Measuring Creativity in Co-Writing with AI: Rhyme Density and the Limits of Computational Proxies

Ibùkún Ọlátúnjí[1]* **Mark Sheppard**[2]*

[1]Computational Foundry, Swansea University, Crymlyn Burrows, Skewen, Swansea SA10 6JW, UK
[2]University of Kent, Canterbury, Kent CT2 7NZ, UK

Abstract

Evaluating creativity in AI-assisted writing remains a persistent challenge, often relying on subjective user ratings with limited reproducibility. This paper investigates rhyme density as a computational proxy for creative quality in human–AI collaborative poetry tasks. Drawing on data from a mixed-methods study (N=30), we analyse how rhyme-centric metrics relate to participant evaluations of fluency, creativity, and accuracy. We contribute (1) an empirical analysis of rhyme-centric features in AI-assisted writing, (2) a critical examination of their limitations in capturing human judgements, and (3) a framework for future evaluation methods that integrate computational rigour with human-centred insight. This work advances debates in computational creativity by offering practical strategies for assessing co-creative output.

Introduction

This study investigates the role of computational writing tools in poetry and lyric composition, examining how different tool types influence both the creative process and the resulting stylistic outcomes. Building on research into co-creative systems (Karimi et al. 2018) and generative writing tools (Arnold, Volzer, and Madrid 2021), we design a set of structured tasks in which participants use support, co-creative, and autonomous tools. We analyse the resulting texts using quantitative metrics (e.g., lexical distance, rhyme density) alongside participant evaluations, providing a multi-dimensional account of how these tools shape creative expression. We introduce a novel task design that integrates divergent idea generation into subsequent writing tasks, enabling joint evaluation of both creative phases. This allows us to study not only final outputs, but how creative intent is transformed into poetic form, an approach that bridges process and product-based perspectives on creativity. We focus on rhymed poetic forms as a structured yet expressive domain for computational analysis.

Contribution This work makes three key contributions: (1) empirical insights into how tool types affect user experience and stylistic expression; (2) a novel task design linking divergent and convergent phases of creativity; and (3) a framework for evaluating co-creativity through both computational and human-centred measures.

Related Work

Karimi et al. propose a framework for characterising human roles in creative systems (Karimi et al. 2018), distinguishing between : (1) *creativity support tools*, which assist but do not generate content (e.g., a rhyming dictionary (WriteExpress 2022)); (2) *co-creative systems*, which collaborate with users during generation (e.g., a poetry writing companion (Google Research 2022)); and (3) *fully autonomous systems*, which produce complete texts from prompts without further human input (e.g., ChatGPT (OpenAI 2024)).

System Type	Description	Examples
Creativity Support	Assists with specific writing tasks through language processing tools.	Rhyming dictionaries, Thesauri
Co-Creative Systems	Collaborates with the writer by generating text suggestions, acting as a co-writer.	Verse by Verse, Deepbeat
Autonomous Systems	Generates complete texts independently based on user prompts.	GPT-4, Bard, LLaMA-2

Table 1: Creative Systems Framework

Writing Assistants: Language model–based systems have become increasingly common in both research and commercial settings (OpenAI 2024; Google Research 2022). However, users often disengage due to shallow interaction dynamics and unclear collaboration roles (Rezwana, Maher, and Davis 2021). As Arnold et al. observe, a common pattern is for models to generate text that users then adopt or adapt. While this workflow aligns with system capabilities, it “does not necessarily support all writing approaches that writers may find useful.”(Arnold, Volzer, and Madrid 2021). Such systems also risk blurring authorship,

*Corresponding authors: 2030349@swansea.ac.uk, ms2403@kent.ac.uk

as they ”feed words to writers suggesting that the writer claim the system’s words as their own, especially with regards to possible plagiarism claims.” (Arnold, Volzer, and Madrid 2021)

Poetry: We focus on poetry because it draws on phonemic awareness, vocabulary, and world knowledge (Rojcewicz 2004; Hadaway, Vardell, and Young 2001; Wassiliwizky et al. 2017). Its formal constraints, such as metre and rhyme, enable measurable comparisons between human and AI-assisted outputs. Studying poetic composition offers insight into how computational tools shape stylistic choices and structural form. To frame this investigation, we adopt Karimi et al.’s taxonomy of support, co-creative, and autonomous systems (Karimi et al. 2018) (see Table 1).

