond directly to the cognitive demands outlined in the Hayes-Flower and SRSD frameworks (Hayes & Flower, 1980; Mohammed & Khalid, 2025; Ya'u & Mohammed, 2025; Yan et al., 2024). These feedback functions are not merely technical affordances but can be pedagogically framed as dialogic and empowering when intentionally integrated into writing instruction. For example, Dong (2024) introduced a ChatGPT Feedback Engagement Framework that conceptualizes AI-generated responses not as automated output but as pedagogically meaningful dialogic interactions, particularly for learners who benefit from structured support, such as second-language (L2) learners or students with LD.

However, the pedagogical use of ChatGPT also raises important ethical and epistemological concerns. In an exploratory study, Price et al. (2024)

demonstrated that ChatGPT-generated feedback can reflect and reinforce stereotypical assumptions about learner groups, including those labeled as "gifted" or "special education." This finding underscores the need to critically examine the latent biases embedded in AI-generated language, particularly when engaging with vulnerable student populations such as learners with disabilities.

Unlike corrective AI tools (e.g., Grammarly, QuillBot) that focus on grammar and syntax, ChatGPT provides dialogic, content-oriented feedback that supports higher-order writing processes such as idea development, cohesion, and argument structure (Steiss et al., 2024; Yan et al., 2024; Zhai et al., 2024). Despite this potential, however, empirical research on the use of ChatGPT with students formally diagnosed with LD remains scarce. For example, a recent systematic review by Imran and Almusharraf (2023) confirmed that most studies focus on general or higher education contexts, rarely addressing the needs of marginalized learners. This is a critical gap, as students with LD often benefit disproportionately from timely, individualized, and structured feedback (Graham et al., 2025). The capacity of ChatGPT to provide low-stakes, adaptive practice, simulate peer or teacher feedback, and model effective revisions positions it as a potentially powerful tool for inclusive writing instruction, provided its implementation is guided by pedagogical intent and ongoing critical oversight.

Clarifying Productive Writing Skills

To evaluate whether AI-generated feedback can meaningfully support writing development in students with LD, it is necessary to clarify the specific construct under investigation. The present study focused on productive writing skills, a construct encompassing the capacity to generate, structure, and refine original written texts for communicative purposes. These skills differ from transcriptional or receptive language abilities, as they require active conceptualization, organization of ideas, lexical selection, syntactic planning, and genre-specific structuring (Berninger & Amtmann, 2003). For students with LD, this demands cognitive flexibility, working memory resources, and iterative feedback – areas in which they often experience marked impairments (Graham et al., 2016; Hooper et al., 2002). Yet, existing research on effective support strategies for productive writing remains sparse and lacks integration of current technological advances.

Positioning the Present Study in the Research Landscape

As mentioned, there is a noticeable research gap at the intersection of AI-supported writing instruction and special education. Existing reviews on AI in education tend to focus on general education student populations, often overlooking the differentiated pedagogical needs of students with LD (Imran & Almusharraf, 2023; Zhang & Tur, 2024). Similarly, the literature on writing instruction for students with LD rarely addresses digital or AI-mediated scaffolding. Studies exploring feedback-based writing interventions in this population remain fragmented and methodologically heterogeneous, with few employing experimental or longitudinal designs (Graham et al., 2014; Troia, 2011).

Meta-analytic evidence confirms that students with LD benefit most from writing interventions that are both explicitly structured and responsive to individual needs. In particular, the SRSD model, which combines cognitive strategy instruction, modeling, and scaffolded feedback, has demonstrated robust effects across multiple studies (Graham et al., 2014). Similarly, Swanson and Sachse-Lee (2001) highlighted the importance of adaptive instructional components that account for the cognitive diversity of this population. Despite this evidence, however, only a limited number of studies have examined whether and how technology-mediated feedback systems, such as AI-generated responses, align with these established pedagogical principles. The heterogeneity within the LD population, especially with regard to co-occurring attentional, linguistic, or executive function difficulties, further underscores the need for differentiated, context-sensitive feedback solutions.

While AI tools may not yet be fully equipped to provide such support, nevertheless, promising findings are beginning to emerge. For example, in an empirical study conducted in Saudi Arabia, Alsahli et al. (2025) found that ChatGPT facilitated learning gains among students with special educational needs, particularly when used to provide scaffolded writing support aligned with curricular goals. A comprehensive synthesis of these strands – writing development, AI-generated feedback, and inclusive pedagogy – is, therefore, both timely and necessary.

Research Aim of the Present Study

Against this backdrop, the present study investigated the effectiveness of a ChatGPT-assisted writ-

ing intervention for students with LD, implemented within a single-case ABABE design. Alternating between baseline (A) and intervention (B) phases, this design allows for within-subject comparison and causal inference (Babbie, 2020; Kazdin, 2019). The core research question was as follows: *To what extent does ChatGPT-based feedback, embedded within a structured SRSD-informed intervention, improve the productive writing skills of students with LD?*

By focusing on this underrepresented group, the study sought to contribute both to the theoretical understanding of AI as a pedagogical tool and to the practical development of inclusive instructional strategies. In line with Sustainable Development Goal 4 (UNESCO, 2020), the study also aimed to support global efforts toward equitable, high-quality education for all learners. Initial empirical evidence supports this orientation: Ibrahim and Ajlouni (2024) demonstrated that ChatGPT can foster engagement and standard-aligned progress among students in special education settings, suggesting that AI tools may contribute meaningfully to inclusive education when applied with instructional care.

Methods

The study employed a single-case experimental design (ABABE) comprising two baseline phases (A1, A2), two intervention phases (B1, B2), and a follow-up phase (E). This design enabled within-subject comparisons and was chosen for its compatibility with small, heterogeneous samples in special education research (Kazdin, 2011). The ABABE design was selected because it allows for the demonstration of experimental control through the replication of effects across two intervention phases (B1 and B2), whereas the return-to-baseline phase (A2) provides evidence that changes in the dependent variables are functionally related to the intervention rather than attributable to maturation, practice, or other confounding variables. The inclusion of a follow-up phase (E) additionally permits evaluation of the maintenance of any observed effects. Phase transitions were determined on a time-based schedule, with each phase comprising five sessions. This decision was guided by practical constraints (school schedules and participant availability) rather than data-based stability criteria. While data-based phase change rules would have strengthened experimental control, the consistent five-session structure across all phases ensured procedural comparability.

Across 21 sessions of 45 minutes each, tasks were implemented in quiet, distraction-free school environments. To ensure internal validity, baseline and intervention phases were structurally aligned in duration, task format, and procedural components. The defining distinction was the integration of ChatGPT-generated feedback during intervention phases.

