Empirical Studies of Writing and Generative AI: Introduction to the Special Issue

Chris M. Anson & Kirsti Cole

Department of English, North Carolina State University | USA

Abstract: This special issue of the *Journal of Writing Research* brings together seven empirical studies of the relationship between writing and generative AI, examining what can be systematically observed and measured about the functioning of generative AI in educational and professional writing contexts. Collectively, the studies demonstrate the necessity and value of methodological pluralism for investigating a complex, rapidly evolving phenomenon. In their contributions, the researchers use experimental comparisons, mixed-methods intervention designs, corpus-based analyses, computational linguistic techniques, and qualitative interpretive approaches. Taken together, these methods enable lines of inquiry that no single approach could sustain: comparisons of AI and human performance in professional writing tasks; analyses of how writers at different ages and levels of expertise engage AI tools; examinations of how assessment systems register and respond to AI-generated prose; and investigations of how human readers interpret texts with ambiguous authorship. By foregrounding both the affordances and limitations of different methodological traditions, the articles present a multifaceted approach to the study of writing and generative AI.

Keywords: writing and generative AI, qualitative inquiry, corpus analysis, mixed-methods research, computational linguistics, experimental comparisons

journal of WRITING RESEARCH

Anson, M.C., & Cole, K. (2026). Empirical studies of writing and generative AI: Introduction to the special issue. *Journal of Writing Research, 17(i3),* 371-385. https://doi.org/10.17239/jowr-2026.17.03.01

Contact: Chris Anson, Department of English, North Carolina State University | USA – canson@ncsu.edu

The appearance of widely accessible generative AI (genAI) systems in November 2022 prompted immediate and extensive discussion in writing studies about pedagogical responses, assessment practices, and the future of writing instruction. Although breakthroughs in deep learning, transformers, and large-scale text generation had been occurring for some years before the release of ChatGPT, it was not until the widespread availability of genAI that educators, especially those responsible for teaching or supporting students' writing abilities, were suddenly confronted with serious existential questions: How would systems that effortlessly produced what convincingly appeared to be "natural language" affect the teaching and learning of writing? What would happen to the relationship between writing and thinking? How might students subvert the important cognitive, interpersonal, and social challenges of producing written discourse by using readily available genAI systems? What were the implications of AI-based analyses of student writing for large-scale assessment, a process previously fraught with the failures and inconsistencies of machine-scoring programs (Bridgeman, Trapani & Yiga, 2012; Herrington & Moran, 2006; NCTE, 2013)? How might genAI either assist or impede learning to write in a second (or third) language? What would the future hold for professionals skilled in the curation and editing of texts?

Much of the early discourse in response to these developments was characterized by speculation about potential positive and negative impacts based on reactions that one article described as "amazement and trepidation" (Anson & Straume, 2022), along with advocacy positions that preceded systematic understanding (e.g., in the U.S., the Association for Writing Across the Curriculum's 2024 policy statement and the Modern Language Association and Conference on College Composition and Communication's 2024 joint task force statement), and calls for pedagogical responses before empirical grounding. While such responses were necessary and important as writing teachers called for guidance, institutions lacked policies, and the field required ethical deliberation. The responses also revealed a significant gap: the need for rigorous empirical inquiry into how writers, readers, and evaluators engage with these systems across diverse contexts. In one sense, the advent of genAI had the potential to catalyze research on writing to a greater extent than any other moment in the field's history. However, what began as nascent systems that often generated flawed outputs, poorly assembled information, hallucinations and made-up references, biases that reflected those of human discourse and relations, and styles that did not match the accepted norms for certain genres improved and expanded at lightning speed. The extraordinary development and improvements of genAI, including the creation of multiple competing platforms, posed challenges for scholars who worried that their analyses and findings would soon be dated.

But the speed of genAI's development should not deter research into its current nature and implications. This special issue responds to the need for ongoing inquiry into the relationship between writing and genAI, especially in educational contexts. Rather than beginning with assumptions about those relationships, the seven studies collected here ask what we can systematically observe and measure about how genAI functions in specific educational and professional settings. The collection demonstrates the value of

methodological pluralism in addressing complex phenomena. These studies employ experimental comparison, mixed-methods intervention research, corpus analysis, computational linguistics, and qualitative interpretation. Together, they illuminate questions that no single methodological approach could address, including: How do AI systems perform compared to human experts in professional writing contexts? How do writers of different ages and expertise levels engage AI tools? How do assessment systems respond to AI-generated text? How do human readers make sense of ambiguously-authored prose? Each method brings its own affordances and constraints; collectively, they model the kind of multifaceted inquiry that writing and genAI require.

1. Methodological rigor across contexts

This special issue aligns with the Journal of Writing Research's commitment to methodological rigor while spanning diverse contexts and populations. It features research on professional writers, elementary students, K-12 learners, undergraduates, and readers, examining how genAI functions in workplace writing, personal diary keeping, academic writing, and assessment contexts. Importantly, these studies not only examine production (how writers use AI) but also reception (how humans interpret ambiguously- authored texts) and assessment (how AI scoring systems function and how AI can support large-scale qualitative analysis). This breadth reflects the reality that genAI's implications extend across the full ecology of literate activity. The studies are organized here thematically rather than chronologically, grouped to show how different methodological approaches address related questions about human-AI interaction in writing. The issue begins with studies examining genAI as a potential collaborator or scaffold in writing production (Janssen et al.; Liao; Sørhaug), moves to assessment of undergraduate academic writing practices (Madsen-Hardy), then turns to broader assessment contexts where genAI enables new forms of analysis and introduces new challenges (Zhang et al.; Sorapure), and concludes with a study of how readers navigate uncertainty about textual authorship (Velasquez et al.).

