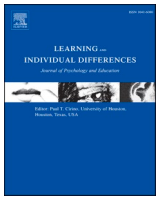
Learning and Individual Differences 118 (2025) 102626

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif



What is creative in childhood writing? Computationally measured linguistic characteristics explain much of the variance in subjective human-rated creativity scores

Birsu Kandemirci [a], Roger E. Beaty [b], Dan Johnson [c], Bonamy R. Oliver [d], Yulia Kovas [e], Teemu Toivainen [e,*]

[a] The University of Manchester, United Kingdom of Great Britain and Northern Ireland
[b] Pennsylvania State University, United States of America
[c] Washington and Lee University, United States of America
[d] UCL Institute of Education, London, UK
[e] Goldsmiths, University of London, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Creativity
Assessment
Computational
Narrative
Development

ABSTRACT

The present study investigated the extent to which linguistic features of children's stories (analysed using automated techniques), predicted human-rated Creative Expressiveness and Logic scores (both assessed with the Consensual Assessment Technique). A sample of 160 children (Mage = 8.99 years, SD = 0.3) wrote stories based on three pictures. Eleven linguistic characteristics were measured: Length, Grammar, Originality, Controlled Lexical Diversity, Uncontrolled Lexical Diversity, Divergent Semantic Integration (DSI), Referential Cohesion, Narrativity, Syntactic Simplicity, Word Concreteness and Deep Cohesion. The results showed that 51 % of the variance in Creative Expressiveness was explained by Length, DSI, Originality, Grammar, and Controlled Lexical Diversity (sr^2 = 0.01 to 0.14). In comparison, 28 % of the variance in Logic scores was accounted for by DSI, Grammar, Controlled Lexical Diversity, Syntactic Simplicity, and Narrativity (sr^2 = 0.01 to 0.06). These findings offer insights for educational practices by identifying the linguistic characteristics relevant to children's creative writing as opposed to logical narration.

1. Introduction

The assessment of creative content of educational outputs, such as written narratives, is needed to recognise the skills which contribute to verbal creativity at various points of development. The process leading to creative products, such as a creative written text, is context dependent and an outcome of many skills and resources, such as domain-specific skills, creative skills, and intrinsic motivation (Amabile, 2019). Recently, computational approaches have been developed to measure linguistic characteristics which correlate with human rated creativity assessments in text and verbal tests (e.g., Beaty et al., 2022; Johnson et al., 2022; Organisciak et al., 2023; Zedelius et al., 2019). These approaches overcome some limitations of subjective creativity measurements (such as Consensual Assessment Technique (CAT); Amabile, 1983) which are time consuming and labour intensive due to the ratings from multiple judges. The results have supported the use of computational approaches in the scoring of cognitive creativity tasks (Beaty et al., 2022; Organisciak et al., 2023). A strong positive correlation has also been reported between human rated creativity and computational approaches in short narratives among adult samples (Johnson et al., 2022).

To build on this work, the present study investigated the predictive value of linguistic characteristics on human-rated creative content in childhood writing. This is important for two reasons. First, it is important to identify the linguistic characteristics that impact the human-rated assessment of creative content in childhood writing. This identification can provide educators with valuable information on specific linguistic features and underlying skills that can be targeted to support and enhance creative expression in children's writing. Second, it is useful to understand whether various linguistic characteristics, such as the diversity of used words and grammatical correctness, are differently relevant in the production of creative content among young writers. This

* Corresponding author.
E-mail address: t.toivainen@gold.ac.uk (T. Toivainen).

https://doi.org/10.1016/j.lindif.2025.102626
Received 27 September 2023; Received in revised form 20 December 2024; Accepted 2 January 2025
Available online 11 January 2025

understanding can help in tailoring educational approaches to better foster creativity in children's writing. Often, the quality of childhood writing is evaluated in relation to technical merits (e.g., story development, organization of content, sentence structure and vocabulary, and mechanics of writing; Graham et al., 2017), with less focus on creative content.

Research literature has used a broad definition of creativity that includes various behaviours (e.g., exploration and trial-and-error), processes (e.g., design and artistic processes), and outcomes (e.g., new ideas and artistic outputs). The Standard Definition of Creativity describes creativity as a process leading to an original (or novel) and effective (or appropriate) outcome (Runco & Jaeger, 2012). This definition is generic, especially when used as a basis to operationalise creativity and to evaluate creativity in educational contexts. Several studies have addressed this issue and progress has been made in agreeing on a more detailed creativity definition (e.g., Plucker et al., 2004; Puryear & Lamb, 2020). Plucker et al. (2004) highlighted creativity as, "…a perceptible product that is both novel and useful as defined within a social context" (Plucker et al., 2004, p. 90). In other words, social context plays an important role in influencing how creativity can be measured, since as the social context changes, so does what is viewed to be original and effective, hence creative. For instance, assessment of creativity may differ for a story written by a 9-year-old child compared to a professional author. In addition, Ivcevic (2022) cautions against a generic *creativity* measurement and urges the specification of creativity aspects a particular study focuses on. Similarly, the creative domain (e.g., storytelling, drawing, composing music) must be specified since different domains may require different assessment methods and levels of prior knowledge (e.g., Baer, 2016; Han & Marvin, 2002). In addition to domain differences, creativity measures can be grouped together based on whether they focus on the creative *product* (the creative output), *process* (e.g., cognitive skills related to creative idea production), *person* (features of creative individuals), *press* (the interaction between the creative individual and the environment), or *potential* (level of ability to be creative; e.g., Lubart, 2017; Rhodes, 1961).

Stories, written by children or adults, are examples of creative *products*. Stories can provide fruitful data for creativity investigations among developmental samples (e.g., Cavanaugh et al., 2017; Hennessey & Amabile, 1988; Mohamed & Maker, 2011). A benefit of using stories in child samples is that most children are interested in storytelling and engage in storytelling as part of their educational routine (Hennessey & Amabile, 1988). In addition, stories can be investigated both linguistically and based on subjective assessments.

The developmental stage, such as that of 9-year-olds, is interesting to investigate in relation to creative content in writing. In fourth grade, children are able to write coherent sentences, but the linkages between ideas may still be lacking (McCutchen & Perfetti, 1982). Additionally, the focus in writing shifts away from basic skills, such as spelling (Graham et al., 2017). In England, the primary education curriculum for writing, a nationally set standard, mentions creativity, alongside composition, vocabulary, grammar, and punctuation (Department for Education, 2014). Creative narration can also be integrated into other subjects beyond English, for example in the Creative Arts (Hall & Thomson, 2017), History (Cooper, 2018) and Geography (Catling & Willy, 2010). Additionally, children are exposed to creative narration and its production in various contexts outside of school, such as at home (Spagnola & Fiese, 2007), through media (Linebarger & Piotrowski, 2009), or when interacting with their peers (Nelson, 2010). However, concerns have been raised about the benefits of focusing on formal grammar in primary education and how it may negatively influence writing motivation (Wyse, 2001). Students in primary education are already aware of the importance of National Curriculum assessments and the content they focus on (Reay & Wiliam, 1999). Additionally, excessive curriculum content and standardised assessments have been viewed by teachers as having a negative impact on overall teaching quality in English (Webb & Vulliamy, 2006).