Participants and Sampling Strategy

Six ninth-grade students (four male, two female; average age 15.8 years) from inclusive and special education schools in North Rhine-Westphalia, Germany, took part in the study. All participants were born in Germany, fluent in German, and formally diagnosed with LD, particularly affecting productive writing. Three had a migration background (two of Russian, one of Turkish descent). Pseudonyms were used to ensure anonymity. Detailed participant characteristics are summarized below.

All students attended Grade 9 and were aged between 15 and 17 years. Cognitive ability was not formally assessed as part of this study; however, all participants had received a formal diagnosis of LD (*Förderschwerpunkt Lernen*) through the German special education assessment system, which includes standardized cognitive and achievement testing as part of the diagnostic process. No participant had a co-occurring diagnosis of attention deficit hyperactivity disorder (ADHD) or autism spectrum disorder (ASD) according to available school records.

Participant selection followed a two-stage process: (a) teacher nomination based on observed writing difficulties and (b) screening with the Hamburg Writing Test (HSP 5–10 EK; May et al., 2023). The HSP evaluates core writing skills and was selected to confirm sufficient foundational competence for participation. The procedure ensured that only students with pronounced but modifiable writing difficulties were included. All procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki and the Belmont Report (World Medical Association, 2013). Written informed consent was obtained from the legal guardians of all participating students, and assent was obtained from the students themselves.

Procedures

Baseline Phases (A1, A2)

At baseline, students engaged in mindfulness activities (e.g., focused breathing, sensory awareness), followed by unscaffolded reflective writing. These simple, cognitively undemanding tasks (Sweller, 1988) were designed to establish unbiased

baselines of spontaneous writing behavior. Structural consistency across sessions enabled reliable phase comparisons.

Intervention Phases (B1, B2)

Structured writing tasks based on Hayes and Flower's (1980) process model (planning, drafting, revising) were introduced. Each session began with a writing prompt to elicit creative or reflective responses. Students drafted a text and then received targeted feedback from ChatGPT based on six pretested standardized prompts derived from key dimensions of the writing process as conceptualized by Hayes and Flower (1980) and Hayes (1996); see Table 1. Feedback was documented on worksheets, with students selecting up to three actionable suggestions for revision.

To illustrate the standardized procedure and support replicability, a typical ChatGPT interaction went as follows. After responding to the prompt "Imagine you are the principal of your school for one day. What would you change and why?," one student wrote: "I would cancel homework because it is boring and students already learn enough in school." ChatGPT generated the following feedback: "Your opinion is clear. To strengthen your argument, consider adding an example or consequence. For instance, explain how students could use the extra time instead." Using the prioritization worksheet, the student selected this suggestion and revised the sentence accordingly.

This four-step cycle – prompt, drafting, AI feedback, guided revision – was applied consistently across all intervention sessions using one of six standardized ChatGPT prompts. To ensure accessibility and instructional clarity, all students worked with preformulated prompts designed to elicit targeted feedback from ChatGPT. Prompts addressed key dimensions of text quality, including goal clarity, structure, creativity, and thematic depth. All questions were framed in simplified language to support comprehension and consistency. Table 1 provides an overview of the six dimensions, the exact prompts used during the intervention, and their pedagogical purposes.

No additional teacher feedback was provided to isolate the AI's contribution to text revision.

Follow-Up Phase (E)

The follow-up session replicated the previous task format but excluded scaffolding and AI support, thereby assessing the durability and transferability of writing improvements (Horner et al., 2005).

Intervention Fidelity and Accessibility Enhancements

Treatment fidelity was monitored using post-session checklists covering procedural adherence, clarity of instruction, and material use (Kratochwill et al., 2010; O'Donnell, 2008; Scheibel et al., 2022). (The

Table 1
Dimensions and Purposes of ChatGPT-Generated Feedback Questions

Dimension	Question	Purpose
Clarity of Goal	"ChatGPT, in simple language, is my goal clear? Please suggest how I can make it better."	Helped participants evaluate and refine the communicative intent of their texts.
Sequence of Ideas	"ChatGPT, in simple language, are my ideas expressed in a good sequence? Please suggest how I can improve the order."	Aimed to improve the logical flow and coherence of narratives.
Text Structure	"ChatGPT, in simple language, is my story well structured? Please suggest improvements."	Guided participants in refining the organizational framework of their texts.
Stylistic Quality	"ChatGPT, in simple language, does my text sound good? Please suggest how to enhance it."	Addressed tone, sentence variety, and word choice to improve readability.
Originality and Creativity	"ChatGPT, in simple language, is my text original and creative? Please suggest improvements."	Encouraged participants to develop more unique and imaginative writing.
Thematic Depth	"ChatGPT, in simple language, does my essay have depth? Please suggest how I can add more."	Supported the creation of richer, more nuanced content.

Note. All prompts were preformulated by the research team and delivered via preconfigured tablet templates to ensure consistency across sessions and to reduce cognitive demands on participants. Prompts were designed in simplified language to support accessibility for students with LD. Each dimension corresponded to a theoretically grounded writing component derived from the Hayes and Flower (1980) cognitive process model and the SRSD framework (Graham & Harris, 2005). Students selected up to three feedback suggestions per session for guided revision.

following reporting follows the SCRIBE 2016 guidelines for single-case intervention studies; Tate et al., 2016). Across all 21 sessions, the average fidelity adherence rate was 94%, indicating high procedural integrity throughout the study. Initial difficulties in processing ChatGPT feedback led to two accessibility adaptations: (a) tablets featured pre-installed templates with standardized prompts, and (b) prioritization worksheets helped students identify and implement key suggestions.

Instruments and Materials

Writing Prompts and Tools

Writing tasks were based on counterbalanced prompts of comparable thematic and linguistic complexity designed to elicit expressive-reflective texts. Complexity was defined in terms of syntactic demands, vocabulary range, and abstractness of the prompt topic. For example, one representative writing prompt was: "Describe a moment when you felt proud of something you accomplished. What did you do, and how did it make you feel?"

This prompt was classified as expressive-reflective due to its personal-emotional frame, open-ended structure, and the demand for temporal and causal coherence. Differentiation across prompts was achieved by systematically varying (a) temporal perspective (past vs. hypothetical future), (b) emotional valence (positive vs. challenging experiences), and (c) required text length (short vs. extended reflection). All prompts were piloted to ensure accessibility across reading levels and cultural backgrounds.

Materials included brainstorming scaffolds, structured feedback logs, and fidelity monitoring forms. To enhance transparency and replicability, all tools and materials used in the intervention are summarized in Table 2.

