frontiers | Frontiers in Education

TYPE Systematic Review
PUBLISHED 12 October 2023
DOI 10.3389/feduc.2023.1240962

OPEN ACCESS

EDITED BY
Mohammad Khalil,
University of Bergen, Norway

REVIEWED BY
Chen-Yao Kao,
National University of Tainan, Taiwan
Faisal Saeed,
Kyungpook National University,
Republic of Korea

*CORRESPONDENCE
Manoli Pifarré
✉ manoli.pifarre@udl.cat

RECEIVED 15 June 2023
ACCEPTED 18 September 2023
PUBLISHED 12 October 2023

CITATION
Ul Haq I and Pifarré M (2023) Dynamics of automatized measures of creativity: mapping the landscape to quantify creative ideation. *Front. Educ.* 8:1240962.
doi: 10.3389/feduc.2023.1240962

Dynamics of automatized measures of creativity: mapping the landscape to quantify creative ideation

Ijaz Ul Haq and Manoli Pifarré*

Faculty of Education, Psychology and Social Work, University of Lleida, Lleida, Spain

The growing body of creativity research involves Artificial Intelligence (AI) and Machine learning (ML) approaches to automatically evaluating creative solutions. However, numerous challenges persist in evaluating the creativity dimensions and the methodologies employed for automatic evaluation. This paper contributes to this research gap with a scoping review that maps the Natural Language Processing (NLP) approaches to computations of different creativity dimensions. The review has two research objectives to cover the scope of automatic creativity evaluation: to identify different computational approaches and techniques in creativity evaluation and, to analyze the automatic evaluation of different creativity dimensions. As a first result, the scoping review provides a categorization of the automatic creativity research in the reviewed papers into three NLP approaches, namely: text similarity, text classification, and text mining. This categorization and further compilation of computational techniques used in these NLP approaches help ameliorate their application scenarios, research gaps, research limitations, and alternative solutions. As a second result, the thorough analysis of the automatic evaluation of different creativity dimensions differentiated the evaluation of 25 different creativity dimensions. Attending similarities in definitions and computations, we characterized seven core creativity dimensions, namely: novelty, value, flexibility, elaboration, fluency, feasibility, and others related to playful aspects of creativity. We hope this scoping review could provide valuable insights for researchers from psychology, education, AI, and others to make evidence-based decisions when developing automated creativity evaluation.

KEYWORDS

review, creativity process, ideation, evaluation, artificial intelligence

1. Introduction

Creativity as a 21st century skill is increasingly becoming an explicit part of educational policy initiatives and curricula (Plucker et al., 2023). Creativity is a multifaceted concept, and research in this area has made remarkable progress in understanding the different components embedded in creativity phenomena, such as idea generation through collaborative creative (co-creative) processes (Sawyer, 2011, 2022). Furthermore, research also revealed the significance of another important component of creativity: creativity evaluation (Guo et al., 2023), which is the ability to accurately identify creative ideas, solutions, or characteristics among individuals to understand their creative strengths and potential (Kim et al., 2019). In the educational context, creativity evaluation is an essential step for teachers and students because it is helpful to monitor, refine, and implement creative ideas, which could improve students' creative performance in the creative process (Rominger et al., 2022).

Creativity evaluation poses a challenging problem in creativity research. Creativity evaluation mainly involves four dimensions: fluency (number of meaningful ideas), flexibility (number of different categories), elaboration (detailed ideas), and novelty (uniqueness of ideas) (Bozkurt Altan and Tan, 2021). To evaluate these creativity dimensions, various manual creativity evaluations (paper-based) and psychological tests have been commonly used (Rafner et al., 2022). Examples are the Torrance Tests of Creative Thinking (Torrance, 2008), Creativity Assessment Packet (CAP) (Williams, 1980), and Divergent Production abilities (DP), (Guilford, 1967). Other ways to evaluate creativity include a rating scale (Gong and Zhang, 2017; Birkey and Hausserman, 2019), a survey and questionnaire (De Stobbeleir et al., 2011; Gong et al., 2019), using a grading rubric (Vo and Asojo, 2018), and subjective scoring of creativity dimensions (George and Wiley, 2020). However, these manual creativity evaluations face some challenges, e.g., being error-prone (experts' ratings do not always agree on what is creative) and time-consuming (Said-Metwaly et al., 2017; Doboli et al., 2020). These challenges can be tackled using automated creativity evaluation supported by AI techniques which can also enrich co-creation by providing real-time feedback to guide students to develop novel solutions (George and Wiley, 2020; Kenworthy et al., 2023).

Artificial intelligence (AI) focuses on enabling machines to perform tasks that typically demand human intelligence. Within AI, machine learning (ML) algorithms learn from data to make predictions. Notably, computer vision is used for analyzing figural data, and NLP is used for analyzing textual data. Given our focus on textual ideas, NLP enables machines to comprehend, interpret, analyze, and generate human language (Braun et al., 2017). NLP contains a variety of approaches and techniques such as text similarity, text classification, topic modeling, information extraction, and text generation, each with its computational techniques spanning from statistical methods to predictive and deep learning models. NLP provides different opportunities to compute variables related to creativity dimensions. Among these, the following five variables could be computed in the vector space provided by NLP: (1) Contextual and semantic similarity are applied to measure the uniqueness of ideas and originality (Hass, 2017; Doboli et al., 2020); (2) text clustering could identify different categories in the text; (3) text classification is used to compute novelty (Simpson et al., 2019); (4) keyword searching is mainly used to compute elaboration (Dumas et al., 2021); and (5) information retrieval could be applied to score the level of idea elaboration (Vartanian et al., 2020). These implications of NLP in co-creative processes can be used to automatically evaluate creativity and support co-creation by providing feedback (Bae et al., 2020; Kang et al., 2021; Kovalkov et al., 2021).

Considering the above implications of NLP, current research focuses on studying how different computational techniques can measure creativity dimensions (Doboli et al., 2020). Research on this topic has been very productive and has designed other computational techniques to measure creativity dimensions, e.g., (1) novelty is measured by keyword similarity (Prasch et al., 2020), part of speech tagging (Karampiperis et al., 2014; Camburn et al., 2019), and different ML classifiers, such as Bayesian classifiers, random tree, and Support Vector Machine (SVM) (Manske and Hoppe, 2014; Simpson et al., 2019; Doboli et al., 2020); (2) originality dimension is measured by Latent Semantic Analysis (LSA) (Dunbar and Forster, 2009), Global Vectors for word representation (GloVe) (Dumas et al., 2021), and part of speech tagging (Georgiev and Casakin, 2019); (3) fluency dimension is measured by LSA (Dumas and Dunbar, 2014; LaVoie et al., 2020); (4) elaboration dimension is measured by parts of speech tagging (Dumas et al., 2021); and (5) level of details dimension is measured by text-mining methods (Camburn et al., 2019).

This study aims to tackle the following four main challenges that current research faces when designing computational techniques to measure creativity: (1) a range of computational techniques evaluating various creativity dimensions; (2) there is no consensus about the use of a specific technique for computing a specific creativity dimension; (3) some of the studies do not expose and argue the rationale that supports the use of a specific technique to compute a specific creativity dimension, e.g., evaluation of the category switch dimension of creativity using LSA (Dunbar and Forster, 2009); and (4) the need to consider the limitations of computational techniques that could affect the evaluation of creativity dimensions (Olivares-Rodríguez et al., 2017; Doboli et al., 2020). Considering these challenges, as per our knowledge, no existing literature review addresses the above four challenges. Therefore, this exploration led us to two research questions: (1) What NLP approaches and techniques are used to automatically measure creativity? and (2) What creativity dimensions are computed automatically, and how? These research questions enable us to address the previous four challenges in automatic creativity evaluation. Furthermore, these research questions help to understand the concept of NLP approaches and creativity dimensions, their applications in evaluating creativity dimensions, identify research gaps and limitations, and propose alternative solutions for advancing the evaluation and promotion of creativity. Therefore, we chose a scoping review because it helps to understand key concepts and identify knowledge gaps (Munn et al., 2018) to inspire innovation and improve the education of future generations through advanced technologies.