One of the most frequently used techniques to measure creativity of products is the Consensual Assessment Technique (CAT; Amabile, 1982). The method relies on the agreement between independent judges' subjective evaluations of creativity of products. That is, judges are asked independently to rate the creative value of an output, with a high consensus (i.e., a high interrater reliability) among the judges seen to suggest validation of the creativity of the output. An important aspect of the CAT measurement is that no formal instructions are given to the judges regarding the features which should be considered when conducting the assessments of creativity. Instead, the judges rely on their subjective individual assessment on which products they view to be more creative in comparison to others. Traditionally it was seen as preferable for judges to have some expertise in the domain of the creative product that they are assessing (Kaufman et al., 2007). However, one study found no difference between undergraduate Psychology students' and primary school teachers' assessment of creative content in children's stories (Toivainen et al., 2017).

The CAT has been used to measure creativity of various outputs, such as collages, poems, paintings, and stories (e.g., Kaufman et al., 2007). Alongside its advantages, the CAT has also been criticised for its shortcomings. First, it requires extensive time and effort to measure creative outputs with the CAT, which is due to the laborious nature of the process (e.g., Baer & McKool, 2009). Second, certain personality traits of the judges, such as novelty preference (Storme & Lubart, 2012) or level of openness (Silvia, 2008), might influence the creativity assessment. Third, to assess complicated or lengthy creative outputs, in comparison to concisely formulated ideas, can lead to a lower interrater reliability among the judges (Forthmann et al., 2017). A reason for this may be the fatigue associated with higher cognitive load when assessing more complex creative outputs.

A previous study used the CAT to assess children's creative story writing at age 9 (Toivainen et al., 2021), and a subsample of data from this study were used in the current study. The study investigated the relation of creativity with educationally relevant measures of cognitive ability, English grades, and motivation to write — all measured at age 9. The data were a subsample of a longitudinal twin sample, the Twins Early Development Study (TEDS; Rimfeld et al., 2019). The study reported positive associations between creative content in the stories with intelligence, motivation to write and English grades for writing (rs = 0.24–0.40). The study also reported weak longitudinal associations between creativity at age 9 and end of school exams at age 16, above and beyond intelligence, motivation to write and English grades for writing at ages 12 and 14 ($sr^2 = 0.02$). The study also reported that Logic scores, based on human-rated assessment of logical narration of the same stories, were associated with intelligence, motivation to write and English grades (Toivainen et al., 2021). However, the study did not explore the specific linguistic characteristics of the stories which may be associated with the creativity scores.

Linguistic characteristics have been used in previous research to predict what constitutes good writing, for example in the context of essay writing (e.g., Crossley et al., 2011; McNamara et al., 2015). These studies used techniques such as Coh-Metrix (Graesser et al., 2004; a measure focusing on text difficulty, structure, and cohesion) and Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2014; focusing on the word categories in a text). A study, which evaluated college students' creative short stories using Coh-Metrix and LIWC, reported that linguistic characteristics such as narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion predicted the subjective judgements of image, voice, and originality of creative stories (Zedelius et al., 2019).

In a study investigating children's creativity in oral storytelling, Hennessey & Amabile (1988) recruited 9 independent judges to evaluate children's stories on 10 dimensions: *creativity*, *liking*, *novelty*, *imagination*, *logic*, *emotion*, *grammar*, *detail*, *vocabulary*, and *straightforwardness*. The same dimensions were used in a later study focusing on children's handwritten stories (Toivainen et al., 2021). As some of the dimensions

were highly intercorrelated, two principal component scores, *Creative Expressiveness* and *Logic*, were created (Toivainen et al., 2021). Creative Expressiveness was considered as a measure of creative content in the stories, whereas the Logic score reflected the level of logical narration. These component scores were also used in the current study, as the dataset used in the current study was derived from this 2021 study (see Method section for the details).

Hennessey & Amabile (1988) argued that it might not be possible to devise a method to measure novelty and appropriateness objectively and therefore "(…) it may be wise – at least for research purposes – to rely on explicitly subjective judgements of creativity by observers (…)". However, while some of the dimensions listed above are subjective in nature, some linguistic characteristics can be computationally and objectively measured as indicators of creative content in children's stories. Such objective and automated techniques can make creativity measurement more applicable in educational settings by eliminating the need for multiple independent judges to evaluate the creativity of stories.

To date, it is unclear if the human-rated creative content in childhood writing is more dependent on technical skills, such as grammar or more influenced by the use of diverse ideas and words. This information would have educational implications. For example, if diverse ideas and original words contribute to the creative value of childhood storytelling, perhaps they should be recognised in their own right, instead of focusing solely on technical skills. Previous research has reported that formative feedback enhances writing quality (Mackenzie et al., 2020), however, this should not only be focused on writing skills over creative content, if the aim is to scaffold creative expression in writing.

The current study investigates linguistic features of children's stories which may contribute to subjective creativity assessments by independent human judges. To our knowledge, no previous studies have investigated the influence of linguistic characteristics on the creative content in children's writing — e.g., is it only due to grammatical correctness or is the diversity of the used words more relevant? To understand the relevance of linguistic characteristics on creative content in childhood writing might have benefits in educational context when the aim is to enhance creative storytelling. When evaluating the creativity of a product, the use of multiple methods of measurement can provide researchers with a more comprehensive understanding of the output's creative value (Lubart et al., 2010). This is important given that it is not always clear what thought processes judges go through when they evaluate the creative value of a product (Caroff & Massu, 2022). It is likely that linguistic characteristics play a role in the subjective assessment of creativity of written stories. For example, a judge may attribute higher creative value to a story if it is written with rich, lexically diverse, and elaborate language, with little or no grammatical mistakes or typos. The writer's proficiency in language is also likely to impact the judge's decision about the writer's creative skills.

To assess the relationship between human-rated Creative Expressiveness and Logic scores and computationally measured linguistic characteristics, the present study used three linguistic tools: Computerised Language Analysis program (CLAN; MacWhinney, 2018), Distributional Semantic Integration modelling (DSI; Johnson et al., 2022) and Coh-Metrix (Graesser et al., 2004). We recognise that other linguistic tools are also available. To provide information on the predictive power of these methods in human rated creativity across varied forms of texts and samples, studies need to explore their contribution to human rated creativity, alongside other metrics.

CLAN is a program which evaluates various aspects of narratives transcribed in the CHAT (Codes for the Human Analysis of Transcripts) format (MacWhinney, 2018). The CHAT format allows CLAN to analyse the narratives in terms of their linguistic features such as frequency of words and type-to-token ratios. Primarily, CLAN is used to analyse conversational interactions and young children's language learning but given the affordances of the program and its user-friendly layout, this free program has been used in a variety of studies both with children and adults (MacWhinney, 2018). CLAN produces detailed linguistic analyses following application of lines of codes but also requires the user input during the CHAT transcription process. We used CLAN to measure Length (the number of words used in the stories), Lexical Diversity (different word types used in the stories, controlled and uncontrolled for the length of the story), Grammar (the ability to follow the grammatical and spelling rules of the language) and Originality (proportion of low-frequency words based on the overall pool of words used within the dataset) of the stories.

DSI is a new automated measurement technique that evaluates the semantic divergence of words in a piece of creative writing (Johnson et al., 2022). Informed by the distributional semantics theory (Firth, 1957), DSI measures the extent to which the writer combines divergent words in a piece of text. To do this, DSI creates pair-wise comparisons of all words in the text. To assign a value to estimate the distance of the words in a semantic space, DSI produces a semantic distance value. The semantic distance value indicates the level of uniqueness of the words appearing in a piece of text, based on a large English corpus. All semantic distance values, based on all the word pairs in the text, are averaged to come up with the total DSI score. The DSI model used in the present study (out of six), is the Bidirectional Encoder Representations from Transformers (BERT) model. A benefit of BERT is that it also generates context-dependent numerical representations of words (i.e., embeddings). For example, the word "bark" would be assigned with different numerical representations depending on the context in a sentence (e.g., loud dog bark vs. rough bark on a tree). Previous research has reported moderate to strong predictive power of the DSI output on human-rated creativity assessment of short narratives, written by adults (up to $R^2 = 0.72$; Johnson et al., 2022). Tools such as DSI are helpful to reduce human input and associated subjectivity in creativity measurement. In the present study, we used DSI to evaluate children's ability to combine semantically divergent concepts in a story.