Writing Assessment Tools

Two dependent variables were assessed: writing fluency and writing quality. Writing fluency was operationalized as the Total Words Written (TWW; Hosp et al., 2016), a validated metric for productivity in written expression. Writing quality was measured using an adapted version of the Teacher Evaluation of Story Elements (TESE; Troia & Graham, 2002). While originally developed for narrative writing, the TESE rubric was systematically revised for expressive-reflective text types based on the Hayes and Flower (1980) model of the writing process. The adapted rubric comprised six dimensions, each rated on a 5-point Likert scale (1 = *minimal proficiency*; 5 = *exceptional performance*). For the purposes of this study, the dimensions were operationalized as follows:

- **Goal Clarity:** Presence of a clearly identifiable communicative intention or emotional focus, independent of narrative closure.
- **Structure:** Logical organization of ideas, including thematic coherence and use of transitions to structure reflection.
- Creativity: Degree of original formulation, personal expression, and use of figurative language or imagery.
- **Linguistic Quality:** Syntactic accuracy, lexical variety, and grammatical correctness.
- **Thematic Depth:** Engagement with abstract, complex, or multi-layered aspects of the chosen theme.

Table 2
Materials and Tools Used in the Intervention

Category	Description
Assessment Instruments	The Hamburg Writing Test (HSP) was utilized to evaluate baseline writing skills, including orthography, grammar, and composition.
Digital Tools	Tablets with preconfigured ChatGPT accounts were provided for drafting and revising. Researchers monitored the accounts to ensure data integrity.
Supportive Materials	• Worksheets for brainstorming and integrating feedback • Preloaded templates with standardized ChatGPT queries • Checklists to guide session steps and ensure consistency
Documentation Tools	Treatment fidelity forms were employed to monitor adherence to the intervention protocol and document the accuracy of its implementation.

Note. All materials were developed specifically for this study and piloted prior to implementation. The Hamburg Writing Test (HSP 5–10 EK; May et al., 2023) was used exclusively for baseline screening and participant eligibility, not as an outcome measure. ChatGPT accounts were accessed via a standardized interface and monitored by the research team to ensure data integrity and prevent unintended use beyond the study protocol. (Brainstorming scaffolds and feedback logs are available from the authors upon reasonable request.)

- **Authenticity:** Evident personal involvement, emotional congruence, and subjective credibility of the text.

Both dependent variables (TWW and TESE) were assessed at every session across all phases, resulting in 21 measurement points per participant. TWW was counted immediately after each writing session, whereas TESE scores were rated independently by three trained raters within one week of each session. These revised scoring criteria reflect the genre shift from narrative to expressive-reflective writing and were aligned with both cognitive models of composition and genre-specific demands. Rating consistency was ensured through structured training procedures (see below).

Rater Training and Interrater Reliability

Three trained raters independently evaluated all texts. Training consisted of two calibration sessions using anchor texts and score norming, with consensus-building procedures in place for rating discrepancies. Anchor texts were drawn from pilot data and normed using rubric-based exemplars. Interrater agreement exceeded 85% across all dimensions; discrepancies were resolved through consensus-based discussion.

Interrater reliability was assessed using percentage agreement, calculated as the number of exact-match ratings divided by the total number of rated items. Agreement ranged from 85% to 96% across the six TESE dimensions. While percentage agreement provides a basic index of consistency, we acknowledge that it does not account for chance agreement. Future studies should supplement this with more robust indices such as Cohen's kappa or intraclass correlation coefficients (ICC) to strengthen psychometric rigor.

Data Analysis

Quantitative data (TWW and TESE scores) were analyzed using the R package scan (Wilbert & Lüke, 2022). Phase comparisons (A1→B1, B1→A2, A2→B2, A1→B2) were examined using non-overlap of all pairs (NAP; Parker et al., 2011) and percentage exceeding the median (PEM; Parker & Hagan-Burke, 2007). For the B1→A2 comparison, the decreasing = TRUE command was specified due to the expected performance drop.

Effect sizes were interpreted as follows: Weak = 0–.65; Moderate = .66–.92; and Strong = .93–1.0 Additionally, piecewise regression analysis (Manolov & Moeyaert, 2017) was used to estimate slope and level changes across phases:

$\gamma_i = \beta_0 + \beta_1(time_i – time_1) + \beta_2 phase_i + \beta_3(time_i – timen_a+1) \times phase_i$ where γ_i represents the outcome at time i, and $timen_a+1$ refers to the first time point of the intervention phase. The parameters are defined as follows: β_0 = intercept (initial performance level), β_1 = slope within the baseline phase (time trend), β_2 = level change at the onset of the intervention, and β_3 = slope change indicating the difference in trend between baseline and intervention. While the single-case design focuses on individual trajectories, aggregated trends were analyzed for exploratory insight.

Visual analysis followed the six-feature framework specified in the What Works Clearinghouse single-case designs technical documentation (Kratochwill et al., 2010). Specifically, phase comparisons were guided by evaluation of (a) level (mean performance within and across phases), (b) trend (direction and slope of data within phases), (c) variability (range and consistency of data points within phases), (d) immediacy of effect (change in level between the last data points of one phase and the first data points of the next), (e) overlap (proportion of data points in adjacent phases that share the same range), and (f) consistency of data patterns across similar phases (e.g., A1 vs. A2; B1 vs. B2). These criteria informed all interpretive conclusions reported in the results section.

Social Validity

Social validity was assessed using a nine-item questionnaire adapted from Wolf (1978) and Hurley (2012), focusing on perceived task relevance, motivational value, usability, and self-efficacy. Items included statements such as "The writing task helped me express my thoughts clearly" and "Using ChatGPT made it easier to improve my text." Participants rated their agreement on a four-point Likert scale (0 = *never* to 3 = *always*). The questionnaire was administered immediately after the final session to capture students' retrospective evaluations of the full intervention. Responses were analyzed descriptively to complement outcome data and assess learner-oriented feasibility.

Psychometric Quality Criteria

Reliability was evidenced by interrater agreement (>85%). Construct validity was supported by theoretical alignment of the adapted TESE rubric with Hayes and Flower's process model (Hayes & Flower, 1980). Content validity was ensured through expert review of writing. Specifically, two writing researchers with expertise in educational psychology and one licensed language therapist independently reviewed all prompts for cognitive appropriateness, linguistic

accessibility, and alignment with expressive-reflective writing goals. Ecological validity was considered high, given the authentic school-based implementation. Social validity, as evaluated through student perceptions, further substantiated the intervention's contextual relevance and practical feasibility. Due to the small sample size and idiographic nature of the design, generalizability remains limited. However, analytic generalization through replication logic is pursued.