Study Design

The study includes three components: (1) a pre-task questionnaire to capture participants’ writing experience and expectations; (2) writing tasks with three tool types—support, co-creative, and autonomous; and (3) a post-task questionnaire evaluating fluency, creativity, accuracy, and tool usefulness. Each participant completed a base poem without computational assistance, followed by three tool-assisted tasks, enabling within-subject comparison across tool types.

Study Design

Pre-study Questionnaire. The questionnaire collected demographic information and writing experience to serve as a baseline for later comparisons. The study involved 30 participants, each completing all writing tasks in a within-subjects design.

Divergent Association Task (DAT). The DAT (Olson et al. 2021) was used as a pre-test to measure verbal creativity. Participants were instructed to generate 10 words as different from each other as possible. The task, which takes under four minutes, provides an efficient and validated measure of divergent thinking. Semantic *distance* between the words was quantified, with higher DAT scores associated with greater creativity. These scores were later explored in relation to participant performance in the writing tasks.

Writing Tasks. Participants completed four writing tasks to assess how different tools influenced creative output: (i) *Baseline*, no tools used; (ii) *Support Tool*, e.g., a rhyming dictionary; (iii) *Co-Creative Tool*, offered suggestions in collaboration with the user; and (iv) *Autonomous Tool*, generated complete texts independently.

Task Instructions. To ensure consistency, all writing tasks were based on the theme of *place*:

- ”What does the place you live mean to you?”
- ”What is your favourite location in the place you live?”
- ”How does the place you live make you feel?”

Participants were also asked to (i) incorporate as many DAT words as possible and (ii) use rhyme in their compositions. This combination encouraged both divergent thinking (idea generation) and convergent thinking (structuring ideas within poetic constraints), allowing us to observe how participants balanced creative exploration with formal structure.

Study Component	**Description**
Study Design	Mixed-methods design including pre-study questionnaire, Divergent Association Task (DAT), four writing tasks, and post-study questionnaires.
Participants	**N = 30** participants recruited through direct requests to acquaintances, including journalists and writers.
Pre-study Questionnaire	Age and previous writing experience (low, medium and high)
Divergent Association Task (DAT)	Verbal creativity measured by generating 10 unrelated words; used as baseline creativity indicator.
Writing Tasks	• *Baseline:* Without computational tools. • *Support Tools:* Tools assisting specific aspects (e.g., rhyme). • *Co-Creative Tools:* Tools providing suggestions or partial content. • *Autonomous Tools:* Fully automated text generation.
Post-study Questionnaire	Participant subjective ratings (1–5 Likert scale) of fluency, creativity, and accuracy of their writing outputs produced using the tools, and ratings of tool usefulness.
Scoring	DAT computed via word distance metrics; Rhyme Density scores by phoneme Cosine similarity.

Table 2: Summary of Study II design and methodological components.

Ethics All participants gave informed consent, and the study ensures anonymity and data protection in line with current ethical guidelines.

Rhyme Density

To enable consistent comparison across diverse outputs, we compute rhyme density scores for both participant-generated and tool-assisted compositions. Rhyme density serves as a shared, quantifiable anchor across conditions, helping to normalise differences in length, structure, and output style. Beyond this, we treat rhyme density as part of a broader framework for examining creativity as a combination of divergent and convergent thinking. The Divergent Association Task (DAT) provides a pre-task measure of ideational fluency, while the use of DAT-generated keywords reflects how participants transform divergent ideas into creative content. Rhyme density, as a formal constraint, captures convergent effort during composition. Together, these measures allow us to examine how divergent thinking and convergent structuring manifest in creative writing, and how these dimensions vary across tool conditions.