2. Research objectives

This scoping review aims to meet the following two objectives.

1. To identify and categorize different ML approaches used in automatic creativity evaluation, highlighting their application scenarios and limitations of computational approaches and techniques. This categorization could contribute to a deeper understanding of the contribution that different ML approaches can make to automatic creativity.
2. To analyze the definition and computation of different creativity dimensions used in automatic creativity evaluation research. This analysis can help establish a joint agreement on creativity dimensions and their computation, which will pave the way for advancements in automatic creativity evaluation.

3. Method

This section describes the sampling method we used to collect and compile the state-of-the-art approaches to automatic creativity evaluation. Our methodological framework follows the PRISMA technique (Dickson and Yeung, 2022) by conducting a scoping review to find relevant and significant research papers by identifying the following four core concepts.

1. Creativity: The articles must be related to creativity, especially the creative process (Sawyer, 2011).
2. Measurement/evaluation/assessment of creativity dimensions.
3. Technology: We selected those studies that are assisted or evaluated with technology support. This core concept aims to review the technological support for creativity evaluation and explore future research in the creative process.
4. Domain: We focused on the creativity process applicable in the educational sector that helps to enhance students' creativity. Other fields such as medicine, finance, and business were excluded from the search query.

Exploring the current literature considering the above four core concepts, peer-reviewed journals and conference papers are included in this mapping study. Regarding the time span, we searched from 2005 to 2021, although interestingly, according to our inclusion–exclusion criteria, the oldest study included is from 2009, and most are from recent past years. It indicates that automatic creativity evaluation has recently grabbed researchers' attention and is still an open and active research problem.

We excluded articles focused on the person's or organization's creativity evaluation. We excluded domains other than education, e.g., medicine and finance. Articles in other languages apart from English published before 2005 and articles with no technological role and creativity were also excluded.

For this mapping study, we extracted articles published in Scopus with the search query: [(creativ* OR "Creative Process" OR "Novelty" OR "Flexibility" OR "Fluency" OR "Elaboration" OR "Originality") AND (Measur*OR Evaluat* OR Asses* OR Calcul* OR Analys* OR Scor* OR Qunat*) AND (Automat* OR Comput* OR Machin* OR Natural* OR Artificial* OR Deep learning OR Mathemat* OR Mining) AND (E-learning OR educa* OR Learn* OR School OR students*)].

The search query resulted in 364 research articles. By applying the inclusion and exclusion criteria while reading the title, abstract, keywords, and conclusion, the search is filtered to 65 articles. Furthermore, the authors read, checked, and discussed the selected articles and conducted all the screening stages to answer the two research questions. The consensus among the authors developed by solving discrepancies since member checking is a well-established procedure to build up "trustworthiness" in qualitative research (Toma, 2011). After this process, a total of 26 articles were finally included in this scoping review. The overall article selection procedure through the PRISMA technique is depicted in Figure 1.

4. Results

4.1. Approaches and techniques used in automatic creativity evaluation (RQ1)

The compilation of computational approaches and techniques in automatic creativity evaluation research to answer the first research question gives the following three results;

The first result reveals that creativity evaluation research spreads over three different NLP approaches, namely, (1) text similarity, which measures the relatedness and closeness among words, sentences, or paragraphs presented in a numerical space; (2) text classification, which is a supervised learning approach (needs data training) that requires ML algorithms [such as the K-nearest neighbor (KNN) algorithm and random forest] to analyze text automatically and then to assign a set of predefined tags or categories; and (3) text mining that uses NLP to examine and transform extensive unstructured text data to discover new information and patterns. These three NLP approaches and their computational techniques identified in the studies included in this review are displayed in Figure 2.

As a second result, the scoping review shows that text similarity is the most common approach (69% of the reviewed studies), followed by text classification (27%), and text mining is less commonly used (only 4% of the studies), as shown in Figure 2.

As a third result, our scoping review has identified and categorized the computation techniques used in the three NLP approaches (text similarity, text classification, and text mining) and the creativity dimensions that were evaluated automatically. In the following sections, we present the mapping that we have built after a thorough analysis of all the studies included in the scoping review.

Regarding the text similarity approach, NLP converts textual ideas into a numerical vector space. To do this conversion, the studies revised the use of a wide range of techniques that could be classified into the next three categories: string-based similarity, corpus-based similarity, and knowledge-based similarity. These three categories and their computational techniques identified in the reviewed studies are shown in Figure 3, and Table 1 maps automatic creativity evaluation studies into the three categories and techniques used.

In the first category, string-based similarity (6% of the text similarity approach of reviewed studies) matches exact keywords or alphabet strings, e.g., Longest Common Substring (LCS) or N-gram (a subsequence of n items from a given sequence of text). The string similarity of ideas with the existing ideas in the database is computed by using keyword matching (Prasch et al., 2020).

In the second category, corpus-based similarity is mostly used (72% of textual similarity), and the results are presented in Table 1. The corpus-based similarity is classified into two sub-categories: On the one hand, the statistical-based models, e.g., LSA, present corpus in the word-document matrix as words in row vectors and each document as a column vector, and weighting schemes and dimension reduction schemes are applied before calculating the cosine similarity among word vectors (Martin and Berry, 2007; Wagire et al., 2020). On the other hand, the deep learning-based models (both word and sentence embeddings) use supervised (which need to be trained on data), semi-supervised,