Results

Total Words Written (TWW)

Visual analysis based on the six-feature framework (level, trend, variability, immediacy of effect, overlap, and consistency; Kratochwill et al., 2010) indicated that all students exhibited substantial improvements across intervention phases. Pronounced level changes were observed from baseline to intervention phases (Figure 1), with low variability within

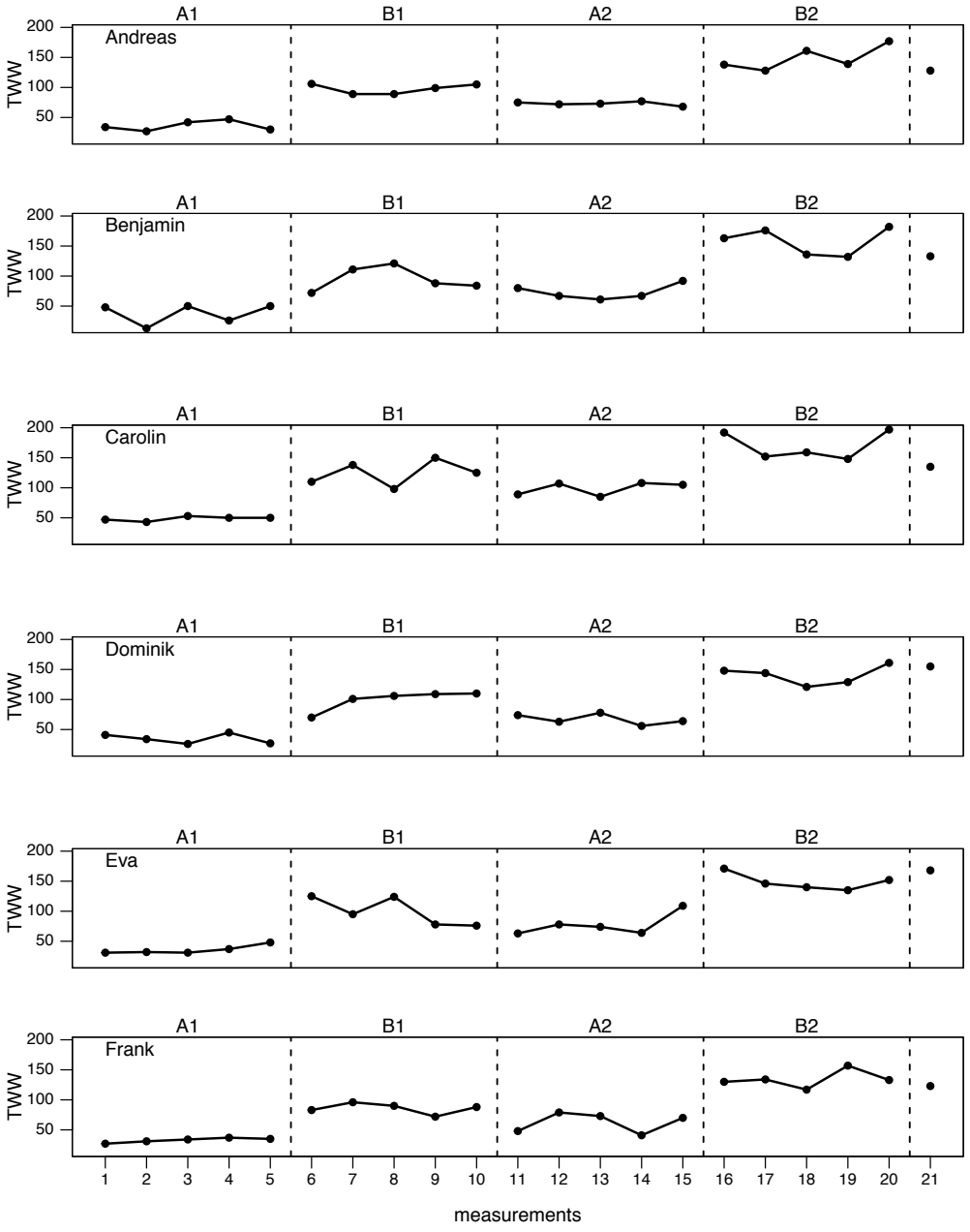

Figure 1
TWW (Total Words Written)

intervention phases and minimal overlap between adjacent A and B phases. The immediacy of effect was evident at each phase transition, as performance levels changed markedly from the last data points of the preceding phase to the first data points of the subsequent phase.

All students exhibited pronounced increases in total word output across intervention phases. Mean TWW scores rose substantially from A1 to B2 for each participant (see Table 3), with the highest individual session score increasing from 53 words in A1 to 197 words in B2. Clear level effects were observed from the respective baseline phases to the intervention phases, with a decline in scores for all students from the B1 to the A2 phase. That is, when the ChatGPT-based writing training was removed, the students' performance deteriorated. In the B2 phase, a marked improvement was again evident, accompanied by pronounced level effects. The follow-up survey revealed stable results for half of the students, while the scores for the other half showed a slight decline but remained considerably higher than in the initial baseline phases. The descriptive data (see Table 3) corroborated the findings of the visual inspection. That is, for each student, scores increased substantially across phases, with a decline from the B1 to the A2 phase. The data indicated an increase from a maximum score of 37.00 in the A1 phase to 197.00 in the B2 phase, highlighting substantial improvement.

Table 3
Descriptive Data for TWW and TESE

	Andreas	Benjamin	Carolin	Dominik	Eva	Frank
TWW						
M (A1)	36.00	37.40	48.60	34.60	35.80	32.80
SD	(8.34)	(17.00)	(3.78)	(8.39)	(7.26)	(3.90)
Max (A1)	47	50	53	45	48	37
M (B1)	97.60	95.20	124.20	99.20	99.60	85.80
SD	(8.30)	(20.19)	(20.89)	(16.69)	(23.90)	(9.01)
Max (B1)	106	121	150	110	125	96
M (A2)	73.00	73.40	98.80	67.00	77.60	62.20
SD	(3.39)	(12.50)	(10.92)	(8.89)	(18.69)	(16.66)
Max (A2)	77	92	108	78	109	79
M (B2)	148.69	157.80	169.60	140.60	148.80	134.20
SD	(19.93)	(22.83)	(23.14)	(1.82)	(13.95)	(14.45)
Max (B2)	177	182	197	161	171	157
M (E)	128	133	135	155	168	123
TESE						
M (A1)	4.20	4.40	5.60	5.60	4.20	4.40
SD	(4.27)	(4.04)	(4.56)	(5.46)	(2.17)	(4.04)
Max (A1)	10	10	10	13	6	10
M (B1)	15.60	20.80	25.20	14.40	20.40	11.60
SD	(3.21)	(3.63)	(2.86)	(3.49)	(3.51)	(1.82)
Max (B1)	19	24	29	18	25	14
M (A2)	12.20	16.00	20.20	10.20	14.80	7.40
SD	(1.92)	(3.67)	(3.27)	(3.49)	(3.90)	(2.51)
Max (A2)	14	20	24	15	20	10
M (B2)	19.40	26.20	27.60	19.80	24.60	17.40
SD	(1.52)	(4.27)	(1.67)	(1.30)	(3.10)	(1.52)
Max (B2)	22	30	29	22	29	19
M (E)	16	24	24	17	20	13

Note. Mean (*M*) and standard deviation (*SD*) values represent phase averages across five measurement points per phase (A1, B1, A2, B2) and one measurement point for the follow-up phase (E). Maximum values (Max) indicate the highest single-session score within each phase. A decline in scores from B1 to A2 was expected, as the ChatGPT-based writing intervention was withdrawn during this phase. TWW reflects writing productivity. TESE reflects text quality rated on a 5-point Likert scale.