Coh-Metrix is an automated tool for analyzing over 200 linguistic metrics in text and discourse (Crossley et al., 2011). In previous research, Coh-Metrix has been applied for example in the investigations of linguistic characteristics influencing essay quality assessment (Crossley et al., 2011; Graesser et al., 2004; McNamara et al., 2015). A previous study also investigated the relationship between a measure of referential cohesion, produced by Coh-Metrix, and human-rated originality scores (Zedelius et al., 2019). The study reported a negative correlation between Referential Cohesion and Originality scores of creative writing texts by 133 undergraduate students.

The research questions for the study are:

1) Which linguistic characteristics of children's stories best predict human-rated creativity?
2) How much variance do linguistic characteristics explain in human-rated creativity scores?

We hypothesised that all eleven linguistic aspects of stories (length, controlled lexical diversity, uncontrolled lexical diversity, grammatical accuracy, originality of the words used, the DSI scores, Referential Cohesion, Narrativity, Syntactic Simplicity, Word Concreteness and Deep Cohesion) would predict the Creative Expressiveness assessed by independent human judges. To contrast the predictive power of linguistic characteristics on Creative Expressiveness, we also included a human-rated Logic score as an outcome.

2. Method

2.1. Participants

The participants in the current study were 160 children who took part in the Twins Early Development Study (TEDS). TEDS, initiated in 1994–96, is a large, longitudinal, and representative twin study in the UK (Rimfeld et al., 2019). The participants in the current study were selected from a smaller TEDS subsample ($N = 1306$) whose written

childhood stories were subjectively rated for creativity as part of an earlier study (Toivainen et al., 2021). One twin per twin pair was randomly selected to exclude the inflated inter-individual similarity observed in twins. The mean age for the participants was 8.99 years (*SD* = 0.27) and the final sample included 99 girls and 61 boys. For all children, English was the main language spoken at home. The TEDS study received ethical approval from the King's College London Ethics Committee.

2.2. Materials

2.2.1. Stories based on a picture sequence

At the age of 9, TEDS participants were shown three coloured pictures of animals and buildings in a farm (see Fig. 1). The participants were given the following instructions: '*We would like you to make up a story for us. On the next page you will see three different pictures, 1, 2 and 3. Together they make a little story about a farm. Try to think hard about what you see in the pictures. After you have looked at them carefully, write your story on the next page of this book. Have fun making your story interesting, creative, or even funny!*' The children completed the task in their homes, supervised by their parents. There was no time limit for the task. In Toivainen et al. (2021) study all the stories were transcribed using a word processor to minimize the effects of children's handwriting on creativity assessment.

The present study used three pieces of software to explore the linguistic characteristics of the stories: Computerised Language Analysis (CLAN; MacWhinney, 2018), Divergent Semantic Integration (DSI; Johnson et al., 2022) and Coh-Metrix (Graesser et al., 2004). To process the data using CLAN, the transcribed stories were edited following the CHAT format. For instance, capital letters at the beginning of the sentences were edited to be lowercase to avoid CLAN from counting these as proper nouns (for details of CHAT formatting see MacWhinney, 2018). During this process, no changes were made to the original scripts. For instance, if a child made a grammatical mistake (e.g., "the pig *goed* to the barn" instead of "the pig *went* to the barn"), the words were inputted as written down by the children. Similarly, the typos that the children made were transcribed as the participants wrote them (e.g., *littel* instead of *little*). Similarly, no changes were made in the text when it was analysed with DSI and Coh-Metrix.

2.3. Human ratings

2.3.1. Creative expressiveness and logic scores

As part of the previous study, the stories were rated for their creativity and nine other dimensions by a group of judges (Toivainen et al., 2021). In the previous study, the stories were divided in 5 blocks, due to a large sample size (n = 1306). Stories in each block were rated for 10 story dimensions by 5 judges, with a total of 25 judges. The inter-rater reliabilities were acceptable for most dimensions. They were slightly below the recommended 0.70 for Straightforwardness (0.63) and Logic (0.66). Due to high intercorrelations between story dimensions, two principal component scores, *Creative Expressiveness* and *Logic*, were created based on the scores for 6 (Creativity, Imagination, Novelty, Liking, Emotion and Detail) and 3 (Straightforwardness, Logic, and Grammar) story dimensions, respectively. One dimension, Vocabulary, was excluded from the principal component scores due to the similar loading on both components. For detailed information on the coding and creation of the component scores, see Toivainen et al. (2021). In the current study, the Creative Expressiveness scores were used as a human-rated measure of creativity and Logic scores as a human-rated measure of logic. The Creative Expressiveness and Logic scores were not significantly correlated.

The present study used the composite score of Creative Expressiveness, instead of relying only on the specific creativity dimension, due to the high correlations between creativity and other dimensions in the previous study. However, the Creative Expressiveness score is a very close proxy for the creativity dimension score with a correlation of r = 0.91. The Logic scores were included in the study to provide a comparison point to investigate if the same linguistic features, and to what extent, predicted creative vs logical content in childhood writing.

2.4. Computational ratings

2.4.1. Length (CLAN)

The length of each story was calculated using the Frequency (freq) function on CLAN. When "freq" is applied to a file, it calculates the type (number of different types of words, such as words with different roots and words that share the same root but have different affixes), token (number of words, including repetitions), and type-to-token ratio (the proportion of type to token). To illustrate, in the sentence "I have three dolls; one doll with black hair and two dolls with blonde hair", there are 15 tokens (number of words, including repetitions) and 13 types (as the words "hair" and "dolls" are repeated), and type-to-token ratio is 13/15 = 0.87. We used the token value for the length variable as it calculates the number of words (including the repeating words) in each story.

2.5. Lexical diversity (CLAN)

The study employs two measures of lexical diversity: Controlled and Uncontrolled Lexical Diversity. Controlled Lexical diversity, controlled for the story length, is a measure of Moving Average Type to Token Ratio (MATTR; Covington & McFall, 2010; MacWhinney, 2018). Given that the children produced their stories with no time limitation, the length of the stories varied greatly (from 13 to 479 words). To account for the differences in story lengths when calculating the lexical diversity scores, the present study specified the number of words ('a window') which were analysed as separate entities (Covington & McFall, 2010). For example, in a story of 300 words, a 10-word-window can be created to calculate the lexical diversity within that window. This window is then moved by one word, and therefore a new value is calculated among words 2–11, and then another one among 3–12 and so on. Following this, CLAN calculated the mean value of all 10-word-windows. The CLAN code for MATTR is "+b(n)" with "n" being the size of the word window that is used, e.g., 10 in the case of the current study. The 10-word window was selected as it is small enough to allow the shortest story's MATTR to be calculated while also allowing for enough

Fig. 1. The picture sequence for participants' stories.

variability within a window. The uncontrolled lexical diversity scores are the Type to Token Ratio scores (TTR), which were not controlled for the length of the stories.