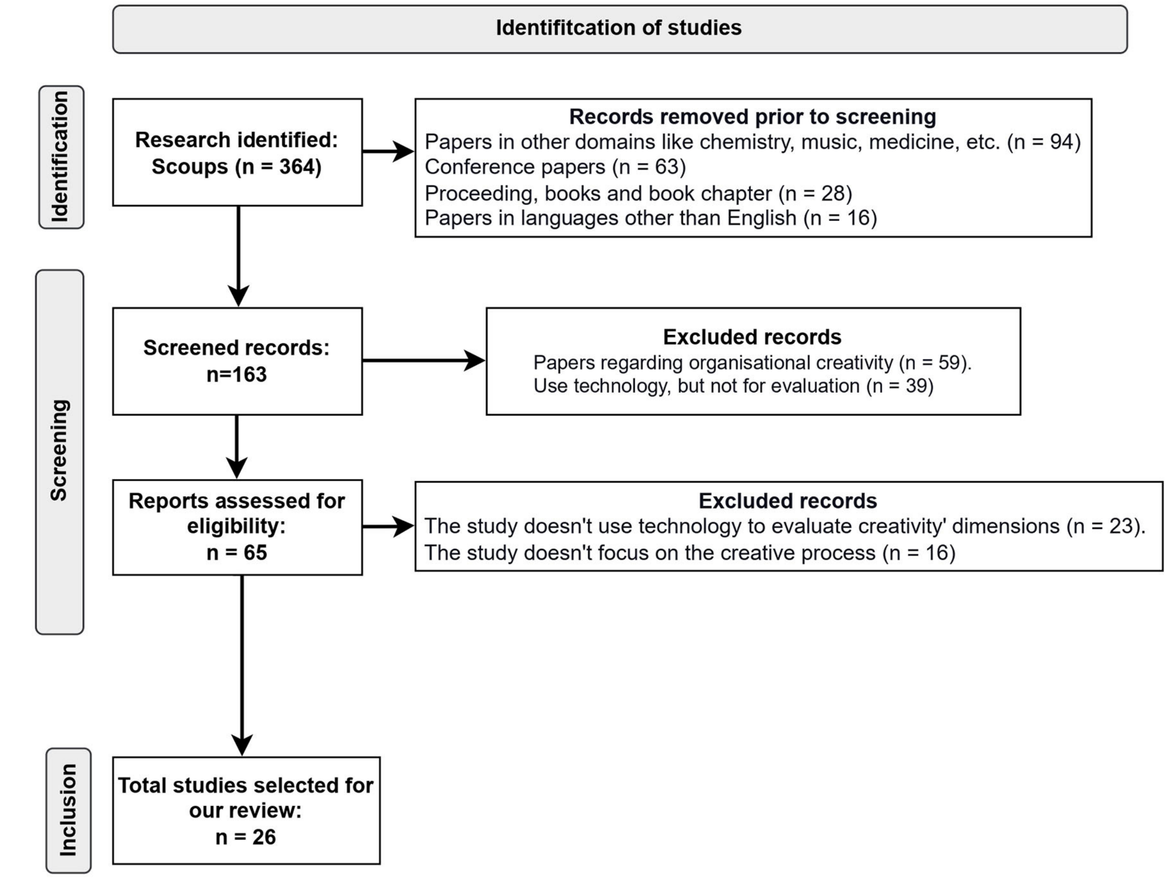
Identifitcation of studies
Identification
Screening
Inclusion
Research identified: Scoups (n = 364)
Records removed prior to screening
Papers in other domains like chemistry, music, medicine, etc. (n = 94)
Conference papers (n = 63)
Proceeding, books and book chapter (n = 28)
Papers in languages other than English (n = 16)
Screened records: n=163
Excluded records
Papers regarding organisational creativity (n = 59).
Use technology, but not for evaluation (n = 39)
Reports assessed for eligibility: n = 65
Excluded records
The study doesn't use technology to evaluate creativity' dimensions (n = 23).
The study doesn't focus on the creative process (n = 16)
Total studies selected for our review: n = 26

FIGURE 1
Screening procedure of the articles using the PRISMA technique.

or unsupervised methods (no prior training) that are trained on a large corpus, e.g., Wikipedia and common crawl dataset. Deep learning models such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) use knowledge from large datasets, encode the data, and find similarities in words or sentences. The GloVe model showed reliable results as compared with the experts' scores, especially for single-word creativity tasks (Beaty and Johnson, 2021; Johnson and Hass, 2022).

In the third category, knowledge-based similarity (used in 22% of text similarity approaches in reviewed studies, as presented in Table 1) using the knowledge of ontologies represents the textual data on a semantic network graph consisting of nodes representing semantic memory and lines. Ontologies are the dictionaries of millions of words and are lexically associated, e.g., WordNet, Wikipedia, and DBpedia.

Text classification is the second NLP approach used by 27% of the reviewed studies in automatic creativity evaluation depicted in Figure 1. Classification is an ML technique that categorizes text into predefined categories. The classification consists of four main steps: (1) data collection, pre-processing (data acquisition, cleaning, and labeling), and data presentation (feature selection, dividing into training and testing datasets); (2) applying classifier models; (3) evaluation of classifiers; and (4) prediction (output of the testing data). These four steps are influential factors when applying text classification in automatic creativity evaluation. Table 2 gives an overview of the classification approach, the datasets, classifiers, evaluations, and creativity dimensions in creativity evaluation research.

Text mining is the third approach in automatic creativity evaluation, which is the practice of analyzing a vast collection of textual data to capture key concepts, trends, patterns, and hidden relationships. In the scoping review, text mining is used (Dumas et al., 2021). The studies used four mining techniques, e.g., all words count, stop list inclusion (defined terms that are not meaningful), counting part of speech, and applying inverse document frequency (a technique to extract rare and important documents).

4.2. Creativity dimensions are computed automatically (RQ2)

In the studies included in this scoping review of automatic creativity evaluation, we differentiated 25 different creativity dimensions. These 25 dimensions of creativity are displayed in the second column (Manifestation) of Table 3. We analyze the similarities in the conceptual definition and computational approach employed in various studies that consider different

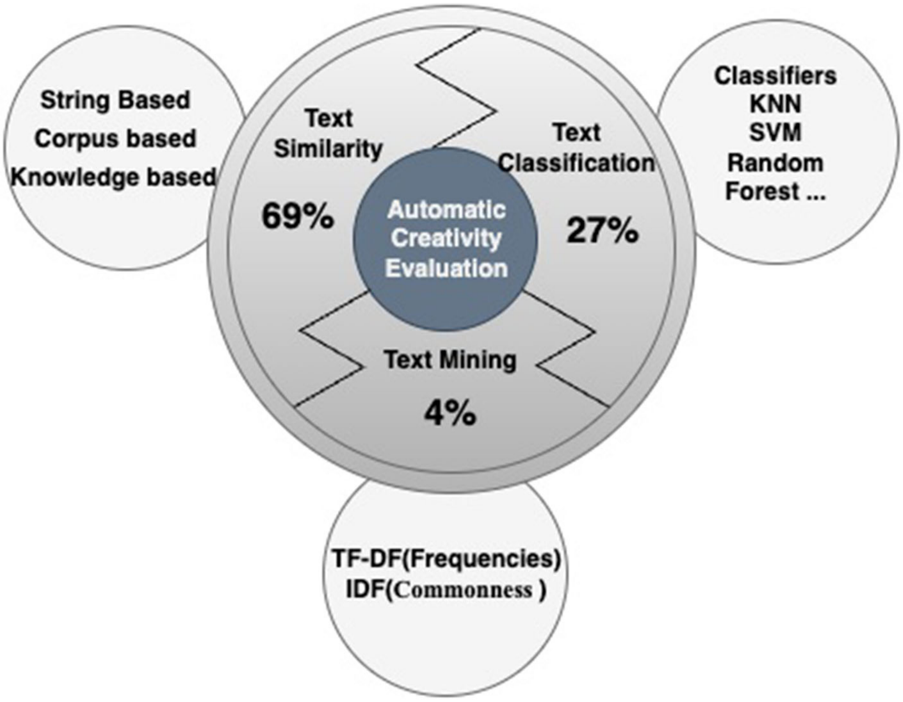

FIGURE 2
Different NLP approaches in creativity evaluation.