What these studies collectively reveal is that writing with genAI is not a single phenomenon but a constellation of practices shaped by writer expertise, developmental stage, task demands, institutional context, and technological affordances. They demonstrate that genAI interactions involve cognitive processes (attention, decision-making, metacognition), social processes (collaboration, scaffolding, apprenticeship), developmental processes (skill acquisition, transfer), affective dimensions (confidence, agency, ownership), and assessment concerns (validity, fairness, explainability). No single methodological approach can capture this complexity, which is precisely why this collection's diversity is its strength.

2. GenAI as collaborator: Three contexts

The first cluster of studies examines how writers engage AI systems as tools, collaborators, or scaffolds—and what these engagements reveal about writing processes and development

across three distinct contexts: professional editing, elementary diary writing, and K-12 classroom writing.

The opening article interrogates the capacities of genAI to mirror the behaviors of human writers in the analysis and improvement of texts. **Janssen et al.** conducted an experimental comparison of professional editing performance, asking how ChatGPT (versions 3.5 and 4) compares to experienced human editors in improving organizational writing for readability. Working with three professional editors (all with over 20 years of experience from the same Dutch agency) and four organizational letters/emails (13-29 sentences each), the researchers used screen capture and retrospective stimulated recall to track human editing processes, while testing ChatGPT with three prompt conditions: a simple prompt ("make this text reader-friendly"), a B1 readability level specification, and an 8-step expert workflow mimicking human editor processes. Readability was measured using LiNT, a Dutch text analysis tool that captures word difficulty, sentence complexity, active voice usage, and structural elements. The findings reveal both convergence and persistent difference. ChatGPT with the B1 prompt achieved comparable readability improvements to human editors on quantitative metrics, with no factual errors. However, human editors demonstrated advantages that surface-level metrics couldn't capture: greater flexibility and context-sensitivity, more nuanced tone management, and strategic adaptation across document types. Importantly, the study showed that ChatGPT output quality was highly sensitive to prompt design. The simple prompt introduced factual errors and inconsistencies, while the structured B1 prompt performed well. Human editors averaged 13 minutes per task, a time that could be reduced with AI-generated drafts as starting points.

Building on Hayes's (2012) model of writing that incorporates technology and collaboration, Janssen et al. position ChatGPT as a potential "virtual collaborator" while carefully distinguishing organizational writing (collective, functional) from educational writing (personal, expressive). Their critical insight is that effective AI use requires existing rhetorical knowledge—writers must understand audience, purpose, and genre to craft appropriate prompts and evaluate AI output. This positions prompt engineering not as a separate skill but as an applied rhetorical practice involving, for example, attention to the heuristic as opposed to command dimensions of prompts; choices of language (imperatives, hedging, metadiscursive markers); and sensitivity to the relationship between genre specification and potential outputs. In short, prompt engineering is a *learned practice* that shapes machine-generated discourse through the invocation of genre, rhetorical awareness, specification of audience, and anticipation of constraints with strategies to overcome them. The results demonstrate the importance of machine-human interaction, undermining the assumption that genAI does all the work for the writer (see Anson and Cole, in press).

But what sorts of affordances does genAI promise for collaborations with much younger, novice learners? Moving from professional to elementary contexts, **Liao** conducted a 12-week mixed-methods intervention study examining how a genAI writing companion affected Taiwanese elementary students' (Grades 3-5) writing interest and behaviors in Chinese-language diary keeping. The study combined pre/post quantitative measures of writing

interest across four dimensions (curiosity, immersion, meaningfulness, and long-term interest development) with qualitative observations and interviews. Critically, students wrote by hand while the AI companion provided scaffolded support for idea generation and character lookup. Liao found statistically significant increases in curiosity, immersion, and meaningfulness, but no significant improvement in long-term interest development. Participation patterns revealed that diary frequency declined with grade level (3→5), consistent with increased academic demands. During the AI-supported phase, students showed increased entry length and idea diversity, with improvements in narrative coherence and descriptive expression. The AI companion functioned as just-in-time lexical scaffolding—faster than dictionary lookup for difficult Chinese characters—reducing extraneous cognitive load while allowing students to focus on composition. Importantly, students valued autonomous topic selection, and the AI companion worked within student-chosen themes rather than directing them.

Drawing on scaffolding theory (Wood, Bruner, & Ross, 1976; Vygotsky, 1978), Self-Determination Theory's emphasis on autonomy as fundamental to intrinsic motivation (Ryan & Deci, 2000), and Hidi and Renninger's (2006) four-phase interest development model, Liao argues that the AI companion reduced extraneous cognitive load while preserving agency—critical for sustained engagement. The study also draws on embodied cognition research to argue that handwriting combined with occasional AI support reinforces lexical retention better than full digital composition. However, the 12-week timeframe may be insufficient for deeper attitudinal shifts, and the distinction between novelty effect and genuine pedagogical impact remains unclear. The study raises important questions about how to balance AI support that sparks initial curiosity with strategies that build enduring individual interest.