2.5.1. Grammar (CLAN)

To assess the grammatical correctness with CLAN, the stories were read by two independent coders to evaluate their grammatical and linguistic accuracy. The coders were provided with this instruction: "Read the stories and flag any diversions from the rules of English language, including grammatical mistakes and spelling errors." To highlight these diversions, the coders added a code at the end of the relevant words (*@n*) which, in CLAN, refers to neologism. This code is used to highlight made-up or grammatically incorrect words (MacWhinney, 2018). By flagging the words which violated the grammatical rules, we were able to calculate the number of them in each story using the code freq +s*@n. Once we found the number of occurrences per story, we calculated the proportion of the words in the story without language violations to all the words. For instance, if there were 50 words in the story and there were 10 instances of violation to the language, the story had a success rate of 80 % in terms of following the rules of the language.

2.5.2. Originality (CLAN)

The originality score is the sum of the frequency scores for each word in a story based on a pool of the words used by all participants, controlled for the story length. First, all words in 160 stories were combined into one file using the CLAN code freq +o3 + u + d2 which created an overall frequency output for each word based on the number of times this word was used across 160 stories. Second, the originality score for each word was calculated by dividing the individual word frequency value by the total number of words in the sample (e.g., for word pig: 406 / 23,585 = 0.01721). Third, the total originality value for each story was calculated by summing up the originality scores for each word in individual stories. Finally, these originality scores were divided by the word count of the story to account for the differences in story lengths. The 50 most common words and their frequencies for 160 stories are listed in Table 1.

2.5.3. Divergent semantic integration (DSI)

The DSI score measures how well the writers connect divergent ideas in their text (Johnson et al., 2022). The DSI score can be run with R or Python code (see the code and supporting materials in Johnson et al., 2022). The BERT model used in this study, out of six DSI models, generates context-dependent word embeddings based on data extracted from the BooksCorpus (800 million words) and Wikipedia (2.5 million words, see Devlin et al. (2019), for more details). The word embeddings are fitted at the sentence level in order to capture the contextual differences. Compared to other DSI models, the BERT model is more powerful due to its ability to produce 24 distinct word embeddings for every word in a sentence or piece of text. Each of these embeddings is associated with different weights, which determine the significance of each word's representation relative to every other word within the sentence. In the present study, we used word embeddings of layers 6 and 7 to calculate the DSI scores, which have been indicated to be particularly attuned to the structure and meaning of language (Jawahar et al., 2019). The computation of DSI scores is based on the pairwise cosine semantic distance between all word embeddings derived from both layers.

2.5.4. Referential Cohesion (Coh-Metrix)

The Referential Cohesion measure evaluates how well a text maintains referential cohesion through the repetition and overlap of key content words and concepts (Graesser et al., 2014). A source of cohesion in text is the overlap of words and concepts between adjacent sentences and paragraphs (McNamara et al., 2010). The Referential Cohesion measure in Coh-Metrix is a standardised principal component score of various referential cohesion indices, such as noun overlap, argument overlap, and content word overlap across sentences and paragraphs (Graesser et al., 2011).

2.5.5. Narrativity (Coh-Metrix)

Narrativity is a standardised principal component score which measures how much a text aligns with a narrative or storytelling structure (Graesser et al., 2011; Graesser et al., 2014). Texts high in narrativity tend to feature characters, events, and a temporal sequence.

2.5.6. Syntactic simplicity (Coh-Metrix)

Syntactic simplicity is a standardised principal component score which assesses the complexity of sentence structures in a text (Graesser et al., 2011; Graesser et al., 2014). Short sentences with simple and familiar syntactic structures are easier to comprehend. In contrast, complex sentences often include embedded syntactic elements.

2.5.7. Word concreteness (Coh-Metrix)

Word concreteness is a standardised principal component score measuring how tangible and easy to process the words in a text are (Graesser et al., 2011; Graesser et al., 2014).

2.5.8. Deep cohesion (Coh-Metrix)

Deep cohesion is a standardised principal component score which evaluates the logical and conceptual connections in a text that help readers understand relationships between ideas (Graesser et al., 2011; Graesser et al., 2014).

Access to the Coh-Metrix tool can be requested from The Science of Learning and Educational Technology (SoLET) Lab at Arizona State University: https://soletlab.asu.edu/coh-metrix/.

2.5.9. Control variables

We included Age, Sex, Cognitive Ability, Motivation to Write and

Table 1
The frequencies for the 50 most common words in the stories.

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
The	1955	He	356	**Owls***	181	Had	141	**Barn***	108
And	1028	One	308	In	174	His	141	With	108
To	552	Were	250	Up	169	Two	130	You	106
A	528	So	230	Then	167	Him	129	Out	105
Of	454	There	211	**Farm***	157	**Birds**	128	Down	103
Was	451	All	195	Other	156	I	125	Got	102
Pig*	406	It	192	Said	151	Little	125	**Eagles**	96
On	394	But	190	Day	149	Top	123	Is	96
They	389	**Bats**	187	That	149	Came	118	Get	93
Pigs*	370	**Owl***	181	Not	146	**Flew**	109	**Roof**	92

Note. The words in bold represent the items and actions presented in the three-picture stimulus that participants based their stories on.

* CLAN counts proper nouns (e.g., Pig) as different words to regular words (e.g., pig). When the proper words are included in the calculation (the ones spelled exactly the same except for the capitalised letter, e.g., not pigs), the frequency of "pig" reaches 493, "pigs" 373, "owl" 223, "owls" 183, "farm" 173, and "barn" 115.

English Writing Grade as our control variables. The cognitive Ability score refers to a combined score of participants' performance at age 9 on two verbal and two non-verbal tests of cognitive skills measured by Vocabulary and General Knowledge tests from the WISC-III (Kaplan et al., 1999) and Figure Classification and Shapes tests from the Cognitive Abilities Test 3 (Smith et al., 2001). Motivation to Write was a mean of two questions. Children were asked 'How much do you like writing' (1–5) and parents 'How much does your child like writing' (1–5). The items were developed by the TEDS research team (Spinath et al., 2006). English Writing Grades were assigned by teachers who provided marks on the level of attainment in writing in terms of the National Curriculum (NC) in a scale of 1 to 5. The assessment criteria were based on grammar, punctuation and spelling (NC level Key Stage 2).

2.6. Statistical analyses

The analyses, which were hierarchical regressions with stepwise selection of independent variables, were conducted using the IBM SPSS Statistics, Version 27. The stepwise selection of independent variables was carried out for the second step of the regression. In SPSS, stepwise regression relies on the Ordinary Least Squares (OLS) method to estimate the coefficients at each step of the selection process, ensuring that the best linear predictors are identified based on the specified criteria. For the stepwise regressions, we used Stepwise Selection which combines forward selection and backward elimination, adding or removing predictors based on the chosen criteria. The default significance level for adding a variable to the model is 0.05 and to remove 0.10.

Methods and analyses for this project were pre-registered using the open science framework (https://osf.io/tqnrx/?view_only=fda10c781bc04cc7834af6410c7e1b28). Amendments to the measures and statistical analyses were made based on the reviewers' recommendations.

3. Results

Table 2 presents descriptive statistics for each measure included in analyses.

The frequency distributions for Age, Length, Grammar, Controlled Lexical Diversity (MATTR), Referential Cohesion, Syntactic Simplicity, Word Concreteness and Deep Cohesion indicated non-normality due to the presence of outliers. Upon further inspection, these outlier scores were determined to be accurate representations of the measured characteristics and were retained in the data set. To ensure the robustness of our findings, we reran the analyses with the outliers removed. The results showed minimal changes when the outliers were excluded (see Supplementary materials).