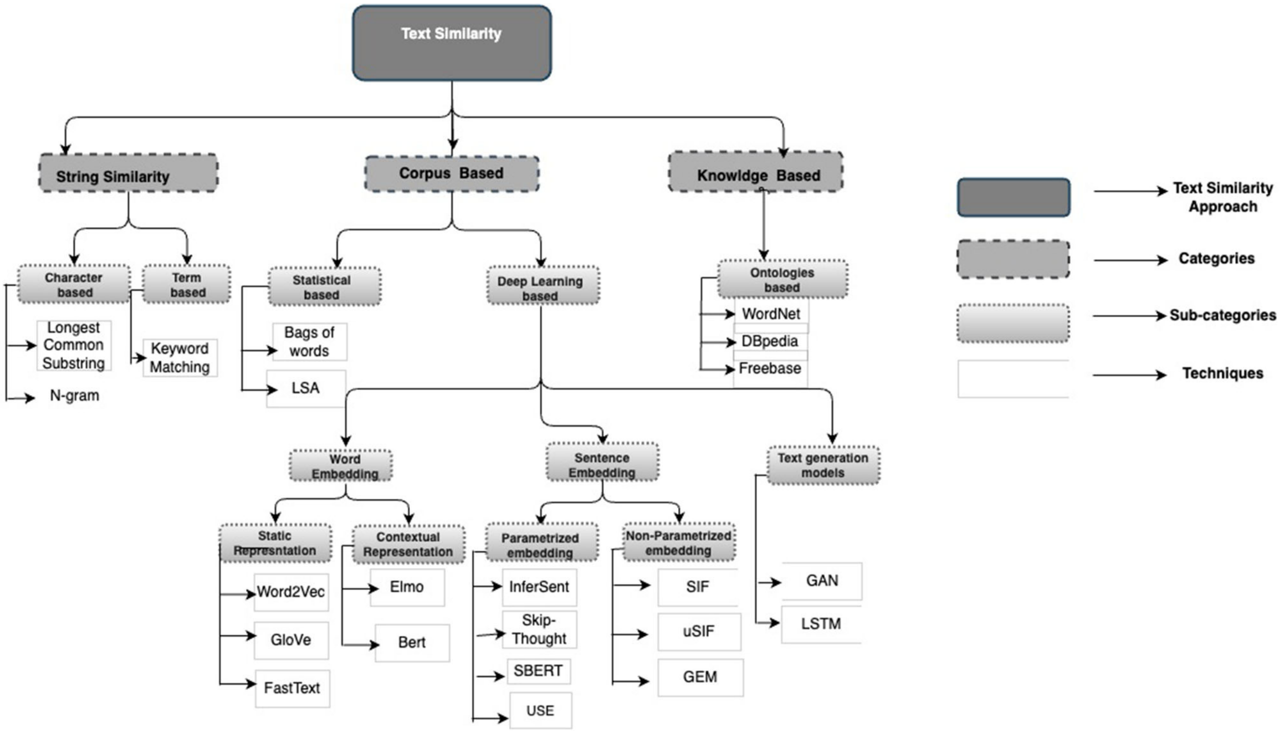

FIGURE 3
Text similarity approaches, categories, sub-categories, and their computational techniques.

dimensions for assessing creativity. This analysis allows us to categorize these 25 manifestations of creativity into seven core creativity dimensions, namely, novelty, value, flexibility, elaboration, fluency, feasibility, and others related to playful aspects of creativity such as humor or recreational efforts, which are displayed in the first column of Table 3 (Core Dimension).

TABLE 1 Categorizing of review studies in text similarity approaches and percentages of studies included in the review that use each approach.

Text similarity categories	Sub-categories	Vectorization techniques	Dimensions	Studies
String-based 6%		Keyword matching	Novelty, usefulness	Prasch et al., 2020
Knowledge-based 22%		Part of speech tagging	Novelty, level of details	Camburn et al., 2019
		Part of speech tagging	Originality, value, overall value, feasibility	Karampiperis et al., 2014
		Clustering in the knowledge graph	Novelty, surprise, rarity, recreational effort	Georgiev and Casakin, 2019
		Semantic network	Flexibility	Cosgrove et al., 2021
Corpus-based 72%	Statical based	LSA	Category switch, variety, original, prune originality, common use	Dunbar and Forster, 2009
		LSA	Fluency, originality	Dumas and Dunbar, 2014
		LSA	Similarity, fluency	LaVoie et al., 2020
		Vectorization of linguistic features	Similarity	Zuñiga et al., 2017
	Deep learning	Word2Vec	Originality, flexibility, fluency	Sung et al., 2022
		GloVe	Originality	Acar et al., 2021; Beaty and Johnson, 2021; Dumas et al., 2021
		GloVe	Similarity of text	Olson et al., 2021
		GloVe	Diversity (Novelty)	Johnson and Hass, 2022
		Universal sentence encoder	Novelty	Kenworthy et al., 2023
		GAN	Novelty, value, surprise	Franceschelli and Musolesi, 2022
		LSTM	originality	Marrone et al., 2022

TABLE 2 Text classification-based creativity evaluation studies.

Datasets	Classifiers	Evaluation	Dimensions	Studies
In 4,099,877 solutions from Project Euler Website	Linear regression and SVM	Comparison with expert rating	Novelty, usefulness, quality	Manske and Hoppe, 2014
Two datasets were used: 1. ideas: 1,480; 2, domain dataset: a collection of 1,144 sports datasets from Wikipedia	SVM, neural networks (NN), logistic regression, decision trees, KNN, and Naive Bayes	F-measure is a measure of a test's accuracy. Precision and recall are calculated	Novelty	Doboli et al., 2020
Semeval-2017 jokes, 4,030 short texts, and VU Amsterdam Metaphor Corpus	Bayesian approach	Bayesian approach is compared to the best-worst scaling method	Novelty, humor	Simpson et al., 2019
Internet movies database and Rotten Tomatoes dataset contained textual, image, and numerical attributes	SVM, random forest, ridge regression, Bayesian regression, and K-nearest regression	Correlation analysis	Novelty, value, influence, unexpectedness	Shrivastava et al., 2017
User queries Wikipedia as knowledge source	Random trees	Sensitivity, Specificity	Diversity	Olivares-Rodríguez et al., 2017
203 responses present in the multiplex lexical network	Logistic regression, random forest, and SVM classifiers	Entropy	Fluency	Stella and Kenett, 2019
1,214 recipes, 2,130 ingredients, and 235 cooking techniques	K-neighbor classifier, SVM, multi-layer perceptron classifier, and the random forest	The scoring function of classifiers, random forest, has the best results. No other evaluation	Novelty, adaptiveness, style, transcendence, realization	Jimenez-Mavillard and Suarez, 2022

Furthermore, the results obtained to answer research question two are illustrated in Figure 4, which displays the percentage of the seven core creativity dimensions identified in this review. These results show that novelty is the most evaluated dimension in the studies compiled in this scoping review.

TABLE 3 Characterization of 25 creativity dimensions into seven core creativity dimensions (first column) and creativity dimensions manifested (second column) based on similarities in definitions (third column) and computation (fourth column).

Core dimension	Dimension manifestation	Dimension definition	Dimension computation	Study
Novelty	Novelty	Novelty is an idea with respect to prior ideas or deviation from existing solutions	Textual similarity of a given solution to all existing or previous solutions	Manske and Hoppe, 2014; Prasch et al., 2020; Kenworthy et al., 2023
		A measure of how unique a concept is relative to others	Span (Path length) is the sum of distances of each entity or unique words from the central entity or topic (e.g., predefined hierarchical topical categories of Wikipedia).	Camburn et al., 2019
		The deviation from existing knowledge/experience	Average semantic distance between the dominant terms included in the textual representation of the story, compared to the average semantic distance of the dominant terms in all stories	Karampiperis et al., 2014
		-	Pairwise text similarity using linguistic features	Simpson et al., 2019
		Novelty is defined as a unique solution	From surprise and relevance score surprise term is computed from document term frequency in idea data, and relevance term is calculated from domain dataset (sports was collected from Wikipedia)	Doboli et al., 2020
		How an artifact is different from others	Calculation of the distance between a given artifact and the other artifacts in a descriptive space	Shrivastava et al., 2017; Franceschelli and Musolesi, 2022
		Novelty to originality score and defines that the creative method led to more innovative products	The classifier models learn from ingredients and techniques and classify them as novel or not novel in the case study of culinary products	Jimenez-Mavillard and Suarez, 2022
	Originality	Similarity to existing ideas	Semantic distance between the responses	Dunbar and Forster, 2009; Song et al., 2020; Beaty and Johnson, 2021
		Originality is referred as a novelty	The semantic distance among ideas	Dumas and Dunbar, 2014
		Statistically infrequent responses	Semantic distance between the responses	Acar et al., 2021
		A response that is more unusual within a given context would be more Original.	Semantic distance between a given responses	Dumas et al., 2021
	Similarity	The similarity of meaning between multiple texts	The similarity of the new response was measured with topic clusters, rubrics, and example responses	LaVoie et al., 2020
		The similarity of the original poem to the translated poem	Similarity distance is calculated between original (English language) and translated poems (Spanish)	Zuñiga et al., 2017
		Similar contexts have smaller distances	Semantic distance between different words	Olson et al., 2021
	Diversity	Semantic distance among user queries	Semantic similarity is estimated of each user-issued query to the k most relevant concepts for the challenge using distance formulas	Olivares-Rodríguez et al., 2017
		The degree to which participants engaged in semantic context search	Semantic diversity refers to the degree to which the contexts surrounding words vary in their meanings	Johnson and Hass, 2022