The overlap measures revealed strong, significant effects for the nonoverlap of all pairs (NAP) in the A1 to B1 phases (100.00, $p < .01$) across all students. A similar pattern was observed for the A2 to B2 comparison as well as the A1 to B2 comparison. In contrast, the B1 to A2 comparison demonstrated that while scores decreased significantly, some overlap between phase points remained. All other phase comparisons (A1-B1, B1-A2, A2-B2) were also conducted for both TWW and TESE but revealed no additional differentiations beyond what is reported here. In every case, effect sizes remained consistently high (NAP and PEM ≥ 98.00), except for the B1-A2 comparisons, which showed medium-range effects in several students. To enhance clarity and avoid redundancy, only the results for the A1-B2 comparison are presented in Table 4.

For the percentage of exceeding medians (PEM), the results mirrored those of the NAP for each individual student. The regression analysis (see Table 5) revealed highly significant level effects from the A1 to B1 phase, with a beta coefficient of 59.367 ($p < .001$), a decline in scores from B1 to A2 (B = -30.517), and a subsequent positive level effect from A2 to B2 (B = 67.100, $p < .001$). No slope effects were detected in the dataset, as such effects were not anticipated given the strong level effects.

Teacher Evaluation of Story Elements Rubric (TESE)

The dependent variable TESE initially exhibited variable baseline data with some existing trends which can be seen through visual inspection (see Figure 2). Compared to TWW, it is worth noting the lower scaling of the variables and graphs in this context. Nevertheless, the level effects from A1 to B1, from B1 to A2, and from A2 to B2 were clearly similar to those observed for TWW. In the B2 phase, students exhibited notably stronger performance and higher text quality compared to the initial baseline phase, despite declines in the follow-up survey for all students. However, these follow-up scores remained substantially higher than the initial values.

The descriptive data (see Table 3) supported the findings of the visual analysis. That is, all students showed substantial increases in scores from A1 to B2, with a noticeable decline from B1 to A2, coinciding with the cessation of the ChatGPT-supported writing intervention. Maximum scores increased from 6.00 in the A1 phase to 30.00 in the B2 phase.

Despite the visual decline in follow-up scores, the descriptive statistics indicated that these scores remained relatively stable across students. With regard to the overlap measure NAP, strong effects were observed for all students from A1 to B1 (100.00, $p < .01$), with the exception of Benjamin, who exhibited medium effects, though these were directly at the cutoff for strong effects (92.00, $p < .05$). Similar to TWW, the B1 to A2 phase exhibited some overlap in the phase data, with all students showing a decline in performance after the intervention was removed. Strong effects were consistently observed for the A2 to B2 and A1 to B2 comparisons (98.00–100.00, $p < .01$).

The PEM analysis revealed a similar pattern to that of TWW, with strong effects across all students (see Table 4). The regression analysis further highlighted significant level effects for TESE. These included an increase from A1 to B1 (B = 12.583, $p < .001$), a decline from B1 to A2 (B = -3.783, $p < .05$), and

Table 4
Overlap Indices for TWW and TESE

	Andreas	Benjamin	Carolin	Dominik	Eva	Frank
TWW A1-B2						
NAP	100.00	100.00	100.00	100.00	100.00	100.00
(*p*)	(<.01)	(<.01)	(<.01)	(<.01)	(<.01)	(<.01)
PEM	100.00	100.00	100.00	100.00	100.00	100.00
TESE A1-B2						
NAP	100.00	100.00	100.00	100.00	100.00	100.00
(*p*)	(<.01)	(<.01)	(<.01)	(<.01)	(<.01)	(<.01)
PEM	100.00	100.00	100.00	100.00	100.00	100.00

Note. NAP = nonoverlap of all pairs; PEM = percentage of exceeding the median; TWW = Total Words Written; TESE = Teacher Evaluation of Story Elements Rubric. All additional phase comparisons (A1-B1, B1-A2, A2-B2) yielded comparable results with minimal variance. (Full data available upon request.) Effect sizes were interpreted as follows: weak = 0-.65, moderate = .66–.92, strong = .93–1.00 (Parker et al., 2011).