Table 2
Descriptive statistics.

	N	Min score	Max score	Mean	SD
Creative expressiveness	160	−2.2767	2.4000	0.0027	1.0681
Logic	160	−2.7968	2.7261	−0.0527	1.0229
Age	160	8.55	10.05	8.99	0.2671
Cognitive ability	160	−2.0850	1.8310	0.1375	0.8819
Writing motivation	160	1	4	2.01	0.79
English Writing Grade	160	1	4	2.96	0.70
Length	160	13	479	147.43	82.09
Grammar	160	53.85	100	93.86	5.91
Originality	160	0.9756	0.9928	0.9851	0.0035
Controlled Lexical Diversity (MATTR)	160	0.7600	0.9860	0.9270	0.0322
Uncontrolled Lexical Diversity (TTR)	160	0.3260	0.8750	0.5764	0.0986
DSI	160	0.7789	0.8592	0.8206	0.0150
Referential cohesion	160	−1.91	5.69	0.76	1.21
Narrativity	160	−2.79	2.69	0.55	0.94
Syntactic simplicity	160	−6.92	2.70	−0.55	1.59
Word concreteness	160	−0.18	7.92	2.02	1.34
Deep cohesion	160	−4.53	3.91	0.05	1.50

Note. Creative Expressiveness, Logic and Cognitive ability scores are standardised z-scores.

Table 3 presents the correlations between the study variables. Cognitive ability, Writing Motivation, English Writing Grade, Length, Grammar, Originality, Controlled Lexical Diversity, DSI, Narrativity, Syntactic Simplicity and Deep Cohesion had positive correlations with Creative Expressiveness ($rs = 0.17$–0.71). Uncontrolled Lexical Diversity ($r = -0.45$), Referential Cohesion ($r = -0.27$) and Word Concreteness ($r = -0.46$) were negatively correlated with Creative Expressiveness. Cognitive ability, Writing Motivation, English Writing Grade, Length, Grammar, Controlled Lexical Diversity, Narrativity and Syntactic Simplicity had a positive correlation with Logic ($r = 0.17$–0.53). Uncontrolled Lexical Diversity ($r = -0.18$), DSI ($r = -0.18$), Referential Cohesion ($r = -0.19$) and Word Concreteness ($r = -0.27$) had negative correlations with Logic.

To address which linguistic characteristics contribute, and to what extent, to the variance in human-rated creativity assessment of childhood writing, we ran a hierarchical regression with stepwise selection of independent variables. The stepwise selection of independent variables was carried out for the second step of the regression. This model provides information on the total variance explained by the linguistic predictor variables, as well as their separate contributions to human-rated creativity.

3.1. Predictors of Creative Expressiveness

To investigate the predictive value of linguistic characteristics on Creative Expressiveness, a hierarchical regression with stepwise selection of independent variables was conducted. In the first step, the control variables Age, Sex, Cognitive Ability, Motivation to Write and English Writing Grade were added to the model. In the second step, the linguistic predictor variables of Length, Grammar, Originality, Controlled Lexical Diversity, Uncontrolled Lexical Diversity, DSI, Referential Cohesion, Narrativity, Syntactic Simplicity, Word Concreteness and Deep Cohesion were entered to the model. Table 4 summarises the results. The step 2 results include only the predictor variables which were statistically significant predictors of Creative Expressiveness (see Methods section for the selection criteria). To test assumption of multivariate normality, the histogram of standardised residuals indicated that the errors were approximately normally distributed. The assumption of homoscedasticity was checked with the standardised residual plots against the unstandardised predicted values. The multicollinearity values (VIF) were also within an acceptable range (see Supplementary materials).

In total, the linguistic measures explained an additional 51 % of the variance in Creative Expressiveness, over and beyond Age, Sex, Cognitive Ability, Motivation to Write and English Writing Grade. Out of eleven linguistic variables, the final model included Length ($sr^2 = 0.14$), DSI ($sr^2 = 0.07$), Originality ($sr^2 = 0.03$), Grammar ($sr^2 = 0.02$), and Controlled Lexical Diversity ($sr^2 = 0.02$) as statistically significant predictors of Creative Expressiveness ($F(10, 149) = 27.51$, $p < .001$, $R^2 = 0.65$).

3.2. Predictors of Logic

To investigate the linguistic measures which explain variance in the Logic scores, we ran a hierarchical regression with stepwise selection of independent variables. In the first step, the control variables Age, Sex, Cognitive Ability, Motivation to write and English writing were added in the model. In the second step, the linguistic predictor variables of Length, Grammar, Originality, Controlled Lexical Diversity, Uncontrolled Lexical Diversity, DSI, Referential Cohesion, Narrativity, Syntactic Simplicity, Word Concreteness and Deep Cohesion were entered to

Table 3
Correlations between the predictor and control variables.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Creative expressiveness	1																
2. Logic	0.12	1															
3. Age	0.07	−0.01	1														
4. Cognitive ability	0.25**	0.29**	0.16*	1													
5. Writing motivation	0.17*	0.17*	0.04	0.01	1												
6. English Writing Grade	0.29**	0.38**	0.14	0.36**	0.21**	1											
7. Length	0.71**	0.25**	0.11	0.20*	0.23**	0.18*	1										
8. Grammar	0.29**	0.53**	0.04	0.25**	0.17*	0.43**	0.29**	1									
9. Originality	0.46**	0.15	0.04	0.10	0.05	0.22**	0.34**	0.06	1								
10. Controlled Lexical Diversity	0.26**	0.30**	0.05	0.28**	0.05	0.25**	0.24**	0.13	0.56**	1							
11. Uncontrolled Lexical Diversity	−0.45**	−0.18*	−0.01	0.05	−0.13	−0.11	−0.70**	−0.34**	0.07	0.22**	1						
12. DSI	0.52**	−0.18*	0.00	0.25**	−0.01	0.08	0.37**	−0.17*	0.45**	0.46**	−0.02	1					
13. Referential Cohesion	−0.27**	−0.19*	−0.10	0.22**	−0.14	−0.16	−0.17*	−0.13	−0.36**	−0.38**	−0.17*	−0.29**	1				
14. Narrativity	0.38**	0.30**	0.04	0.13	0.08	0.19*	0.38**	0.24**	0.43**	0.24**	−0.30**	0.13	0.09	1			
15. Syntactic simplicity	0.18*	0.32**	0.06	0.11	0.15	0.10	0.21**	0.17*	0.10	0.03	−0.18*	−0.14	−0.48**	0.16*	1		
16. Word concreteness	−0.46**	−0.27**	−0.02	−0.17*	−0.12	−0.27**	−0.46**	−0.24**	−0.43**	−0.21**	0.41**	−0.21**	0.33**	−0.59**	−0.51**	1	
17. Deep cohesion	0.38**	0.16*	0.08	0.20*	−0.01	0.23**	0.33**	0.17*	0.31**	0.16*	−0.30**	0.24**	0.09	0.63**	0.03	−0.36**	1

* $p < .05$.
** $p < .01$.

Table 4
Hierarchical regression, with stepwise selection of independent variables, predicting Creative Expressiveness.