(Continued)

TABLE 3 (Continued)

Core dimension	Dimension manifestation	Dimension definition	Dimension computation	Study
	Rarity	A rare combination of properties	The sum of weights on the min-weight closure of the cluster graph is compared to the maximum sum of weights in the story	Karampiperis et al., 2014
	Common use	Common uses of objects in Object use tasks	Each response was compared to the most common use of the corresponding object (collected previously from Common Use Judges)	Dunbar and Forster, 2009
	Surprise or unexpectedness	Unexpectedness or surprise defines how different the artifact or some of its attributes are from expected behavior	The similarity of a given artifact with other artifacts	Shrivastava et al., 2017; Franceschelli and Musolesi, 2022
	Influence	How impactful or inspiring it has been	The similarity of an artifact with other artifacts occurs later	Shrivastava et al., 2017
Value	Value	A measure of how artifact is valued by domain experts for artifact	Datapoint is highly valuable if its combination of correlated dimensions leads to a better rating prediction.	Shrivastava et al., 2017; Franceschelli and Musolesi, 2022
	Overall value	Overall value of the outcome of the designs from design ideation	Semantic analysis of verbalizations can be promising to measure the semantic value.	Georgiev and Casakin, 2019
	Quality	Quality is related to reliability, maintainability, extensibility, and adaptability.	Quality and Usefulness are computed from two metrics. Static Code Metrics: Line of codes Dynamic Code Metrics: number of visited lines	Manske and Hoppe, 2014
	Usefulness	The correct solutions to programming tasks		Manske and Hoppe, 2014
	Adaptiveness and Style	Adaptiveness is the solution to solve a problem. Style is elegance and other aesthetic qualities	Adaptiveness as useful solutions style as quality	Jimenez-Mavillard and Suarez, 2022
Elaboration	Elaboration	The degree to which they explain and embellish their responses	Counting based on: (1) Unweighted Word, (2) Stop listed Inclusion, (3) Part of Speech Inclusion, and (4) Inverse Frequency Weighting.	Dumas et al., 2021
	Level of details	Level of details of the ideas	Count of named entities. Examples of entities are person, place, things, money, etc	Camburn et al., 2019
Flexibility	Flexibility	Semantic memory structure	1. Cosine similarity to estimate the edges between nodes in semantic network. 2. Number of similar clusters	Sung et al., 2022
	Category switch	Number of changes in the category of use between responses	The similarity scores between successive response pairs were averaged for each object	Dunbar and Forster, 2009
	Variety	Measure the variety of responses produced by each person	The similarity scores between every single pair of responses for an object were also averaged as a measure of the variety of responses produced by each person	Dunbar and Forster, 2009
Fluency	Fluency	Number of ideas	Counting the number of ideas	Dumas and Dunbar, 2014; Stella and Kenett, 2019; Sung et al., 2022
Feasibility	Feasibility	Feasibility can be materialized or achieved in real practice	Polysemy, abstraction, and IC are highly correlated to the feasibility score	Georgiev and Casakin, 2019
	Transcendence and realism	Transforming into reality	Development of the product and its communication with the other products	Jimenez-Mavillard and Suarez, 2022

(Continued)

TABLE 3 (Continued)

Core dimension	Dimension manifestation	Dimension definition	Dimension computation	Study
Other	Humor	Humor is funniness	Pairwise comparison of text	Karampiperis et al., 2014
	Recreational effort	Difficult to achieve	The number of different clusters that each story contains as compared to the maximum number of clusters in a story of the whole group	Simpson et al., 2019

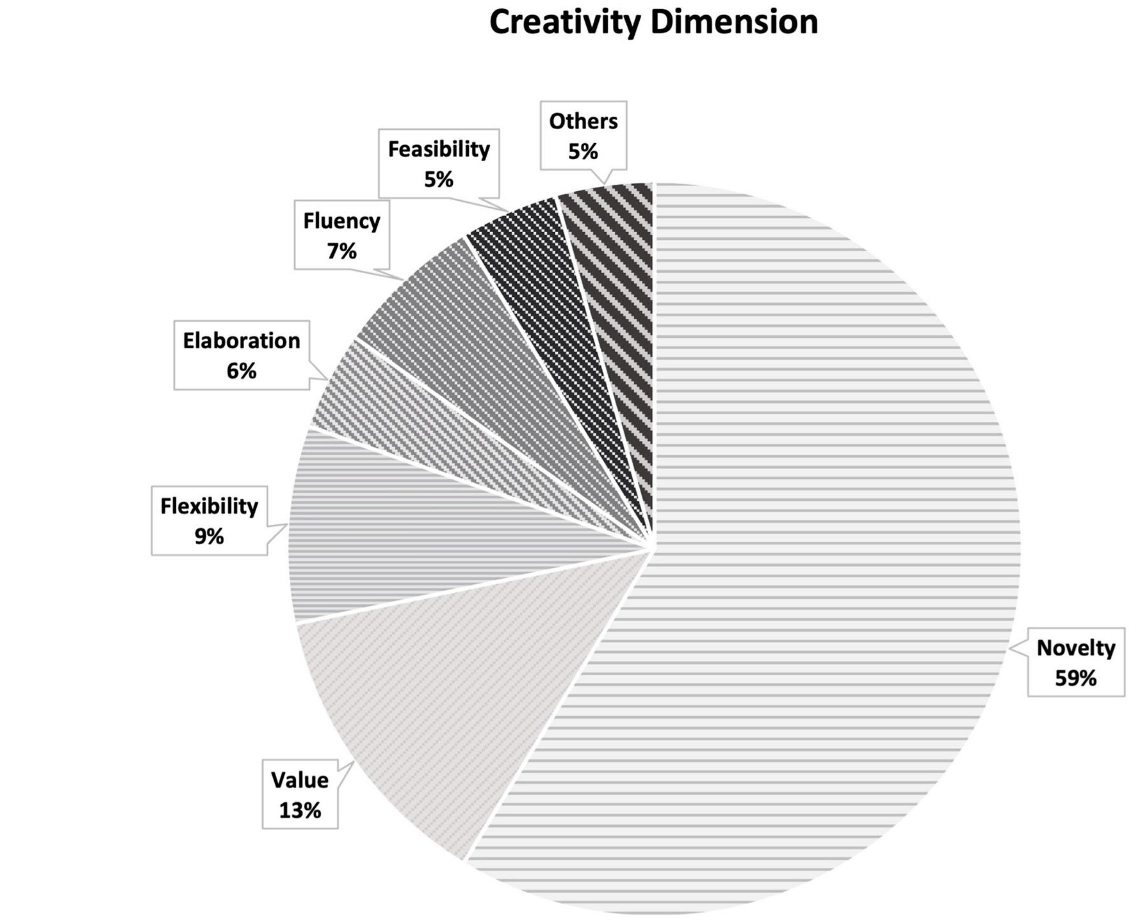

FIGURE 4
Percentage distribution of each core creativity dimension in the reviewed studies.