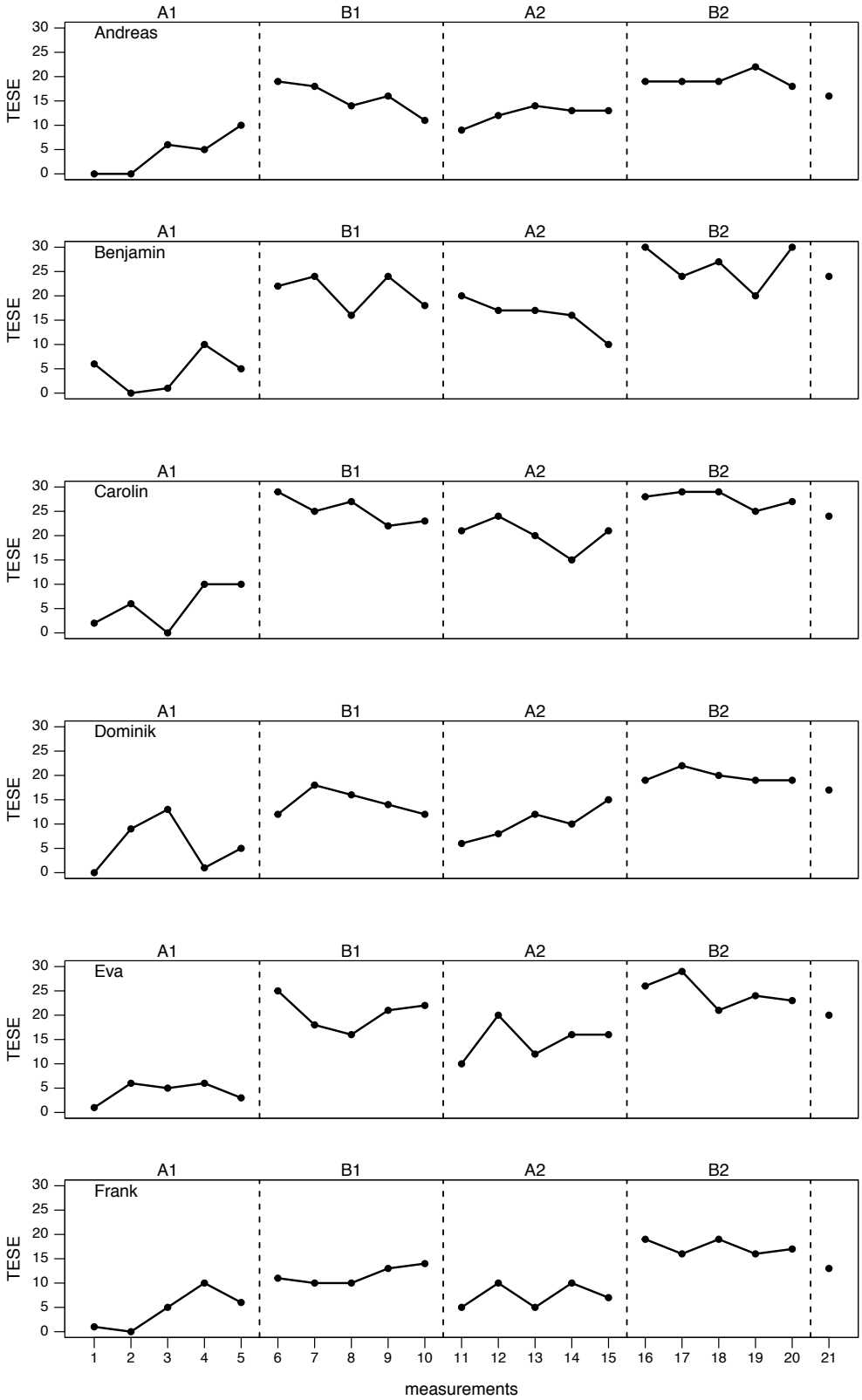

Figure 2
TESE (Teacher Evaluation of Story Elements Rubric)

another increase from A2 to B2 (B = 10.017, $p < .001$). No overall slope effects were detected (see Table 5).

Social Validity

The social validity data (see detailed data in Table 6) indicated that the majority of students enjoyed participating in the intervention. However, Dominik was the only one who reported that he did not particularly enjoy the ChatGPT-based writing intervention. He also provided the lowest ratings overall. The results from the checklist were varied – some students found it highly beneficial, while others did not find it as helpful. Additionally, students mentioned that the records did not always serve as a source of inspiration. Nevertheless, all students expressed the belief that after the intervention they were able to

Table 5
Regression Analysis Level 2 for TWW and TESE (Across All Groups)

		B	*SE*	*t*	*p*
TWW	Level A1-B1	59.367	7.505	7.910	<.001
	Level B1-A2	-30.517	7.986	-3.821	<.001
	Level A2-B2	67.100	8.362	8.024	<.001
	Slope A1-B1	-1.267	2.574	-0.492	.63
	Slope B1-A2	1.583	2.739	0.578	.57
	Slope A2-B2	-0.417	2.868	-0.145	.89
TESE	Level A1-B1	12.583	2.132	5.903	<.001
	Level B1-A2	-3.783	1.578	-2.398	<.05
	Level A2-B2	10.017	1.503	6.665	<.001
	Slope A1-B1	-1.967	0.731	-2.690	.01
	Slope B1-A2	0.833	0.541	1.540	.13
	Slope A2-B2	-0.633	0.515	-1.229	.22

Note. Results reflect piecewise regression analyses conducted across all six participants. *B* = unstandardized regression coefficient; *SE* = standard error; *t* = t-statistic; *p* = two-tailed significance value. TWW = Total Words Written; TESE = Teacher Evaluation of Story Elements Rubric. Level effects indicate abrupt changes in performance at the onset of a new phase; slope effects reflect gradual within-phase trends. The absence of significant slope effects is consistent with pronounced level effects dominating the data pattern. Analyses were conducted using the R package scan (Wilbert & Lüke, 2022).

Table 6
Social Validity Results per Item and Mean Values for Each Item Across Participants

Item	Andreas	Benjamin	Carolin	Dominik	Eva	Frank	Mean
1	2	3	3	0	1	2	1.83
2	1	2	3	1	2	2	1.83
3	1	2	3	1	0	2	1.50
4	2	1	2	1	1	2	1.50
5	2	2	1	3	0	2	1.67
6	3	2	3	3	3	2	2.67
7	3	3	3	1	2	2	2.33
8	3	3	3	1	3	2	2.50
9	2	3	3	2	3	2	2.50

Note. Items were adapted from Wolf (1978) and Hurley (2012). Ratings were assigned on a 4-point scale (0 = never, 1 = sometimes, 2 = often, 3 = always). The questionnaire was administered once, immediately following the final session. Item content: (1) I enjoyed coming to the ChatGPT writing training; (2) the tasks were simple; (3) the checklist helped me to concentrate better; (4) my records spurred me on to write more; (5) I always wanted to know how many words I had written; (6) I can write better texts now than before the training; (7) other teachers should also do ChatGPT writing training with their students; (8) I think it is important to be able to write down your opinion well; (9) the teacher often told me that I had worked well.

write better texts. Furthermore, they reported that the teacher frequently praised their efforts.

Discussion

Summary and Interpretation of Key Findings

This study explored the potential of ChatGPT-based feedback to enhance the productive writing skills of students with LD. Specifically, the research addressed the question: To what extent does ChatGPT-based feedback, embedded within a structured SRSD-informed intervention, improve the productive writing skills of students with LD?

Results showed substantial improvements in both text quality and quantity, lending initial empirical support to the notion that generative AI may help address specific challenges in written expression for students with LD. Regression analyses revealed notable increases in TWW during intervention phases ($p < .001$), alongside improvements in the TESE.

The pronounced level effects during intervention phases and the decline during the return-to-baseline phase (A2) are consistent with the withdrawal logic of the ABABE design, providing evidence that the observed changes were functionally related to the ChatGPT-supported feedback rather than to maturation or practice effects. The further improvement in B2, surpassing B1 values, may reflect a cumulative learning effect: Students may have internalized aspects of the feedback routines from B1, allowing them to engage more effectively with AI-generated feedback in B2. This interpretation aligns with Bangert-Drowns et al.'s (1991) finding that the instructional effect of feedback increases when learners develop schemas for processing and applying evaluative information. From a cognitive load perspective (Sweller, 1988), it is plausible that reduced novelty of the feedback procedure in B2 freed up working memory resources for text generation and revision. Follow-up data showed a slight decline, yet values remained substantially higher than those in the A1 phase. This suggests partial retention of learned strategies and highlights the importance of incorporating mechanisms to support long-term skill transfer, such as periodic refresher sessions or embedded strategy use.

Comparison With Prior Research

The observed gains are consistent with prior research emphasizing the importance of explicit feedback and structured scaffolding for students with LD (Graham & Harris, 2005; Troia & Graham, 2003). More specifically, our results substantiate Dong's (2024) ChatGPT Feedback Engagement Framework, which conceptualizes AI-based responses not as static evaluations but as dialogic tools that promote iterative engagement. In our study, students engaged in a structured four-step cycle (prompt – drafting – AI feedback – revision), thereby operationalizing these principles.