Variable	β	sr^2	R	R^2	ΔR^2
Step 1			0.37	0.14	0.14
Age	0.01	0.00			
Sex	−0.15	0.02			
Cognitive ability	0.18*	0.03			
Motivation to write	0.08	0.05			
English Writing Grade	0.17*	0.02			
Step 2			0.81	0.65	0.51
Age	−0.00	0.00			
Sex	−0.03	0.00			
Cognitive ability	0.02	0.00			
Motivation to write	0.01	0.00			
English Writing Grade	0.09	0.01			
Length	0.47**	0.14			
DSI	0.36**	0.07			
Originality	0.21**	0.03			
Grammar	0.18**	0.02			
Controlled Lexical Diversity	−0.18**	0.02			

** $p < .01$.
* $p < .05$.

the model. Table 5 summarises the results from the regression. To test assumption of multivariate normality, the histogram of standardised residuals indicated that the errors were approximately normally distributed. The assumption of homoscedasticity was checked with the standardised residual plots against the unstandardised predicted values. The multicollinearity values (VIF) were also within an acceptable range (see Supplementary materials).

In total, the linguistic measures explained an additional 28 % of the variance in Logic, over and beyond Age, Sex, Cognitive Ability, Motivation to Write and English Writing Grade. The final model included DSI ($sr^2 = 0.06$), Grammar ($sr^2 = 0.06$), Controlled Lexical Diversity ($sr^2 = 0.06$), Syntactic Simplicity ($sr^2 = 0.02$) and Narrativity ($sr^2 = 0.01$) as statistically significant predictors of Logic ($F(10, 149) = 14.03, p < .001$, $R^2 = 0.49$).

4. Discussion

The present study investigated the predictive power of eleven computationally derived linguistic characteristics on human rated creative content in childhood writing. The results demonstrated that story length, the use of divergent ideas, originality of the words used, grammatical correctness, and lexical diversity (controlled for story length)

Table 5
Hierarchical regression, with stepwise selection of independent variables, predicting Logic.

Variable	β	sr^2	R	R^2	ΔR^2
Step 1			0.45	0.21	0.21
Age	−0.08	0.01			
Sex	−0.15	0.02			
Cognitive ability	0.20*	0.03			
Motivation to write	0.06	0.00			
English Writing Grade	0.27**	0.06			
Step 2			0.70	0.49	0.28
Age	−0.09	0.01			
Sex	−0.05	0.00			
Cognitive ability	0.15*	0.02			
Motivation to write	0.04	0.00			
English Writing Grade	0.11	0.01			
DSI	−0.30**	0.06			
Grammar	0.29**	0.06			
Controlled Lexical diversity	0.28**	0.06			
Syntactic Simplicity	0.17**	0.02			
Narrativity	0.13*	0.01			

** $p < .01$.
* $p < .05$.

predicted the human-rated Creative Expressiveness scores. A comparative regression model showed differences in predictors, as well as in their strength and direction, when explaining variance in the Logic score. This highlights the differences in linguistic characteristics contributing to creative versus logical childhood narratives. In what follows, we discuss the linguistic predictors of Creative Expressiveness in turn.

4.1. Length

Story length was the strongest predictor of Creative Expressiveness scores, suggesting that human judges intuitively attribute higher creative value to longer stories, as shown by previous research (e.g., Hennessey & Amabile, 1988; Kandemirci, 2018). It is likely that story length reflects children's interest and intrinsic motivation to stay on task and produce a detailed story. Length may also influence the creativity assessment specifically among child samples, since children are likely to need more words to describe narratives with more detail, whereas adults can use fewer words to make a text creative by using complex linguistic features, such as metaphors or sarcasm.

4.2. DSI

The DSI score was a positive predictor of creative content in the stories. The DSI scores, which measured the ability to connect divergent ideas in a narrative, were calculated based on the relatedness between the words, based on pair-wise comparisons of the words, against a large English corpus. The predictive value of DSI on Creative Expressiveness was expected; to create an intriguing storyline, it is important to include different ideas in writing. The association between the Logic and DSI scores was negative which may indicate that young writers select closely related words, instead of semantically unrelated words, to create a logical narrative. A systematic review on creativity in narrative writing listed originality and surprise among important characteristics (D'Souza, 2021). Inclusion of different and surprising ideas, which is captured by the DSI scores, might have impacted the judges' decision about the creative value of the story. Going forward, DSI may be a valuable tool when judging the creative value of a narrative given the automated nature of it and its ability to eliminate the dependency on human rating. However more research, including qualitative and mixed methods, is needed to understand the conscious vs unconscious processes in human judgements of creative writing, as well as whether the findings replicate in adult samples.

We also found a small, positive correlation between the DSI scores and the story length. The DSI is designed to capture the average amount of semantic space a text covers. In children's writing, it is likely that a longer text will capture more divergent ideas in semantic space due to the narrative chain which associates diverse concepts as it evolves. A previous study concluded that while the DSI can correlate with text length, it is not an artifact, but the DSI captures divergent semantic coverage independent of text length (Johnson et al., 2022).

The DSI scores were also negatively associated with the Logic scores. This negative association is interesting since it may indicate that DSI scores also include information on unexpected, even illogical narratives which are likely to be viewed as creative. Also, it could, especially in the case of children, be that the more divergent the concepts included in a story are, the more difficult it is to make the story coherent and straightforward.

4.3. Originality

The originality score, which was based on the frequency of the words in a story in comparison to all words in all other stories, positively predicted the judges' creative expressiveness scores. That is, the use of more unusual words across the sample was associated with judgement of higher creative expression. In comparison to previous studies (e.g., Kandemirci, 2018), a benefit of the current study was that the same set of stimuli was used as a basis for all stories (i.e., every participant was given the same sequence of pictures on which to build their stories). This provided an opportunity to identify the originality of words in relation to all the words used by the participants. The less a word was used by all participants, the higher the originality score of the word; and the more rarely used words participants included in their stories, the higher their overall originality scores were. These findings suggest that human judges pick up on more unusual vocabulary use — which may indicate that the child is including concepts beyond the specifics of the pictures to create their story, or simply has a more diverse lexicon. Originality was not a predictor of Logic scores which indicates that a logical narrative does not require words that are not related to the stimuli presented to the children as a prompt for the task.

4.4. Grammar

Grammar scores predicted Creative Expressiveness scores which is in line with previous research, suggesting that clarity and the absence of distracting spelling mistakes are seen as characteristics of creative writing (D'Souza, 2021). However, in relation to creative writing generally, aspects of narratives such as abiding by grammatical rules or having correct spelling and punctuation may not be attributed to the creativity of the narrative (Mozaffari, 2013).

As predicted, grammatical accuracy of the stories (e.g., following grammatical rules and correct spelling of words) was also a significant predictor of the human-rated Logic scores. A study found that grammatical errors and spelling mistakes were the predominant features of elementary and secondary students' written stories that teachers were most likely to correct (Lucero et al., 2018). Additionally, students' success in grammatical aspects and spelling were the best predictors of their overall grades. In the same vein, it is reasonable that the judges might have rated the stories that followed the grammatical rules and included words that were spelled correctly as more logical.

4.5. Lexical Diversity

Lexical Diversity, controlled for the story length, was a predictor of Creative Expressiveness but in the negative direction. This was a surprising finding since the bivariate correlation between Lexical Diversity and Creative Expressiveness was positive. The change in the direction of the effect could be due to the random variation due to small amount of shared variance between the measures. Alternatively, the change of direction could be due to a suppressor effect. In addition, the bivariate correlation between Uncontrolled Lexical Diversity and Creative Expressiveness scores was negative. Interestingly, a similar pattern of associations was observed with Logic scores: the association was positive with Controlled Lexical Diversity and negative with Uncontrolled Lexical Diversity.