5. Discussion

5.1. Approaches and techniques used in automatic creativity evaluation

The scoping review identified three main NLP approaches used in automatic creativity evaluation, namely, (1) text similarity, (2) text classification, and (3) text mining. In the next sections, we discuss the contribution of each computational approach to automatic creativity evaluation, argue their applications, discuss their limitations, identify research gaps, and make further recommendations for automatic creativity evaluations.

Regarding the text similarity approach, the scoping review revealed that it is used in 69% of the studies, which helps understand creative thinking (Li et al., 2023). Our analysis concluded that the widespread use of textual similarity in automatic creativity evaluation is because automatic creativity evaluation is more focused on evaluating originality, novelty, similarity, or diversity dimensions of creativity. The computations of these dimensions involve assessing the similarity of an idea with the existing ideas. The text similarity approach provides a variety of computational techniques to measure the similarity of ideas, as shown in Figure 3.

Concerning the three categories of text similarity, namely, string similarity, corpus-based similarity, and knowledge-based similarity as set out in Table 3, the scoping review shows differences in the process of similarity computation that have an impact on how they are applied. On the one hand, string-based and knowledge-based similarities have limited application in automatic creativity evaluation because string-based only considers syntactic similarity (not semantic) and knowledge-based only extracts from text-specific entities, such as a person's name, place, and money (Camburn et al., 2019). During ideation, the knowledge-based approach might focus on entities rather than technical terms or scientific jargon within the sentence used by sentences solving a scientific challenge. For example, when brainstorming about renewable energy solutions, the knowledge-based approach might not capture specific terms such as "photovoltaics" or "wind turbines." On the other hand, corpus-based techniques are widely used, so in the following, we elaborate on corpus-based techniques.

Regarding corpus-based similarity, it has been commonly used in automatic evaluation because it provides a wide range of computational techniques, from simple statistical to deep learning models, as shown in Figure 2. Considering that a statistical model such as LSA is applied to examine semantic similarity, memory, and creativity (Beaty and Johnson, 2021), it has shown a more reliable scoring technique of originality on divergent thinking tasks than human ratters (Dunbar and Forster, 2009; Dumas and Dunbar, 2014; LaVoie et al., 2020), as shown in Table 1. We argue that LSA uses statistical techniques, including Probabilistic Latent Semantic Analysis (Hofmann, 1999), Latent Dirichlet Allocation (Blei et al., 2003), and Non-Negative Matrix Factorization (Lee and Seung, 1999), which limit its implication because these consider words statistics (e.g., co-occurrence of words) instead of word contextual and semantic meaning. These limitations are addressed by deep learning models, which we discuss below.

Recently, drastic changes in NLP research with the development of deep learning models based on deep neural architectures have unlocked ways to model text with more nuance and complexity. This advancement started with the development of word embedding models such as GloVe or Word2Vec pre-trained, including Wikipedia, news articles, and web pages. These predictive models use a neural network with one or more hidden layers to learn the vector representations of words. The GloVe showed comparable results to human experts' scores in single-word creativity tasks (Beaty and Johnson, 2021; Olson et al., 2021). However, word embedding models do not differentiate between a list of keywords and a meaningful sentence; hence, they cannot capture the semantic and contextual meaning of the whole sentence (idea) in the vector space. The vectorization of the whole sentence is one major innovation in text modeling: The transformer architecture generally outperforms word embedding models on standard tasks, and often by large margins (Wang et al., 2018, 2019), which utilizes a concept called attention (Vaswani et al., 2017). Attention makes it computationally tractable for a transformer model to consider a long sequence of text by selecting the most important parts of the sequence. Attention allows the training of large models on words and the complex contexts in which those words occur. This development resulted mainly in two kinds of categories, pre-trained sentence embedding models and text generation models which are discussed below.

Sentence embedding models vectorize the whole sentence into a vector space that keeps the semantic and contextual meaning of the entire sentence. The sentence embedding models are unsupervised techniques that do not require external data, e.g., Unsupervised Smooth Inverse Frequency (uSIF) (Ethayarajh, 2018) and Geometric Sentence embedding (GEM) (Yang et al., 2018). Some transformers allowed the tuning of parameters or training on their datasets to improve performance (if a large dataset is available), e.g., Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), Sentence Transformer (Reimers and Gurevych, 2019), MPNet (Song et al., 2020), Skip-Thought (ST) (Kiros et al., 2015), InferSent (Conneau et al., 2017), and Universal Sentence Encoder (USE) (Cer et al., 2018). In creativity research, the USE model is used to evaluate the novelty of ideas (Kenworthy et al., 2023). We argue that more exploration is needed to apply different, or combinations of sentence embedding models to evaluate creative ideas in an open-ended co-creation.

Text generation models generate new text that is similar to a given text prompt, such as Generative Pre-trained Transformer (GPT-3) (Brown et al., 2020), Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), and Long Short-Term Memory (LSTM) (Huang et al., 2022). In creativity research, one of the text-generated models, the Generative Adversarial Network (GAN) (Aggarwal et al., 2021), is used by Franceschelli and Musolesi (2022) to evaluate novelty, surprise, and relevance. We present two criticisms regarding using text generation models for evaluating open-ended ideas. First, text generation is specialized to generate text from a given text that could be useful for dialog generation, machine translation, chatbots, and prompt-based learning (Liu et al., 2023). Second, as the model becomes better at generating text with an improved understanding of language, it is more likely to generate text that closely resembles the input data rather than producing more novel or creative outputs. However, we argue that text generation models are not tested on a larger scale in creativity research, so future investigations could help understand these limits.

Finally, two conclusions are drawn from the above discussion. First, for single-word tasks in creativity research, word embedding models can be used, especially the GloVe embedding model, which is widely used. Word embedding models represent words in a high-dimensional vector space, enabling the computation of their contextual and semantic similarity with other words. Second, for open-ended co-creation resulting in ideas of sentence structure, sentence embedding models can be useful in three ways: (a) In open-ended ideation, mostly the ideas are in sentence structure, so these sentence models present the whole sentence in a vector space, capturing the semantic and contextual meaning of the whole sentence; (b) sentence embedding models outperform the word embedding models for textual similarity tasks; and (c) sentence embedding models can also be applied to small datasets and open-ended problems because these models are pre-trained over large corpora. Finally, we recommend not only validating sentence embedding models but also applying text generation models within a broader context of co-creation.

We concluded that sentence embedding models offer a powerful measure that can be used alongside statistical (Acar et al., 2021), word embedding models (Organisciak et al., 2023), and standard subjective scoring methods of the creative process and its output (Kenett, 2019).

Text classification approach refers to the automated categorization or labeling of textual data into predetermined classes or categories using machine learning classifiers. A large dataset is used for text classification, which is divided into training and testing (the usual ratio is 70% training and 30% testing datasets). An ML classifier learns from the training dataset and then uses the knowledge learned during training to categorize the testing dataset. Therefore, integrating text classification into automatic creativity evaluation depends on four key factors: the dataset, the selection of appropriate machine learning classifiers, the accuracy of the ML classifier, and the creativity dimensions being evaluated. These factors in the reviewed studies using the text classification approach are highlighted in Table 2.

Using text classification, it is essential to consider the dataset factor for three reasons: First, the datasets used for classification need pre-processing and labeling. Pre-processing includes removing noisy or irrelevant information, and labeling includes giving a class label to each idea. Second, a large dataset is required to train the ML classifiers. The prediction capability of ML classifiers increases with an increase in the amount of data used for training. All studies reviewed in Table 2 except Stella and Kenett (2019) use more than a thousand ideas or solutions for the classification problem. A smaller dataset may need better or more balanced results. Third, ML classifiers trained on one type of data cannot be applied to another kind of data. For example, classifiers trained on datasets from the linguistic domain cannot be used to test data from the scientific domain.