Similar effects have been reported by Alsahli et al. (2025). Ibrahim and Ajlouni (2024) highlighted ChatGPT's potential in supporting goal-directed writing instruction for students in special education settings. Imran and Almusharraf (2023), in their systematic review of ChatGPT use in higher education, similarly identified AI-mediated feedback as a promising mechanism for scaffolded writing development – findings that may tentatively inform comparable applications at the secondary level. Our results extend this body of work by providing evidence from a rigorously designed single-case study with formally diagnosed LD learners. Together, these studies underline ChatGPT's pedagogical utility when implemented with scaffolding and curricular alignment.

At the same time, critical perspectives remain essential. Zhai et al. (2024) cautioned that AI feedback may reinforce surface-level changes and overwhelm students with lower metacognitive capacities. Price et al. (2024) further warned that generative AI can reproduce implicit stereotypes, particularly in feedback to marginalized learners. In our study, no stereotypical responses were observed, likely due to standardized prompts and structured usage. Nonetheless, these concerns call for ongoing refinement of prompt engineering and continuous ethical monitoring.

Feedback as a Mechanism for Skill Development

Our findings suggest that ChatGPT-based feedback may provide timely, specific, and actionable input, which are all key conditions for effective skill acquisition (AlGhamdi, 2024; Yan et al., 2024; Ya'u & Mohammed, 2025). In line with Hattie and Timperley's (2007) model, feedback appeared to operate primarily at the task and process levels. However, the absence of feedback at the self-regulatory level raises limitations: Current AI tools may not yet be sufficiently equipped to support strategic self-monitoring or the kind of motivational scaffolding that characterizes effective teacher-student interaction (cf. Graham & Harris, 2005).

Nevertheless, students appeared to internalize dialogic routines over time, as evidenced by stronger performance in the B2 phase. These findings support Dong's (2024) conceptualization of recursive feedback engagement and suggest that ChatGPT may serve not only as a corrective device, but also as a catalyst for developing metacognitive writing strategies. Future studies should investigate whether such recursive engagement fosters durable transfer across tasks and contexts.

Participant Engagement and Variability

Social validity assessments revealed generally high levels of engagement and perceived usefulness. Most participants reported increased confidence and enjoyment in writing, supporting the motivational benefits associated with individualized feedback (Troia & Graham, 2003). However, individual variability was notable. One participant, Dominik, reported difficulties in understanding and applying feedback. This variability mirrors the findings of Alsahli et al. (2025), who emphasized the importance of accessibility enhancements and differentiated support when implementing AI-based tools with students who have special educational needs. Our study addressed this through template-based access and prioritization scaffolds, but future implementations should further explore adaptive feedback mechanisms responsive to learners' cognitive and linguistic profiles.

Implications for Educational Practice

The findings of this study have several implications for inclusive writing instruction. Specifically, ChatGPT's capacity to deliver consistent and individualized feedback may offer a promising supplement to teacher-led instruction, particularly, in resource-constrained contexts where individual scaffolding is limited. When aligned with established instructional models such as SRSD and the Hayes-Flower writing process, AI feedback can be meaningfully integrated into existing curricular frameworks (Hayes & Flower, 1980). Preconfigured templates and structured scaffolds, as implemented in this study, support accessibility, procedural clarity, and learner autonomy. However, the scalability of such interventions depends on adequate infrastructure, teacher training, and critical digital literacy.

While the SRSD model provides a robust foundation for writing instruction, its integration with AI-based systems raises fundamental theoretical challenges. For example, core SRSD elements, such as explicit modeling, motivational support, and gradual release of responsibility, are only partially replicable through AI interfaces. This tension highlights the need for future research examining whether, and under what conditions, AI-mediated scaffolding can effectively complement or substitute teacher-led instruction without compromising instructional integrity.

In addition, social validity ratings varied considerably across participants, ranging from consistently positive evaluations (e.g., Carolin, Benjamin) to more reserved responses (e.g., Dominik, Eva), underscoring the necessity of tailoring AI-supported interventions to individual learner profiles.

Limitations of the Study

Despite the positive findings, several limitations must be considered. The single-case design, while methodologically robust for evaluating intervention effects, limits generalizability. In addition, digital infrastructure and student familiarity with technology may influence outcomes. The findings are also context-bound to students with specific types of LD within the German educational system and may not transfer to other settings without modification. While structured prompts helped mitigate variability, they may have constrained creative expression. Furthermore, the ethical implications of AI feedback, particularly regarding equity and bias, warrant continued scrutiny. As noted by Price et al. (2024), stereotype reinforcement remains a latent risk in AI use and must be proactively addressed.

Future Research Directions

Future research should pursue larger-scale, controlled studies to assess generalizability and long-term effects. Longitudinal studies are needed to evaluate the durability of skill gains and determine whether students internalize feedback routines over time. Investigations into adaptive AI systems – responsive to real-time learner needs – could further enhance differentiation. Research should also explore hybrid models that integrate AI with collaborative writing, peer feedback, and traditional instructional strategies. Additionally, studies should examine how ChatGPT-mediated interventions perform across diverse cultural and linguistic contexts, ensuring broad accessibility and contextual relevance. Finally, cross-validation with teacher-generated feedback could help clarify the relative strengths and limitations of AI-based support in inclusive writing instruction.

Conclusion

This study provides initial empirical evidence to suggest that ChatGPT-based feedback may support productive writing development in inclusive settings for students with LD. That is, by providing structured, timely, and individualized feedback, generative AI may represent a promising supplement for addressing specific challenges in the development of productive writing skills, pending replication with larger samples. The observed gains in text quality and quantity, combined with generally positive social validity ratings, lend tentative support to the feasibility of integrating AI-mediated feedback into inclusive pedagogical frameworks. While limitations remain, our findings contribute novel empirical evidence to an emerging field and align with recent calls for equitable, technology-enhanced education. Further research and refinement are essential to ensure that AI-based interventions fulfill their promise as tools for inclusion, empowerment, and sustainable learning development.

References

AlGhamdi, R. (2024). Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study. *Education and Information Technologies*, *29*, 18901–18926. https://doi.org/10.1007/s10639-024-12594-2

Alsahli, M., Alanezi, F., Sh Basri, W., Attar, R. W., Alghamdi, A., Alyahya, N. M., Albagmi, S., Almutairi, S. A., Alsedrah, I. T., Arif, W. M., Alsadhan, A. A., AlShammary, M. H., Bakhshwain, A. M., Almuhanna, A. F., Alnaim, N., & Alhazmi, A. H. (2025). Effectiveness of ChatGPT in facilitating learning for students with special educational needs: An empirical study in Saudi Arabia. *Nutrition and Health*, *31*(4), 1579–1589. https://doi.org/10.1177/02601060241307770

Babbie, E. (2020). *The practice of social research* (15th ed.). Cengage Learning.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Berninger, V. W., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 345–363). Guilford Press.