Following prior work (Hirjee and Brown 2009; Condit-Schultz 2016; Malmi et al. 2016), we define rhyme as the repetition of similar phonemes within or across word sequences. Phoneme similarity is computed using cosine similarity between phoneme embeddings, allowing us to capture both exact and approximate rhymes.

$$\text{sim}(p_1, p_2) = \frac{\vec{v}p_1 \cdot \vec{v}p_2}{|\vec{v}p_1||\vec{v}p_2|}$$

where:
$\vec{v}p_1$ and $\vec{v}p_2$ represent the embedding vectors of phonemes p_1 and p_2, respectively.

Rhyme Density Calculation Rhyme Density (D) is calculated as the ratio between the total number of phoneme pairs , exceeding a predefined similarity threshold (θ), and the total number of phonemes in the analyzed text:

$$D = \frac{\sum_{i,j} M_{ij}}{|P_T|}\ M_{ij} \quad = \begin{cases} 1 & \text{if } \text{sim}(p_i, p_j) \geq \theta \ 0 \\ \text{otherwise} \end{cases}$$

where:
$|P_T|$ denotes the total count of phonemes within the text that is analyzed.

Phoneme Grouping To further analyze rhyme patterns, phonemes are organized into groups based on similarity scores. The phoneme grouping function $G(p)$ is defined as:

$$G(p) = \begin{cases} k & \text{if } \exists p' \in P : \text{sim}(p, p') \geq \theta \ |G| + 1 \\ \text{otherwise} \end{cases}$$

where:
k is the identifier of an existing phoneme group.
$|G| + 1$ indicates the creation of a new phoneme group for phonemes without sufficiently similar matches.

The grouping methodology enables detailed insights into rhyme structure and phonemic repetition, which allows comparisons between human and computational creative outputs.

Results

We used a variety of statistical models to explore relationships between participants' divergent thinking (DAT scores), unaided ability (baseline rhyme density), and their performance across writing tasks involving varying levels of computational support. Table 3 provides an overview of the key models used. Our main findings include:

- A positive correlation between divergent thinking (DAT scores) and rhyme density across all writing tasks.
- Experience significantly predicted improved rhyme performance in tool-assisted tasks (support, co-creative, and autonomous).
- No significant relationship between experience level and baseline (unaided) rhyme performance.

Analysis Type	**Variables/Factors**	**Significance Level**
Pearson Correlation	DAT scores, Rhyme Scores, Experience Levels	$\alpha = 0.05$
ANOVA (One-way)	Experience Levels (groups), Tool Usefulness Ratings	$\alpha = 0.05$
OLS Regression	DAT scores, Base Rhyme Scores, Tool Usefulness Ratings	$\alpha = 0.05$
Subgroup Comparison	Experience, DAT scores, Rhyme Score Groups, Questionnaire Ratings	$\alpha = 0.05$

Table 3: Select Statistical Analyses and Variables.

Analysis

We present a more detailed interpretation of the results below. Selected outputs are summarised in Table 4, with additional qualitative insights included throughout.

Pearson Correlations We calculated Pearson correlation coefficients to measure linear relationships between DAT scores, rhyme density, and experience level across task types. Results revealed a moderate, positive, and statistically significant correlation between DAT scores and rhyme density in all task types (baseline, support, co-creative, autonomous). In contrast, experience level was not significantly correlated with baseline rhyme scores.

ANOVA To examine differences in rhyme performance across experience groups, we conducted one-way ANOVA tests. Results showed no significant differences in baseline rhyme scores based on experience. However, experience level significantly affected rhyme scores for all tool-assisted tasks. This suggests that more experienced participants benefit more from the use of writing tools.

OLS Regression We ran four Ordinary Least Squares (OLS) regression models using rhyme density scores from each task condition as dependent variables. Predictors included DAT score and experience level. Results indicated: (i) Baseline rhyme performance was modestly predicted by DAT score; (ii) experience level significantly predicted rhyme performance in support and co-creative conditions; and (iii) in the autonomous condition, both DAT score and experience were moderate predictors of output quality.