Furthermore, classifier selection and accuracy are also critical. Regarding classifier selection, the working methods of ML classifiers are different and dependent on the nature of the dataset, e.g., SVM works well for multiclass classification, and random forest excels in scenarios involving numerical and categorical features. Similarly, logistic regression works on linear problems; the K-neighbor classifier is best for text, and SVM can also work for multiclass dataset classification. The Bayesian approach is a simple and fast algorithm. The reviewed studies lack arguments for using a specific classifier in their studies. Regarding the accuracy of ML, there is a risk of not getting high accuracy. Different automatic evaluators are used to evaluate model accuracy, such as confusion matrix, entropy, and sensitivity, as shown in Table 2. It is suggested to apply several classifiers, and one with high accuracy can be used for prediction in a similar domain.

Finally, the text classification approach can be applied to evaluate different dimensions of creativity; however, it requires a large, labeled dataset, which limits its application in creativity research. We also argue that the dataset's preparation and labeling might be expensive, which mitigates the advantages of automatic evaluation over manual creativity evaluation, e.g., accuracy, cost, and time. Furthermore, the text classification problems are domain-dependent. So, for creativity tasks, such as object use tasks and alternate use tasks, some public datasets are available that could apply to similar tasks. However, it is not useful for small and open-ended creative tasks because it is not enough to train an ML classifier and is domain-independent. In short, large dataset preparation, labeling, and domain dependence make the text classification approach less reliable and expensive than manual creativity evaluation.

Text mining employs NLP statistical computation to discover new information and patterns. It uses statistical indicators such as the frequency of words, word patterns, and correlation within words. Dumas et al. (2021) implemented four text-mining techniques and measured the elaboration score in Alternate Use Tasks (AUT). Elaboration was computed in four different ways: (1) unweighted word count method: count the number of words; (2) stop listed inclusion: a preliminary agreed list of stop words; (3) parts of speech include verbs, nouns, adjectives, and adverbs; and (4) inverse frequency weighting: commonness of a word in an initial corpus of text.

The above text-mining techniques are the basic statistical operations in NLP. Text mining holds the potential to handle a massive amount of data to discover new information, patterns, trends, relationships, etc., that could be useful in creativity research. Text-mining applications include search engines, product suggestion analysis, social media analytics, and trend analysis.

5.2. Automatically computed creativity dimensions

The scoping review noted 25 creativity dimensions computed automatically. However, our analysis reveals that these creativity dimensions are not sufficiently based on previous creativity research and theory. Therefore, we have found some theoretical and methodological inconsistencies that should be tackled in future research. In this line of argument, first, we highlight that some of the creativity dimensions studied in the scoping review are defined and computed, building links with the challenges or the creativity tasks designed for the experiment but not with a strong theoretical framework. For example, a category switch is defined as the similarity difference between two successive responses in object use tasks (Dunbar and Forster, 2009). Another example is the creativity dimensions of quality (reusability) and usefulness (Degree of completion) that are defined and computed in the context of programming problems (Manske and Hoppe, 2014). Second, another reason for the inconsistency among the dimensions of creativity is the variation in manifestations employed across the reviewed articles. Specifically, it has been observed that dimensions such as novelty (Prasch et al., 2020), similarity (LaVoie et al., 2020), and originality (Beaty and Johnson, 2021) are defined in a similar manner, a strong focus on the similarity between ideas or solutions. Moreover, these dimensions are often measured using semantic textual similarity, although different computational techniques are performed.

To mitigate these shortcomings, this scoping review has thoroughly analyzed the conceptual and computational framework used in each study and contributed to the emergence of seven core creativity dimensions that could be automatically evaluated and bring more consistency to this research area. These seven core creativity dimensions are novelty, elaboration, flexibility, value, feasibility, fluency, and others related to playful aspects of creativity, such as humor and recreational efforts. Following, we discuss each core creativity dimension identified and highlight the key aspects of its conceptual definition and computational approach.

Novelty is the first core dimension in automatic creativity research that is most evaluated in 59% of the reviewed studies. Despite this high interest, our revision indicates multifariousness in defining and measuring novelty. As a consequence of that, the reviewed studies refer to novelty using the following different words or manifestations, namely, (1) uniqueness: the uniqueness of a concept related to the other concepts (Camburn et al., 2019); (2) originality: how different the outcome is from standard/other solutions (Georgiev and Casakin, 2019) or semantic distance among ideas (Beaty and Johnson, 2021); (3) similarity: the similarity of meaning between multiple texts (LaVoie et al., 2020) or similarity distance between the texts (Olson et al., 2021); (4) diversity: the diversity of users' entered queries; (5) rarity: the rare combination or rare ideas (Karampiperis et al., 2014) or unique solution (Doboli et al., 2020); (6) common use: the difference between common and uncommon solutions; (7) surprise: that how much an artifact is different from existing attributes (Shrivastava et al., 2017); and (8) influence or the comparison of an artifact with other artifacts (Shrivastava et al., 2017).

Nonetheless, the diversity in labeling and defining the novelty dimension, our analysis identified the next six characteristics that could be included in defining novelty and assisting its automatic evaluation: (1) deviation from the standard, routine way of solving a given problem (Manske and Hoppe, 2014); (2) semantic distance between ideas (Beaty and Johnson, 2021); (3) similarity of meaning between multiple texts (LaVoie et al., 2020); (4) Semantic similarity of the user query to the concepts in the challenge; (5) combination of properties (Karampiperis et al., 2014); and (6) surprise and unexpected ideas (Shrivastava et al., 2017). These six characteristics involved in the definition of novelty in the studies reviewed give an account of the complexity of defining the novelty dimension and acknowledge the challenges in developing automatic measures for novelty.

Despite these challenges, the scoping review has highlighted some common computing approaches and techniques to measure novelty as a core dimension and they can be synthesized in the next five characteristics: (1) distance of the new solution to the existing solution (Manske and Hoppe, 2014); (2) semantic distance among ideas (Beaty and Johnson, 2021; Olson et al., 2021); (3) semantic similarity of user queries and relevant concepts in Wikipedia; (4) semantic distance between the clusters in a story; and (5) semantic distance between the consecutive fragments of the story (Karampiperis et al., 2014). It concludes that when developing an automatic evaluation of novelty, the semantic distance of a solution to existing solutions should be considered.

Value is the second core dimension identified in automatic creativity evaluation. The scoping review identified the next four concepts related to value (Shrivastava et al., 2017; Franceschelli and Musolesi, 2022): (1) overall value, which relates how an artifact is perceived by society (Georgiev and Casakin, 2019); (2) quality, this concept is mainly used for programming solutions when they embody specific attributes such as reliability, characterized by error-free operation; maintainability, denoting ease of maintenance; extensibility, encompassing scalability and simplified modification; and, adaptability, reflecting the flexibility to integrate new technologies seamlessly (Manske and Hoppe, 2014); (3) the concept of usefulness which is linked to the notion of correctness; and (4) the concept of adaptiveness, it pertains to useful solutions that effectively address specific problems (Jimenez-Mavillard and Suarez, 2022). In sum, these four concepts share a common meaning of usefulness and quality that could be considered the value dimension of creativity. Furthermore, from a computer science perspective, value, quality, usefulness, adaptiveness, and style are the non-functional characteristics related to quality attributes. These quality attributes have different computations depending on the nature of the task, e.g., quality and useful programming solutions are the reusability and scalability of computer programs (Manske and Hoppe, 2014), and usefulness is the degree of completing the task (Prasch et al., 2020). Therefore, the value dimension needs clear definitions and computation metrics like other dimensions.