In addition, the Cognitive Ability explained a small amount of variance in Logic scores but not in Creative Expressiveness. The cognitive ability measures tapped into non-verbal reasoning skill which is likely to support the ability to put together descriptive and logical narratives. Age and Sex did not explain variance within either Creative Expressiveness or Logic.

4.6. Implications of the findings

The findings of this study have practical implications, as well as implications for conceptualisation of creativity. Based on our results, children could be guided to use a diverse language, incorporate remote concepts, and use words other than the presented prompts when engaged with creative storytelling. The identification of these specific linguistic predictors might be easier to apply in teaching practice, as opposed to a generic and non-actionable recommendation to encourage their pupils to be ‘more creative’.

As with production of any creative outcomes, not all outputs will be creative. To increase also the creative content in written, as a starting point, students should be encouraged to write more (Cutler & Graham, 2008). This is not a straightforward task. Despite the difficulties to engage children in writing, several educational interventions are already in use in primary education. The Writing Workshop method (Calkins, 1986) provides a structured yet flexible framework for writing instruction for primary school children. This method has been shown to be an effective instructional method to support young students in their writing by choosing a topic, revising and editing drafts, and sharing their work (Jasmine & Weiner, 2007). This approach, which not only focuses on content production in writing, might also help to increase the self-efficacy of young writers.

Educational tools have also been developed to support creative story telling. Research exploring the linguistic characteristics of creative content in children's writing can inform the development of tools aimed at enhancing creativity in writing. Tales Toolkit is an example of an educational method which supports early years storytelling skills (Jones Bartoli, 2018). Tales Toolkit can be used to narrate stories in a playful context. It enables children to build connections between characters, setting, problems and solutions in a creative manner. Technological advances have enable also digital storytelling, which combines traditional storytelling techniques with digital tools, allowing students to create and share multimedia stories. This approach has shown to improve creative writing and social-emotional skills, while at the same time, engaging also with digital skills (Uslu & Uslu, 2021). In addition, various drama-based interventions are used in the development of creative writing skills (Cremin et al., 2006).

4.7. Limitations and future directions

The present study used eleven linguistic predictors to assess their contribution to human-rated creative content in childhood writing. Future studies should include a wider range of methods to understand with greater detail the linguistic influences which contribute to explaining the subjective creativity assessment of childhood storytelling.

The current study also provided an opportunity to evaluate the practicality and usefulness of automated creativity measures in relation to the time and human labour requirements when assessing creative content in texts. With the use of CLAN, editing the stories to the CHAT format in preparation for the CLAN analyses was found to be relatively time consuming. In addition, certain aspects of the analyses done using CLAN, such as the Grammar measurement still required human input. However, reformatting the stories to be analysed in CLAN was still less laborious and involved reduced human-input compared to CAT. While the DSI has a more automated style when compared to CLAN, the tool is currently incapable of differentiating nonsensical narratives from coherent narratives (Johnson et al., 2022). As a result, nonsensical narratives with no cohesion or a plot might receive high scores on DSI. Similarly to the DSI, the use of Coh-Metrix tool, which was used to produce the Referential Cohesion scores, was easy to use. Due to the easy application, future research could incorporate other linguistic measures, generated by Coh-Metrix, to explore their relationships with human-rated creativity in texts. We acknowledge that this study provides a starting point and that there might be a variety of other automated tools to further improve the results achieved with the usage of tools utilised in the present study.

When transcribing the stories, words that were spelled incorrectly were not eliminated. This approach allowed us to have the best possible representation of children's creations including typos as well as the words they invented that did not exist in dictionaries. While the typos were captured by the grammar evaluation, this approach also brought about the caveat of some typos appearing as original words and inflating the number of original words. To have a consistent approach towards not editing or interpreting the intentions behind children's writing, we adhered to children's original writing.

Automated tools provide valuable supplementary data that can help explore factors influencing human-rated creativity assessments of products. However, more research is needed before these automated analyses can effectively replace human evaluations of creativity. Here, we aimed to present possible overlaps between computational ratings and human judgements to pave the way for a more automated creativity measurement. However, most of the conclusions drawn from the similarities between computational and human ratings of verbal creativity remain speculative, and our correlational analyses are unable to unpick causality in these associations. To have a more concrete understanding of the motivations behind human judges' creativity scores, future research could involve interviews with the judges where they are asked to justify the reasons behind their subjective scores. To the extent that these interviews back up our speculations, we may make the first steps on the path to maximising educational resources.

Educational statement

The study provides practical insights for educators on how to measure and support creativity in children's storytelling. The results emphasize the role of length, grammar, originality, and lexical and semantic diversity in creative childhood narratives. However, the study acknowledges the complexity of creativity assessment and the need for further research to develop more sophisticated tools for measuring creativity in childhood storytelling.

CRediT authorship contribution statement

Birsu Kandemirci: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roger E. Beaty:** Writing – review & editing, Methodology, Formal analysis. **Dan Johnson:** Writing – review & editing, Software, Methodology, Formal analysis. **Bonamy R. Oliver:** Writing – review & editing, Data curation. **Yulia Kovas:** Writing – review & editing, Data curation. **Teemu Toivainen:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgement

We would like to thank Andrew McMillan for his help with the data access; Hannah Hopkins, Addie Esther Subamani, Vivien Hwee Ting Loi, Jacob D Malamatenios and Julia Wick for transcription and coding of the stories on CLAN. R.E.B. is supported by grants from the US National Science Foundation grant (DRL-1920653, DUE-2155070). TEDS is supported by a programme grant (MR/V012878/1) to Professor Thalia Eley from the UK Medical Research Council (previously MR/M021475/1 awarded to Professor Robert Plomin), with additional support from the US National Institutes of Health (AG046938). We gratefully acknowledge the ongoing contribution of the participants in the Twins Early Development Study (TEDS) and their families.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2025.102626.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45*(2), 357.

Amabile, T. M. (2019). *Creativity in context: Update to the social psychology of creativity*. Routledge. https://doi.org/10.4324/9780429501234

Baer, J. (2016). Implications of domain specificity for creativity assessment. In *Domain specificity of creativity* (pp. 103–140). Elsevier. https://doi.org/10.1016/B978-0-12-799962-3.00005-7.

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). IGI Global. https://doi.org/10.4018/978-1-60566-667-9.ch004.

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal, 34*(3), 245–260. https://doi.org/10.1080/10400419.2022.2025720

Calkins, L. M. (1986). *The art of teaching writing*. Heinemann Educational Books Inc.

Caroff, X., Massu, J., Bourgeois-Bougrine, S., Caroff, X., Guegan, J., Mouchiroud, C., … Zenasni, F. (2022). The black box of the consensual assessment technique: Some questions and doubts on the subjective rating of creativity. In T. Lubart, & M. Botella (Eds.), *Homo Creativus* (pp. 193–217). Springer International Publishing. https://doi.org/10.1007/978-3-030-99674-1_11.

Catling, S., & Willy, T. (2010). Teaching primary geography. In *Learning matters*.

Cavanaugh, D. M., Clemence, K. J., Teale, M. M., Rule, A. C., & Montgomery, S. E. (2017). Kindergarten scores, storytelling, executive function, and motivation improved through literacy-rich guided play. *Early Childhood Education Journal, 45*(6), 831–843. https://doi.org/10.1007/s10643-016-0832-8

Cooper, H. (2018). What is creativity in history? *Education 3-13, 46*(6), 636–647. https://doi.org/10.1080/03004279.2018.1483799

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Cremin, T., Goouch, K., Blakemore, L., Goff, E., & Macdonald, R. (2006). Connecting drama and writing: Seizing the moment to write. *Research in Drama Education: The Journal of Applied Theatre and Performance, 11*(3), 273–291. https://doi.org/10.1080/13569780600900636

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Vol. 6738. Artificial intelligence in education* (pp. 438–440). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_62.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology, 100*(4), 907–919. https://doi.org/10.1037/a0012656

Department for Education. (2014). *National curriculum in England: Key stages 1 and 2 framework document*. Department for Education. https://www.gov.uk/government/publications/national-curriculum-in-england-primary-curriculum.