Predicting Perceived Tool Usefulness We conducted OLS regressions to examine whether DAT scores and baseline rhyme performance predicted perceived tool usefulness. Participants with higher baseline rhyme scores tended to rate support and co-creative tools as more useful. However, neither DAT scores nor experience level significantly predicted subjective ratings of tool usefulness.

Analysis	Key Results	Significance
Pearson Correlation	DAT vs. Base Rhyme Score	r = 0.41, p = 0.023
	DAT vs. Support Score	r = 0.39, p = 0.030
	DAT vs. Co-creative Score	r = 0.37, p = 0.039
	DAT vs. Autonomous Score	r = 0.36, p = 0.044
OLS Regression	Support Score predicted by Experience	p = 0.002
	Co-creative Score predicted by Experience	p = 0.020
	Autonomous Score predicted by Experience	p = 0.033
ANOVA	Support Score by Experience Level	F = 6.34, p = 0.005
	Co-creative Score by Experience Level	F = 4.67, p = 0.018
	Autonomous Score by Experience Level	F = 3.99, p = 0.030

Table 4: Summary of Significant Statistical Results.

Discussion

This study examined how rhyme density, experience level, and divergent thinking (DAT scores) relate to performance in computationally-assisted creative writing tasks. Experienced writers benefited most from tool use, aligning with prior research in creativity support systems (Cherry and Latulipe 2014). One explanation may lie in *Cognitive Load Theory*, which suggests that cognitive demands interact with user expertise. Less experienced participants may have been overwhelmed by the dual challenge of constrained poetic tasks and unfamiliar tools, while experienced writers could apply tools more strategically (Sweller 1988b; Van Merrienboer and Sweller 2005; Sweller 1988a)

The weak correlation between experience level and baseline rhyme scores suggests that self-reported writing experience—often in non-fiction—did not reliably translate into poetic skill. Some less experienced participants, likely more familiar with poetic form, outperformed more experienced peers.

Tool performance varied with experience. Support tools most improved rhyme density for experienced writers, while no tool consistently aided novice users—likely due to a skills gap limiting their ability to act on system suggestions. Although DAT scores and experience were positively associated with rhyme performance, neither significantly predicted perceived tool usefulness. Qualitative feedback, especially around expectations and tool interpretation, may help explain these perceptual gaps.

Participants found support and autonomous tools easier to use and more helpful for fluency, while co-creative tools were rated as more cognitively demanding. Of the three, autonomous tools yielded the greatest fluency benefit. However, none led to statistically significant improvements in perceived creativity or accuracy.

DAT scores showed a positive trend with age, and female participants scored higher on average with greater internal consistency. While promising, these trends are difficult to interpret given the modest sample size. The generally high DAT scores may also have inflated observed correlations. Future work should examine whether computational tools support creativity or simply reflect the abilities of an already creative cohort, using a more diverse sample and broader range of divergent thinking levels.

Limitations Several limitations should be noted. First, tools were presented in a fixed order; although participants could choose how to engage with them, usage sequence was not strictly controlled and may have influenced outcomes. Second, apart from the DAT, writing tasks were untimed, introducing variability in pacing and effort. Third, the modest sample size (N = 30) limits the generalisability of statistical findings. Additionally, the study was conducted entirely online without lab-based control, which may have introduced variability in attention, environment, or device setup (Uittenhove, Jeanneret, and Vergauwe 2023; Dandurand, Shultz, and Onishi 2008; McConnell, Hintz, and Meyer 2025). Finally, participant-reported experience may not align with actual writing ability, particularly across genres. Future studies should use a more diverse sample, control task sequencing more tightly, and consider in-person or hybrid protocols to evaluate tool use under consistent conditions.