Another productive avenue for research focuses on the ways that collaborations with genAI might be seen developmentally over the span of students' schooling. Because widespread use of genAI is so recent, we have no significant longitudinal research on this question but can generate avenues for inquiry by comparing the roles and uses of genAI among young students with older ones who have already gained some expertise as writers. Thus, **Sørhaug** shifts the focus to K-12 students' authentic interactions with educational chatbots during classroom writing assignments. Analyzing 108 digital conversations between Norwegian L1 students (grades 6-13) and educational chatbots across various schools and classrooms, Sørhaug employed data-assisted thematic analysis (AI-aided coding with human verification) to examine what kinds of writing support students requested and how these requests aligned with scaffolding principles. Five request categories emerged from student prompts: information requests (most common overall, especially among younger students), structural guidance (organization, genre conventions), example requests (model texts), content creation (text generation), and feedback with follow-up clarification. Age patterns were striking: information seeking decreased with grade level while content creation and feedback requests increased. By grades 11-13, content generation and feedback accounted for over half of all interactions.

The scaffolding analysis reveals both affordances and significant limitations. Structural guidance and example requests align well with genre-based pedagogical approaches where

students receive explicit training and collaborative modeling (Rose & Martin, 2012). However, the chatbots consistently over-provided support. Even when students asked for guidance, chatbots generated complete solutions. When students requested structural advice, chatbots provided formulated sentences. When students sought feedback, chatbots offered both general principles and ready-made revisions. This pattern compromises core scaffolding principles: student ownership, appropriate challenge levels, and gradual release of responsibility (Belland, 2013; Wilkinson & Gaffney, 2015). Sørhaug argues that this over-scaffolding stems from LLMs being designed as text generators and story machines (Sharples & Pérez y Pérez, 2022) rather than pedagogical tools. The systems are optimized to be helpful and accommodating, which in educational contexts becomes excessive helpfulness that compromises student agency. Few students in the corpus employed follow-up strategies to regain control (e.g., asking what modifications were made to their text), suggesting a need to explicitly teach critical evaluation strategies.

Sorgaug's study also raises important questions about the relationship between systems design and pedagogical design. While systems design focuses on functionality, efficiency, and technologically-mediated procedures, pedagogical design emphasizes processes of learning, understanding, and engagement. One critical concern with the advent of genAI is its potential to dictate pedagogy rather than supporting it. Sørhaug's study demonstrates some degree of alignment between the two, but much more research is needed on the ways that AI-based interventions either support or subvert teachers' agency and their own human design principles and strategies. Further experimental research should compare learning outcomes with and without the kind of feedback, modeling, response, and textual outputs afforded by genAI.

Together, these three studies reveal how context profoundly shapes human-AI interaction. Professional writers with 20+ years of experience can leverage AI for efficiency in routine tasks while maintaining rhetorical control. Elementary writers benefit from scaffolding that reduces cognitive load for mechanical tasks (character lookup) while preserving autonomy in topic selection. K-12 writers seek increasingly sophisticated support with age but receive responses that undermine rather than support their development. The common thread is that effective AI integration requires not just capable systems but pedagogically-designed systems attuned to developmental stage and learning goals.

3. Assessing written products

Another important research trajectory focuses on the nature of texts produced with the assistance of genAI and, in the tradition of process studies, how students leverage its affordances in their work. **Madsen-Hardy** provides a detailed observational analysis of how undergraduate students used genAI in pilot writing courses that explicitly permitted and supported experimentation. Working with 50 students (26% EFL background) from six pilot sections at a large private U.S. research university, the study analyzed 50 research papers and 44 chat logs from 34 students. Notably, the course policy allowed up to 50% AI-generated text if marked in blue font, with endnotes describing non-text-generation AI use. Using LLM-

assisted content analysis (ChatGPT-4o for coding, with human verification achieving 85.6-100% agreement on writing samples), Madsen-Hardy examined both what students prompted for and how they integrated AI-generated text. Chat log analysis revealed significant differences between EFL and non-EFL students: EFL students prompted *less* frequently for understanding/clarification (4.8% vs. 34.1%) but *more* for direct writing help. Non-EFL students were more likely to use genAI to understand scholarly articles and then incorporate summaries. Both groups used genAI for search, brainstorming, tutoring, feedback, and argument refinement, though lower-order uses (content generation, source retrieval) were more common than higher-order uses (synthesis, evaluation).

The writing sample analysis produced surprising findings. First, 50% of students included *no* AI-generated text despite explicit permission. Among those who did use AI-generated text, 93% did not use entire paragraphs—instead, they actively selected, revised, and wove genAI content into their own language, producing hybrid texts that reflected significant student agency. EFL students were more likely to use entire paragraphs (17.8% vs. 5.5%) but incorporated fewer genAI passages overall and far fewer AI-generated summaries (2.2% vs. 37.4%).