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). The Guilford Press.

Dong, L. (2024). ChatGPT in language writing education: Reflections and a research agenda for a ChatGPT feedback engagement framework. *Language Teaching Research Quarterly*, *43*, 121–131. https://doi.org/10.32038/ltrq.2024.43.07

European Agency for Special Needs and Inclusive Education. (2020). *Special needs education in Europe: Provision in post-primary settings.* https://www.european-agency.org/resources/publications/special-needs-education-europe

Graham, S., Cao, Y., Kim, Y.-S. G., Lee, J., Tate, T., Collins, P., Cho, M., Moon, Y., Chung, H. Q., & Olson, C. B. (2025). Effective writing instruction for students in grades 6 to 12: A best evidence meta-analysis. *Reading and Writing*, 38, 1–46. https://doi.org/10.1007/s11145-024-10539-2

Graham, S., & Harris, K. R. (2018). Evidence-based writing practices: A meta-analysis of existing meta-analyses. In R. Fidalgo, K. R. Harris, & G. Rijlaarsdam (Eds.), *Design principles for teaching effective writing: Empirically grounded principles* (Studies in Writing, Vol. 34, pp. 13–37). Brill. https://doi.org/10.1163/9789004270480_003

Graham, S., & Harris, K. R. (2005). Improving the writing performance of young struggling writers: Theoretical and programmatic research from the center on accelerating student learning. *The Journal of Special Education*, *39*(1), 19–33. https://doi.org/10.1177/00224669050390010301

Graham, S., Harris, K. R. & Mckeown, D. (2014). The writing of students with learning disabilities, meta-analysis of self-regulated strategy development writing intervention studies, and future directions. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 405–438). The Guilford Press.

Graham, S., Harris, K. R., & Olinghouse, N. G. (2016). Evidence-based writing practices for students with learning disabilities. In R. H. Horner & B. G. Cook (Eds.), *Practitioners' guide to implementing research* (pp. 59–80). Emerald.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Routledge.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.

Hooper, S. R., Swartz, C. W., Wakely, M. B., de Kruif, R. E. L., & Montgomery, J. W. (2002). Executive functions in elementary school children with and without problems in written expression. *Journal of Learning Disabilities, 35*(1), 57–68. https://doi.org/10.1177/002221940203500106

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford Publications.

Hurley, J. J. (2012). Social validity assessment in social competence interventions for preschool children: A review. *Topics in Early Childhood Special Education*, *32*(3), 164–174. https://doi.org/10.1177/0271121412440186

Ibrahim, A., & Ajlouni, A. (2024). Exploring ChatGPT in supporting special education undergraduates in achieving CEC standards: Students' perception. *Journal of Social Studies Education Research*, *15*(5), 87–119.

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, *15*(4), article ep464. https://doi.org/10.30935/cedtech/13605

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy*, *117*, 3–17. https://doi.org/10.1016/j.brat.2018.11.015

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250–266. https://doi.org/10.1080/00461520903213600

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.10.003

May, P., Huber, J., & Schreiber, H. (2023). *Hamburger Schreib-Probe HSP 5–10EK* [Hamburg Writing Sample HSP 5–10EK] (9th ed.). Verlag Hans Huber.

Mohammed, S. J., & Khalid, M. W. (2025). Under the world of AI-generated feedback on writing: mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. *Language Testing in Asia*, *15*, article 7. https://doi.org/10.1186/s40468-025-00343-2

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

Organisation for Economic Co-operation and Development (OECD). (2022). *PISA 2022 results: What students know and can do.* https://doi.org/10.1787/19963777

Parker, R. I., & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification*, *31*(6), 919–936. https://doi.org/10.1177/0145445507303452

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Non-overlap analysis for single-case research. *Behavior Therapy*, *42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Price, M. M., Smith, E., & Smith, R. A. (2024). "Exceptional talent and enthusiasm for math": An examination of storylines circulated by ChatGPT about mathematical learners. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, *12*(6), 1620–1637. https://doi.org/10.46328/ijemst.4471

Scheibel, G., Chen, P.-Y., Zaeske, L. M., Wills, H. P., & Zimmerman, K. N. (2022). Improving implementation fidelity with teacher-directed self-monitoring interventions: A systematic review. *Journal of Positive Behavior Interventions*, *25*(4), 253–269. https://doi.org/10.1177/10983007221137368

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Stanat, P., Hoffmann, L., Richter, D., Marx, A., & Weirich, S. (Eds.). (2022). *IQB-Bildungstrend 2021: Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 9. Jahrgangsstufe* [IQB Education Trend 2021: Competencies in German and mathematics at the end of grade 9]. Waxmann. https://doi.org/10.31244/9783830996064

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Swanson, H. L., & Sachse-Lee, C. (2001). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *34*(2), 114–136. https://doi.org/10.1177/002221940103400202

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., … & Wilson, B. (2016). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016 statement. *Physical Therapy*, *96*(7), e1–e10. https://doi.org/10.2522/ptj.2016.96.7.e1

Troia, G. A. (2011). *Instruction and assessment for struggling writers: Evidence-based practices.* Guilford Press.

Troia, G. A., & Graham, S. (2002). The effectiveness of a highly explicit, teacher-directed strategy instruction routine: Changing the writing performance of students with learning disabilities. *Journal of Learning Disabilities*, *35*(4), 290–305. https://doi.org/10.1177/002221940203500401 01

Troia, G. A., & Graham, S. (2003). The consultant's corner: "Effective writing instruction across the grades: What every educational consultant should know." *Journal of Educational and Psychological Consultation*, *14*(1), 75–89. https://doi.org/10.1207/S1532768XJEPC1401_04

UNESCO. (2020). *Global education monitoring report: Inclusion and education – All means all.* https://unesdoc.unesco.org/ark:/48223/pf0000373718

Wilbert, J., & Lüke, T. (2022). *Scan: Single-case data analysis* (Version 3.2) [R package]. https://CRAN.R-project.org/package=scan

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*(2), 203–214. https://doi.org/10.1901/jaba.1978.11-203

World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Ya'u, M. S., & Mohammed, M. S. (2025). AI-assisted writing and academic literacy: Investigating the dual impact of language models on writing proficiency and ethical concerns in Nigerian higher education. *International Journal of Education and Literacy Studies*, *13*(2), 593–604. https://doi.org/10.7575/aiac.ijels.v.13n.2p.593

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, *11*(1), article 28. https://doi.org/10.1186/s40561-024-00316-7

Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education*, *59*(2), article e12599. https://doi.org/10.1111/ejed.12599

Author Contributions

All authors contributed substantially to the conceptualization, data collection, analysis, and writing of the manuscript. All authors approved the final version of the manuscript.