Devlin, J., Change, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805.

D'Souza, R. (2021). What characterises creativity in narrative writing, and how do we assess it? Research findings from a systematic literature search. *Thinking Skills and Creativity, 42*, 100949. https://doi.org/10.1016/j.tsc.2021.100949

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Blackwell.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity, 23*, 129–139. https://doi.org/10.1016/j.tsc.2016.12.005

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210–229.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Graham, S., Kiuhara, S. A., Harris, K. R., & Fishman, E. J. (2017). The relationship among strategic writing behavior, writing motivation, and writing performance with young, developing writers. *The Elementary School Journal, 118*(1), 82–104. https://doi.org/10.1086/693009

Hall, C., & Thomson, P. (2017). *Inspiring school change: Transforming education through the creative arts*. Routledge. https://doi.org/10.4324/9781315691084

Han, K.-S., & Marvin, C. (2002). Multiple creativities? Investigating domain-specificity of creativity in young children. *The Gifted Child Quarterly, 46*(2), 98–109. https://doi.org/10.1177/001698620204600203

Hennessey, B. A., & Amabile, T. M. (1988). Story-telling: A method for assessing children's creativity. *The Journal of Creative Behavior, 22*(4), 235–246.

Ivcevic, Z. (2022). Conceptual and measurement specificity are key: The case of creativity and emotions. *Creativity Research Journal, 34*(4), 391–400.

Jasmine, J., & Weiner, W. (2007). The effects of writing workshop on abilities of first grade students to become confident and independent writers. *Early Childhood Education Journal, 35*(2), 131–139. https://doi.org/10.1007/s10643-007-0186-3

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *Presented at the Annual Meetings of the Association for Computational Linguistics*. Florence, Italy.

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., … Beaty, R. E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research and Methods*. https://doi.org/10.3758/s13428-022-01986-2

Jones Bartoli, A. (2018). *Using storytelling to promote literacy, communication and socio-emotional development in the early years [report]*. Tales Toolkit. https://research.gold.ac.uk/id/eprint/24937/.

Kandemirci, B. (2018). *The effects of technology and peer collaboration on children's creativity*.

Kaplan, E., Fein, D., Kramer, J., Delis, D. C., & Morris, R. (1999). *Wechsler Intelligence Scale for Children—Process instrument*. San Antonio, TX: Psychological Corp.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*(2), 96–106. https://doi.org/10.1016/j.tsc.2007.04.002

Linebarger, D. L., & Piotrowski, J. T. (2009). TV as storyteller: How exposure to television narratives impacts at-risk preschoolers' story knowledge and narrative skills. *British Journal of Developmental Psychology, 27*(1), 47–69. https://doi.org/10.1348/026151008X400445

Lubart, T. (2017). The 7 C's of creativity. *The Journal of Creative Behavior, 51*(4), 293–296. https://doi.org/10.1002/jocb.190

Lubart, T., Pacteau, C., Jacquet, A.-Y., & Caroff, X. (2010). Children's creative potential: An empirical study of measurement issues. *Learning and Individual Differences, 20*(4), 388–392. https://doi.org/10.1016/j.lindif.2010.02.006

Lucero, M., Fernández, M. J., & Montanero, M. (2018). Teachers' written feedback comments on narrative texts in Elementary and Secondary Education. *Studies in Educational Evaluation, 59*, 158–167. https://doi.org/10.1016/j.stueduc.2018.07.002

Mackenzie, N. M., Scull, J., & Munsie, L. (2020). Analysing writing: The development of a tool for use in the early years of schooling. *Issues in Educational Research, 23*(3), 375–393. https://doi.org/10.3316/ielapa.353067356006239

MacWhinney, B. (2018). *CLAN manual*. https://doi.org/10.21415/T5G10R

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text — Interdisciplinary Journal for the Study of Discourse, 2*(1–3), 113–140. https://doi.org/10.1515/text.1.1982.2.1-3.113

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Mohamed, A., & Maker, C. J. (2011). Creative storytelling: Evaluating problem solving in children's invented stories. *Gifted Education International, 27*(3), 327–348. https://doi.org/10.1177/026142941102700309

Mozaffari, H. (2013). An analytical rubric for assessing creativity in creative writing. *Theory and Practice in Language Studies, 3*(12), 2214–2219. https://doi.org/10.4304/tpls.3.12.2214-2219

Nelson, K. (2010). *Young minds in social worlds: Experience, meaning, and memory*. Harvard University Press.

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS One, 9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist, 39*(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1

Puryear, J. S., & Lamb, K. N. (2020). Defining creativity: How far have we come since Plucker, Beghetto, and Dow? *Creativity Research Journal, 32*(3), 206–214. https://doi.org/10.1080/10400419.2020.1821552

Reay, D., & Wiliam, D. (1999). 'I'll be a nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343–354. https://doi.org/10.1080/0141192990250305

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan, 42*(7), 305–310.

Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., … Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioral and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics, 22*(6), 508–513. https://doi.org/10.1017/thg.2019.56

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 139–146. https://doi.org/10.1037/1931-3896.2.3.139

Smith, P., Fernandes, C., & Strand, S. (2001). *Cognitive abilities test 3 (CAT3)*. Windsor, England: nferNelson.

Spagnola, M., & Fiese, B. H. (2007). Family routines and rituals: A context for development in the lives of young children. *Infants & Young Children, 20*(4), 284. https://doi.org/10.1097/01.IYC.0000290352.32170.5a

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence, 34*(4), 363–374.

Storme, M., & Lubart, T. (2012). Conceptions of creativity and relations with judges' intelligence and personality. *The Journal of Creative Behavior, 46*(2), 138–149. https://doi.org/10.1002/jocb.10

Toivainen, T., Madrid-Valero, J. J., Chapman, R., McMillan, A., Oliver, B. R., & Kovas, Y. (2021). Creative expressiveness in childhood writing predicts educational achievement beyond motivation and intelligence: A longitudinal, genetically informed study. *British Journal of Educational Psychology, 91*(4), 1395–1413. https://doi.org/10.1111/bjep.12423

Toivainen, T., Malanchini, M., Oliver, B. R., & Kovas, Y. (2017). Creative storytelling in childhood is related to exam performance at age 16. *The European Proceedings of Social & Behavioural Sciences EPSBS, 33*(40), 375–384.

Uslu, A., & Uslu, N. A. (2021). Improving primary school students' creative writing and social-emotional learning skills through collaborative digital storytelling. *Acta Educationis Generalis, 11*(2), 1–18. https://doi.org/10.2478/atd-2021-0009

Webb, R., & Vulliamy, G. (2006). The impact of new labour's education policy on teachers and teaching at key stage 2. *Forum, 48*(2), 145. https://doi.org/10.2304/forum.2006.48.2.145

Wyse, D. (2001). Grammar for writing? A critical review of empirical evidence. *British Journal of Educational Studies, 49*(4), 411–427. https://doi.org/10.1111/1467-8527.t01-1-00185

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods, 51*(2), 879–894. https://doi.org/10.3758/s13428-018-1137-1