Drawing on frameworks of critical AI literacy (Gegg-Harrison & Shapiro, 2025) and conceptualizing technology-integrated composing as a continuum rather than binary, Madsen-Hardy argues that when given permission and support, students engage genAI in strategic, selective ways—not wholesale text generation. Half of students determined that using AI-generated text was not in their interest, whether for learning, grades, or other reasons. This challenges both assumptions about student AI dependence and particularly harmful stereotypes about EFL writers being more reliant on AI assistance—a finding with important implications for how AI detection biases may unfairly target multilingual writers.

The study acknowledges important limitations: pilot participants may be atypical (self-selected into experimental sections), the sample is small with uneven EFL/non-EFL distribution, and the analysis examines usage patterns rather than writing quality outcomes. Nevertheless, it provides the first detailed examination of both prompting behavior and textual integration patterns when students have explicit permission and pedagogical support to experiment. The study also opens up a number of important questions about L1 and L2 writers' use of genAI. Currently, much research is focusing on students writing in English as their second language. But genAI models are known to perform less effectively and exhibit cultural biases as they move from dominant to under-resourced languages. Limitations include lack of linguistic nuance, poor representation, and the potential marginalization of speakers of lesser used tongues. Models are most often trained on English-centric data and therefore produce material for speakers of under-resourced or non-standard languages that is culturally and linguistically slanted toward Anglophone contexts (see Lee, Choe, Zou, and Jeon, 2025, for a systematic review of 49 empirical studies).

After the release of ChatGPT and the initial handwringing about student writing, many educators realized that genAI systems could be trained to "read" student work and produce both formative and summative evaluations. One perspective considered how students

themselves might use genAI in ways similar to peer review, to solicit feedback on drafts in progress that they could then use in systematic revision (see, for example, Lo, Wan, and Chan, 2025; McGuire, Qureshi, and Saad, 2024). Another perspective focused on how teachers might employ genAI to replace their own time-consuming and painstaking processes of response and evaluation. Programs soon were marketed to provide a gateway to such evaluative systems, such as Brisk Teaching's "AI Feedback Generator for Teachers" (https://www.briskteaching.com/give-feedback). Finally, writing assessment experts long involved in machine scoring anticipated more robust methods of large-scale writing assessment despite the limitations of, for example, analyzing samples of student writing created in a single, timed sitting under tight constraints of subject matter and composing processes (with studies showing mixed results; see Bui & Barrot, 2025). These developments have raised important ethical questions about the instructional uses of genAI that have policy implications as well as the need for mindful pedagogical integration.

Two studies in this special issue examine how genAI functions in writing assessment contexts, revealing both new possibilities and new challenges for validity and fairness. **Zhang et al.** conducted a systematic investigation of whether AI-generated essays could augment training data for automated essay scoring systems, with particular attention to fairness across racial/ethnic groups. Working with the PERSUADE 2.0 corpus (1,372-2,167 essays per prompt, grades 6-10, seven argumentative/explanatory prompts), the researchers compared scoring models trained exclusively on student essays against "augmented" models trained on combinations of student and GPT-generated essays. The study unfolded across three research questions. First, how similar are AI-generated essays to student essays? Using e-rater to extract surface-level linguistic features (grammar, syntax, mechanics, word usage, vocabulary, text length), the researchers found that AI-produced essays (generated by GPT-4 and GPT-4o with rubric-based prompts and example student essays) showed surface-level alignment in syntax and discourse structure but were slightly shorter with more sophisticated vocabulary. Importantly, while these similarities suggested potential for augmentation, material differences might exist beyond surface features—a limitation requiring qualitative validation.

Second, how accurate are scoring models trained with augmented samples? Fine-tuning models (48M parameters) at varying training sample sizes (20%, 40%, 60%, 80%, 100% of student essays), the researchers found that augmented models performed comparably to base models (Quadratic Weighted Kappa >0.75 for most prompts), with sample size having minimal impact once exceeding ~1,000 essays. Augmentation neither consistently improved nor degraded overall performance, but it significantly increased representation for undersampled score levels (1, 5, and 6), where student essays were sparse. Third, are augmented models fair across racial/ethnic groups? Here the findings are striking. Base models showed bias favoring Asian/Pacific Islander students, with Mean Difference in Standardized Scores (MDSS) exceeding 0.2 in several prompts. Augmentation substantially reduced these biases, bringing MDSS down to ~0.1. Similarly, when base models favored other groups in specific prompts, augmentation mitigated discrepancies. Token importance analysis via DecompX revealed that base models overweighted prompt keywords (e.g., "driverless" in

the driverless cars prompt) and logical connectors. Asian/Pacific Islander students used more content words from prompts, which base models rewarded disproportionately. Augmented models showed more balanced token importance distributions across groups.

The study demonstrates that synthetic data augmentation can address a critical problem in AI scoring: small subgroup sample sizes that lead to biased models. By generating essays across all score levels and prompts, augmentation provides diversity that helps models generalize better across demographic groups. However, significant limitations remain. With only one human rating per essay, true score evaluation was impossible. Quantitative metrics may oversimplify content and style differences. Token-level explainability reveals patterns but not causality—it doesn't explain *why* tokens matter or *how* changes would affect scores. Most critically, expert human review is needed to validate whether AI-generated essays genuinely reflect the rubric at each score level.