Future Work

Future work could investigate how participants' baseline vocabulary affects tool engagement, particularly across native and non-native speakers. Individuals with a broader lexical repertoire may be better equipped to respond creatively to AI-generated suggestions. While DAT scores offer a proxy for divergent thinking, vocabulary profiling tools such as LexTALE or the New General Service List (NGSL) could provide a more direct measure of lexical range (Lemhöfer and Broersma 2011; Brezina and Gablasova 2013). Comparing lexical use with perceived tool usefulness may help isolate how language proficiency shapes co-creative collaboration.

Co-creative tools also warrant refinement to support more intuitive, user-guided interaction. Prior work shows these systems often underperform when users cannot anticipate or steer system responses (Lehmann et al. 2022). More adaptive interfaces, clearer affordances, and responsive feedback mechanisms could lower the barrier for less experienced writers and improve perceived usefulness across user groups.

Finally, creativity evaluation itself should be broadened beyond rhyme and lexical features to include affective tone, rhythm, or originality. A richer, multidimensional framework would better reflect how divergent and convergent processes unfold across tool types, and how users express creativity in response to different levels of assistance.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council.

References

Arnold, K.; Volzer, A.; and Madrid, N. 2021. Generative models can help writers without writing for them. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

Brezina, V., and Gablasova, D. 2013. Is there a core general vocabulary? introducing the new general service list. *Applied Linguistics* 36(1):1–22.

Cherry, E., and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Trans. Comput.-Hum. Interact.* 21(4).

Condit-Schultz, N. 2016. *MCFlow: A Digital Corpus of Rap Flow*. Ph.D. Dissertation, The Ohio State University.

Dandurand, F.; Shultz, T.; and Onishi, K. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40:428–34.

Google Research. 2022. Verse by Verse.

Hadaway, N. L.; Vardell, S. M.; and Young, T. A. 2001. Scaffolding oral language development through poetry for students learning english. *The Reading Teacher* 54:796–796.

Hirjee, H., and Brown, D. 2009. Automatic detection of internal and imperfect rhymes in rap lyrics. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, 711–716.

Karimi, P.; Grace, K.; Maher, M.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems.

Lehmann, F.; Markert, N.; Dang, H.; and Buschek, D. 2022. Suggestion lists vs. continuous generation: Interaction design for writing with generative models on mobile devices affect text length, wording and perceived authorship. In *Proceedings of Mensch Und Computer 2022*, MuC '22, 192–208. New York, NY, USA: Association for Computing Machinery.

Lemhöfer, K., and Broersma, M. 2011. Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods* 44:325–343.

Malmi, E.; Takala, P.; Toivonen, H.; Raiko, T.; and Gionis, A. 2016. Dopelearning: A computational approach to rap lyrics generation. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

McConnell, K.; Hintz, F.; and Meyer, A. 2025. Individual differences in online research: Comparing lab-based and online administration of a psycholinguistic battery of linguistic and domain-general skills. *Behavior Research Methods* 57:22.

Olson, J. A.; Nahas, J.; Chmoulevitch, D.; Cropper, S. J.; and Webb, M. E. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences* 118(25):e2022340118.

OpenAI. 2024. ChatGPT Overview.

Rezwana, J.; Maher, M.; and Davis, N. 2021. Creative penpal: A virtual embodied conversational ai agent to improve user engagement and collaborative experience in human-ai co-creative design ideation. In *Joint Proceedings of the ACM IUI 2021 Workshops, College Station,*.

Rojcewicz, S. 2004. Poetry therapy in ancient greek literature. *Journal of Poetry Therapy* 17(4):209–213.

Sweller, J. 1988a. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2):257–285.

Sweller, J. 1988b. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2):257–285.

Uittenhove, K.; Jeanneret, S.; and Vergauwe, E. 2023. From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*.

Van Merrienboer, J. J. G., and Sweller, J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17:147–177.

Wassiliwizky, E.; Koelsch, S.; Wagner, V.; Jacobsen, T.; and Menninghaus, W. 2017. The emotional power of poetry: neural circuitry, psychophysiology and compositional principles. *Social Cognitive and Affective Neuroscience* 12:1229 – 1240.

WriteExpress. 2022. Rhymer online rhyming dictionary.