These findings support continued concerns about fairness in large-scale assessment, following on many critiques of various methods including holistic and criterion-based analysis, machine scoring, and portfolio assessment, and now AI-based assessment (Ericksson & Haswell, 2009; Hodges, et al., 2019; Jiang, Hao, Fauss, & Li, 2024; Kelly-Riley, Macklin, & Whithaus, 2024; Poe & Elliott, 2019). These concerns are especially important for the assessment of L2 writers (see, for example, Ghanbari, 2019; Plakans and Lee, 2025).

One research method for testing the fairness and accuracy of AI-based large-scale writing assessment is to employ corpus analysis on large datasets. **Sorapure** demonstrates a different application of genAI in assessment: using Retrieval-Augmented Generation (RAG) and AI-assisted thematic analysis to analyze large corpora of student writing. Working with 3,334 students' responses to Collaborative Writing Placement questions at a large research-extensive university (approximately 20,000 total responses to four open-ended questions about writing experience and expectations), Sorapure focused on students' positive self-assessments as writers—an asset-based approach aligned with directed self-placement principles. The two-stage method first employed RAG to efficiently identify relevant responses in the large corpus. Using two embedding models and four search prompts ("I am confident about my writing," "I feel prepared for college writing," "I am a strong writer," "I can write well"), the system retrieved 260 relevant responses (after deduplication) plus 50 random responses for comparison. Three expert faculty raters validated relevancy, confirming that RAG significantly outperformed random selection. This demonstrates RAG's potential for targeted corpus exploration—essentially asking questions of large datasets that would be impractical through human reading alone. Stage two used ChatGPT-4 to conduct thematic analysis following Braun and Clarke's (2006) established phases: data familiarization, coding, theme development, theme review, and theme definition. Comparing AI-generated themes to human coding revealed that ChatGPT-4 could conduct rigorous thematic analysis with human oversight, potentially expediting qualitative research while maintaining interpretive depth.

Sorapure is careful to position this not as genAI replacing human analysis but as human-AI collaboration along a spectrum from "machine-in-the-loop" (humans direct) to "human-in-

the-loop" (AI leads, humans refine). The method requires careful human oversight precisely because genAI cannot reflect on its own biases or step outside its training data. The qualitative research's human element—researchers' perspectives and experiences—remains essential for meaningful interpretation. Nevertheless, the study opens important possibilities for asset-based analysis at scale, enabling researchers to ask questions of large corpora that would otherwise remain unexamined.

Sorapure's contribution to the special issue suggests a further area for the exploration of large-scale analysis of writing: using genAI in the context of multiple disciplines. For example, although genAI does relatively well at analyzing discourse in scientific fields because of their emphasis on technical accuracy and more fixed genres and structures, it performs less effectively in humanities disciplines whose communication involves deep, original arguments, nuance, and discipline-specific kinds of creativity. One study of accuracy, depth, pedagogical alignment, and interpretive appropriateness of genAI assessment of writing in different disciplinary areas, for example, found that it provided feedback on texts in the humanities that was overgeneralized and ambiguous, and did not always interpret intention accurately (Steve, Roland, & Joseph, 2025). More research is needed into unexplored disciplinary contexts where genAI may be less fully trained on idiosyncratic genres and specific writing conventions before educators can be confident that it can provide accurate assessments of student or professional writing.

4. Reading ambiguously authored texts

Although writing assessment involves the aggregated "reading" of corpora, whether by assessment experts or machines, research is needed to study the effects of more general readers of AI-generated texts with those produced by humans. **Velasquez et al.** shift from production and assessment to reception, asking how human readers make sense of texts when authorship is uncertain. In a qualitative, exploratory study, 76 readers (writing instructors, graduate students, undergraduates with varying AI experience and confidence) read three anonymized abstracts from social science undergraduate research: one human-authored, one AI-generated (ChatGPT-4), and one co-written (human original revised by ChatGPT-4). Participants were explicitly told abstracts might be human, AI, or hybrid, then asked to determine authorship and explain their reasoning. The study employed multiple data collection methods to capture readers' evidential processes: surveys (demographics, AI experience, confidence levels), talk-aloud protocols (video/audio recorded), written reasoning, focus groups (collaborative negotiation among synchronous cohort, n=20), and semi-structured interviews. Participants were not asked to "detect" AI accurately but to articulate how they made decisions—what "cues in the text" (Haas & Flower, 1988) they found evidentially significant. Drawing on writing studies scholarship on "felt sense" (Perl, 2004) and "tacit knowledge" (Polanyi, 1967), Velasquez et al. analyzed readers' often pre-linguistic intuitions about texts. Felt sense indexes "an unclear, barely noticeable bodily sensation" (Perl, p. xiii) or "inchoate pushes and pulls" that exist before articulation. In the context of AI-

generated prose, this becomes sensing an "offness"—something is slightly wrong even if readers cannot immediately specify what.

Preliminary findings (coding ongoing) reveal that readers drew on multiple types of evidence: linguistic cues (vocabulary sophistication, sentence structure patterns, transition usage, "generic" vs. specific language), rhetorical cues (clarity, organization, disciplinary convention use), and importantly, affective responses—the "it's giving AI" phenomenon where readers described a text's "vibe" before articulating specific features. One asynchronous participant wrote: "Something about the language pattern. It's hard to say exactly what tipped me off." The synchronous cohort's collaborative negotiations revealed how readers with different expertise levels voiced, tested, and revised hypotheses in real-time, with embodied responses (gestures, facial expressions) indicating tacit knowledge activation. Video analysis of gesture and nonverbal communication is ongoing, as are plans for eye-tracking studies to capture even more fine-grained literacy practices. The study's significance lies not in measuring detection accuracy (it's qualitative, not quantitative) but in revealing the complex, embodied, often tacit processes through which readers navigate textual uncertainty in the AI era. As one of the first empirical examinations of reading practices now that genAI is part of our literate landscape, it extends felt-sense scholarship from writing to reading, showing how readers yoke textual cues with prior knowledge and experience to make meaning under conditions of ambiguous authorship.

Velasquez's contribution raises further important questions about the embodied nature of AI systems. Noller (2025), for example, theorized a relationship between large language models and "4E" cognition, which embraces "embodied, embedded, enactive, and extended" cognitive processes. genAI is seen through the lens of processes and relational phenomena involving the interaction of human agency and the technical dimensions of computers. genAI is not simply a technology but "a co-evolving component of the extended cognitive ecology of human life, shaping and shaped by enactive practices, intentions, and norms" (1). Similarly, Aguilar (2025) calls for a stronger critical awareness of the invisible labor of "humans in the loop" of AI-based products and interactions. Studying Google Translate, Aguilar documents the extent to which AI-generated products are already humanly authored but "erase" traces of writing embodiment (see also Tang, 2025, on writing, embodiment, and intertextuality). In the future, research on genAI will need to broaden its focus from relationships between textual inputs and outputs mediated by technology to more fully investigate the embodied contexts of mobile and wearable devices, spatial environments, and the functions of voice and touch. (See, for example, Nimi, Lu, and Chacon, 2025, on embodied co-creation of genAI in the context of an interactive art installation).

5. Synthesis: What these studies reveal

Taken together, these seven studies demonstrate several key insights about writing, reading, and assessment in the age of genAI.

First, methodological diversity is essential, not optional. Experimental comparison (Janssen) reveals performance differences under controlled conditions but may not capture

naturalistic use. Intervention studies (Liao) enable causal inference about developmental effects but raise questions about sustainability. Observational corpus analysis (Madsen-Hardy, Sørhaug) captures authentic behavior with high ecological validity but cannot determine causality. Computational methods (Zhang, Sorapure) enable pattern analysis at scale but require human interpretation. Qualitative approaches (Velasquez) illuminate meaning-making processes that resist quantification. No method is superior; each reveals different dimensions of complex phenomena. This collection models the kind of pluralistic inquiry that genAI and writing require.

Second, population and context matter profoundly. Professional writers with decades of experience can leverage genAI for efficiency while maintaining rhetorical control (Janssen). Elementary writers benefit from scaffolding that reduces mechanical cognitive load while preserving topic autonomy (Liao). K-12 writers need explicit instruction in critical engagement as they increasingly turn to genAI for text generation (Sørhaug). Undergraduates with permission and support make strategic, selective choices—including choosing not to use genAI at all (Madsen-Hardy). Assumptions that younger, less experienced, or multilingual writers depend more heavily on genAI are not supported by these data. This suggests that blanket policies, whether prohibitive or uncritical, ignore developmental and contextual realities. Effective pedagogical responses must be differentiated.

Third, human expertise, agency, and critical engagement remain central but take different forms. Expertise still matters: professional editors demonstrate advantages in flexibility and rhetorical judgment (Janssen); more experienced undergraduate writers use genAI for higher-order tasks, such as understanding sources, rather than just generating text (Madsen-Hardy); readers draw on tacit knowledge and embodied responses to evaluate authorship (Velasquez). Agency can be preserved: elementary students valued the choice of their own topics (Liao); half of undergraduates, with permission, chose not to use AI-generated text (Madsen-Hardy). But agency can also be compromised: chatbots over-provide support even when students ask for guidance (Sørhaug); older students increasingly request wholesale content generation (Sørhaug). The challenge is not whether humans remain central (they do) but how we support the development of expertise and agency in AI-mediated contexts.

Fourth, assessment validity now requires new frameworks. Writing assessment has always involved questions of validity, reliability, and fairness. GenAI intensifies these concerns in multiple ways. When readers navigate textual uncertainty, they rely on felt sense and tacit knowledge that may be inconsistent and potentially biased (Velasquez). When scoring systems are trained on limited samples, they can introduce or amplify demographic biases (Zhang). At the same time, genAI enables new possibilities: synthetic data augmentation can mitigate scoring biases for underrepresented groups (Zhang); RAG enables asset-based analysis of large placement corpora (Sorapure). Assessment validity now depends not just on evaluating text quality but on understanding agency distribution, collaboration patterns, and potential bias in both human and automated systems.

Fifth—and perhaps most important—what remains unknown far exceeds what these studies reveal. These seven studies raise more questions than they answer, which is exactly what good empirical work should do. Further research, for example, needs to take up questions about writing processes (how they change over extended genAI use beyond limited time frames, and what happens when students iteratively use genAI for multiple tasks across different contexts). Related to process questions are those of development, such as whether genAI scaffolding leads to the improvement or atrophy of different writing skills, or whether early exposure to genAI affects later writing. We also need to know much more about the effects of genAI on different populations (such as learners in different language and cultural groups) and whether findings transfer across institutions, disciplines, languages, and modalities. Finally, many questions about equity remain unanswered: How do genAI systems function for neurodiverse writers, writers with disabilities, and writers who have intermittent or compromised access to digital technologies?

6. Implications and future directions

These studies have implications for writing instruction, assessment, and research, though it's critical to note what empirical evidence can and cannot determine.

For writing instruction, the evidence suggests that effective responses must be differentiated by context, age, and expertise rather than following one-size-fits-all policies. Elementary writers may benefit from AI scaffolding that reduces mechanical cognitive load while preserving autonomy (Liao). K-12 writers need explicit instruction in prompting as a rhetorical practice and critical evaluation of genAI responses (Sørhaug). Undergraduates need opportunities to experiment with genAI support and develop autonomous judgment about when and how to use it (Madsen-Hardy). Professional writers can leverage genAI for efficiency in routine tasks (Janssen). Across all contexts, effective genAI use requires rhetorical knowledge—understanding of audience, purpose, and genre—suggesting that teaching with genAI means teaching rhetoric and critical literacy in addition to responsible and ethical tool use.

Importantly, current LLMs function as text generators rather than pedagogical tools (Sørhaug). They over-provide support, offer complete solutions when asked for guidance, and don't embody scaffolding principles like gradual release of responsibility. This suggests a need for pedagogically designed systems, not just capable generators. In the interim, instruction should focus on stages where genAI support is appropriate (idea generation, structural planning), while being cautious about continuous genAI use throughout the writing process, as over-scaffolding may compromise development.

For writing assessment, the evidence suggests that validity frameworks must expand to account for human-AI collaboration. Detection of AI-generated text is unreliable, subjective, and potentially biased (Velasquez). Assessment must evaluate not just text quality but also agency distribution and collaboration patterns—a complex undertaking. At the same time, genAI enables new assessment possibilities: synthetic data can mitigate scoring biases (Zhang), and RAG can enable asset-based analysis at scale (Sorapure). The key is that these

applications require careful validation and human oversight. genAI can augment but not replace human judgment in assessment contexts.

For writing research, this collection models several principles. Methodological pluralism is essential—experimental, observational, computational, and interpretive approaches each contribute necessary perspectives. Equity must be foregrounded throughout research design, not added as an afterthought—examining fairness across demographic groups (Zhang), challenging assumptions about multilingual writers (Madsen-Hardy), and attending to whose interests are served by particular configurations. Limitations should be embraced as productive rather than merely constraining—they point toward necessary future work.

What's needed moving forward includes:

- longitudinal studies tracking writers' genAI use and development over extended periods, including the extent to which genAI outputs create syntactic and rhetorical structures that tacitly affect learners' own writing;
- cross-context studies comparing educational and professional writing across disciplines and languages;
- intervention studies with developmental measures that systematically test pedagogical approaches; and
- equity-focused research examining genAI's effects on writers with disabilities, in multilingual contexts, and across socioeconomic access patterns.

7. Conclusion

The articles in this special issue demonstrate what empirical inquiry can contribute to urgent conversations about genAI and writing: not definitive answers, but rigorous, contextualized evidence about how writers, readers, and assessment systems engage these tools across diverse settings. Each study attends carefully to its specific population, context, and method while acknowledging limitations and implications. Together, they model the kind of sustained, pluralistic, equity-attentive inquiry that writing and genAI requires. The field needs more such work—not to settle debates about whether genAI "threatens" or "promises" to transform writing, but to ground those debates in systematic observation and measurement of what actually happens when humans and genAI systems interact in acts of literacy. Writing is complex, situated, and irreducibly social. Understanding how genAI functions within writing ecologies requires multiple methodological approaches, attention to population and context, sustained engagement with questions of equity and validity, and willingness to embrace the complexity these studies reveal.

References

Anson, C. M., & Cole, K. (in press). Generative AI does all the work for the writer. In C. Basgier, A. Mills, M. Olegnik, M. Rodak, & S. Sharma (Eds.), *Bad ideas about AI and writing: Toward generative practices for teaching, learning, and communication.* The WAC Clearinghouse and University Press of Colorado.

Anson, C. M., & Straume, I. (2022). Amazement and trepidation: Implications of AI-based natural language production for the teaching of writing. *Journal of Academic Writing, 12*(1), 1-9. https:doi.org/https://doi.org/10.18552/joaw.v12i1.820

Aguilar, G. L. (2025). AI writing is always embodied: Building a critical awareness of the invisible labor of humans-in-the-loop in AI products. *College Composition and Communication, 77*(1), 39-61. https://doi.org/10.58680/ccc202577139

Association for Writing Across the Curriculum (2025). *AWAC Statement on AI and writing across the curriculum*. https://wacassociation.org/statement-on-ai-writing-tools-in-wac/

Belland, B. R. (2013). Scaffolding: Definition, current debates, and future directions. In M. J. Spector, M. D. Merril, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 505-518). Springer. https://doi.org/10.1007/978-1-4614-3185-5_39

Bridgeman, B, Trapani, C. & Yigal, A. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education 25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: a comparative study. *Innovation in Language Learning and Teaching,* 1-16. https://doi.org/10.1080/17501229.2025.2521003

Ericsson, P. F., & Haswell, R. (2006). *Machine scoring of student essays: Truth and consequences*. Utah State University Press.

Gegg-Harrison, W., & Shapiro, S. (2025). From policing to empowerment: Promoting student agency in the context of AI text-generators and AI-detection tools. In C. Wang & Z. Tian (Eds.), *Rethinking writing education in the age of generative AI* (pp. 26–41). Routledge. https://doi.org/10.4324/9781003426936-4

Ghanbari, N. (2019). Promoting fairness in EFL writing assessment: Are there any effects of the writers' awareness of the rating criteria? *Journal of Asia TEFL, 16.*

Herrington, A., & Moran, C. (2006). WritePlacer Plus in place: An exploratory case study. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 114-129). Utah State University Press. https://doi.org/10.18823/asiatefl.2019.16.1.173

Haas, C., & Flower, L. (1988). Rhetorical reading strategies and the construction of meaning. *College Composition and Communication*, *39*(2), 167-183.

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

Hidi, S., & Renninger, A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the writing rubric to inform teacher educators (WRITE). *Assessing Writing*, *40*, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Jian, Y., Hao, J., Fauss, M., & Li, C. (2024). Toward fair detection of AI-generated essays in large-scale writing assessments. In S. M. Olney, I. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Proceedings of the 25th International Conference of AIED,* Recife, Brazil, July 8-12. https://link.springer.com/book/10.1007/978-3-031-64312-5

Kelly-Riley, D., Macklin, T., & Whithaus, C. (2024). Toward fairness in writing assessment. In D. Kelly-Riley, T. Macklin, & C. Whithaus (Eds.). *Considering students, teachers, and writing assessment* (pp. 107-116). The WAC Clearinghouse and University Press of Colorado.

Lee, S., Choe, H., Zou, D., & Jeon, J. (2025) Generative AI (genAI) in the language classroom: A systematic review. *Interactive Learning Environments.* https:doi.org/ 10.1080/10494820.2025.2498537

Lo, N., Wong, A., & Chan, S. (2025). The impact of generative AI on essay revisions and student engagement. *Computers and Education Open, 9,* 100249. https://doi.org/10.1016/j.caeo.2025.100249

McQuire, A., Qureshi, W., & Saad, M. (2024). A constructivist model for leveraging genAI tools for individualized, peer-simulated feedback on student writing. *International Journal of Technology in Education, 7*(2), 326-352.

Modern Language Association and Conference on College Composition and Communication (2024). *Generative AI and policy development: Guidance from the MLA-CCCC task force*. https://cccc.ncte.org/mla-cccc-joint-task-force-on-writing-and-ai

National Council of Teachers of English (2013). *NCTE position statement on machine scoring.* National Council of Teachers of English. https://cdn.ncte.org/nctefiles/resources/positions/machinescoring_booklet.pdf

Nimi, H., Lu, M., & Chacon, J. C. (2025). Embodied co-creation with real-time generative AI: An Ukio-E interactive art installation. *Digital, 5*(4), 1-21. https://doi.org/10.3390/digital5040061

Noller, J. (2025). 4E cognition and the coevolution of human–AI interaction. *Discover Artificial Intelligence, 5*(323), 1-19. https://doi.org/10.1007/s44163-025-00595-0

Plakans, L., & Lee, K. (2025). Fairness, justice, and criticality: Reviewing second language writing assessment. *Language Teaching,* Firstview, 1-28. https://doi.org/https://doi.org/10.1017/S0261444825100876

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English*, *13*(4), 317-336.

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in *Assessing Writing*. *Assessing Writing, 42.* https://doi.org/10.1016/j.asw.2019.100418

Polanyi, M. (1958). *Personal knowledge*. Routledge.

Rose, D., & Martin, J. R. (2012). *Learning to write/reading to learn: Genre, knowledge and pedagogy in the Sydney school*. University of Toronto Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68-78. https://doi.org/10.1037//0003-066x.55.1.68

Sharples, M., & Pérez y Pérez, R. (2022). *Story machines: How computers have become creative writers*. Routledge. https://doi.org/10.4324/9781003161431

Steve, C., Roland, C., & Joseph, O. (2025). Assessing the quality and accuracy of AI-generated feedback in STEM v. humanities education. *Research Gate*. https://www.researchgate.net/publication/393465033

Tang, K-S. (2025). AI-textuality: Expanding intertextuality to theorize human-AI interaction with generative artificial intelligence. *Applied Linguistics, XX*, 1-19. https://doi.org/10.1093/applin/amaf016

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* M. Col, V. Johnson-Steiner, S. Scribner, & E. Souberman (Eds.). Harvard University Press.

Wilkinson, I. A. G., & Gaffney, J. S. (2015). Literacy for schooling: Two-tiered scaffolding for learning and teaching. In L. Corno & E. M. Anderman (Eds.), *Handbook of educational psychology* (pp. 243–258). Routledge.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89-100.