The third core dimension used in automatic creativity evaluation is flexibility. Flexibility refers to one of the key executive functions of creative thinking (Boot et al., 2017), which drives individuals to follow diverse directions, dimensions, and pathways (Acar et al., 2021), more likely to produce highly creative ideas (Zhang et al., 2020). Creativity research defines flexibility in two distinct ways. First, it involves category switching (Dunbar and Forster, 2009; Acar et al., 2019; Mastria et al., 2021), which refers to the ability to transition from one semantic concept to another. Second, flexibility is also measured by the number of semantic categories, varieties (Dunbar and Forster, 2009), or topics generated during the creative process. Owing to variations in the definition of flexibility across creativity research, different computational approaches are employed to compute this dimension. On one side, flexibility as a category switch is a measure of the similarity of one idea to all existing ideas. Therefore, semantic similarity approaches are used to evaluate flexibility, such as LSA (Dunbar and Forster, 2009), network graphs (Cosgrove et al., 2021), and sentence embedding models. On the other side, flexibility identifies semantic categories, varieties, or topics that can be evaluated using text clustering (Sung et al., 2022) or topic modeling techniques [e.g., Latent Dirichlet Allocation (LDA); Chauhan and Shah, 2021] to categorize or extract different topics from the textual ideas. We argue that flexibility as a category switch could be the easiest way to compute because it acquires simple text similarities rather than identifying categories in the text, which involve more variables and algorithms.

Regarding elaboration as a core creativity dimension in automatic creativity evaluation, it is defined as the degree of elaboration to which the participants embellish their responses (Camburn et al., 2019; Dumas et al., 2021) or which gives further details on adding reasoning or cause to an idea. Automatic creativity evaluation captures the level of detail of an idea by counting the number of words used in the idea (Camburn et al., 2019). The scoping review has identified four different methods for evaluating the level of idea elaboration: (1) counting all words in an idea (Counting unweighted measures); (2) counting stop words (words that do not have semantic meanings); (3) counting nouns, verbs, and adverbs; and (4) specifying and counting adjectives (parts of speech inclusion) and uncommon words with high weight (inverse frequency weighting). An idea with more words is considered an elaborated idea. We argue that the above-adopted computation of elaboration may not capture conjunctions (Tuzcu, 2021) or reasoning words (Sedova et al., 2019; Hennessy et al., 2020), adding more explanation to the ideas. Therefore, we suggest the semantic search to specify the words that cause reasoning or words that give reason to the idea, such as because, therefore, and since.

Fluency is defined as the number of ideas generated during an ideation process. This scoping review showed that fluency is one of the core dimensions that finds consensus on its conceptual definition (number of ideas) and computational approach (counting ideas) (Dumas and Dunbar, 2014; Stella and Kenett, 2019). Creativity research claims that when there are more ideas, there is a greater chance of producing original ideas or products (Dumas and Dunbar, 2014). Fluency measurement is easy to implement and is independent of other ideas such as elaboration. Compared to novelty and flexibility, which require comparison with different ideas, fluency can be easily computed for each idea.

Feasibility is defined as the solution that is achievable in real practice (Georgiev and Casakin, 2019). The scoping review found transcendence and realization have been used as manifestations of feasibility as they refer to the achievement in real practice or transforming into reality (Jimenez-Mavillard and Suarez, 2022). These dimensions share the same characteristic of transforming an idea or solution into real practice, which is significant in creativity research. The creativity research highlights the significance of putting ideas into practice; however, the automatic computation of feasibility (Georgiev and Casakin, 2019), transcendence, and realization (Jimenez-Mavillard and Suarez, 2022) does not provide any rationale from the creativity research. Feasibility is mostly a product-oriented dimension and is mostly used in the ideation process, but finding transformable ideas into real practice is still a challenge to address. Therefore, it is a dimension that needs further research to automatically measure feasible, transcendent, and realistic ideas.

Finally, other dimensions associated with the playful aspects of creativity, such as humor (Simpson et al., 2019) and recreational effort (Karampiperis et al., 2014), were identified in the reviewed articles. Humor, representing the funniness of ideas, is typically measured through pairwise text comparison techniques. At the same time, recreational effort is defined as a solution that is difficult to achieve and is measured using clustering methods. These dimensions contribute to the playful nature of creativity, so it is essential to establish clear definitions and develop suitable computational approaches from both psychological and computer science perspectives.

6. Conclusion

This article has the objective of conducting a scoping review of automatic creativity evaluation from creativity and computer science perspectives. To meet this objective, we defined two research questions: The first identifies the NLP approaches and techniques used in automatic creativity, and the second analyzes which and how different creativity dimensions are computed.

The first research question's contributions are multi-fold: (1) identifying the existing ML approaches and techniques in automatic creativity evaluation; (2) categorizing the approaches into different groups for deep compilation, e.g., text similarity, text classification, and text mining. Among these, text similarity is commonly used; (3) classifying creativity evaluation studies into different techniques accordingly, e.g., classifying studies in text similarity approaches using various techniques such as string similarity, corpus-based similarity, and knowledge-based similarity. Our results showed that corpus-based methods are widely used for automatic creativity evaluation. Corpus-based techniques, LSA (Dunbar and Forster, 2009; Dumas and Dunbar, 2014; LaVoie et al., 2020) and GloVe algorithm (Beaty and Johnson, 2021; Olson et al., 2021), have shown a positive correlation with human experts' similarity scores; (4) identifying the limitations of the critical challenge and identifying alternative techniques, for example, statistical and word embedding techniques are generally used, but they cannot capture the semantic and contextual meaning of a whole sentence; and (5) providing a broad overview of all existing automatic creativity to give a deeper understanding of all the approaches. We concluded that word embedding models, especially GloVe, work better for single-word tasks, and for open-ended ideas in sentence structure, sentence embedding models could provide promising results.

The second research question's contributions are also multi-fold: first, we have examined what creativity dimensions are automatically evaluated in the different articles analyzed in this scoping review. In contrast to creativity research, which has standardized tests that evaluate four specific dimensions, 25 different creativity dimensions are found in automatic creativity evaluation. Second, the scoping review has analyzed how these dimensions are defined and measured in automatic creativity evaluation. We found similarities in the definitions and computations of different creativity dimensions. Finally, based on a thorough analysis of the definitions and computations used in the studies, we characterized the 25 dimensions into seven core dimensions. This analysis helps elaborate a coherent and consistent framework about core creativity dimensions and their computation.

The overall contributions of this scoping review bridge the realms of computer science and education. For computer scientists, this review provides insights to refine existing NLP approaches and provides opportunities for developing more novel NLP methods for evaluating and promoting creativity. Meanwhile, educators can use these automatic evaluations as pedagogical tools in real-world classroom practices. The implications of automatic creativity evaluation could help assess and nurture creativity, which is becoming an explicit part of educational policy initiatives and curricula. Ultimately, this scoping review leverages AI as a valuable tool in evaluating and enhancing creativity capable of equipping future citizens with the necessary competencies to generate innovative solutions to the world's complex economic, environmental, and social challenges.

6.1. Limitations and future work

This scoping review has two limitations, which may have conditioned our results. The first limitation could be the search keyword strategy, which may be insufficient to include key articles in our field of study. Second, the exclusion and inclusion criteria may suffer from the omission of relevant studies that could have answered our research questions. We tried to mitigate this risk by carefully constructing an inclusive search string and providing explicit inclusion and exclusion criteria with co-authors' consensus.

In future, concluding from this scoping review, we intend to design experimental research to evaluate the reliability of deep learning models such as sentence embedding models to measure the novelty of ideas in an open-ended co-creative process. Furthermore, we also suggest using text generation models to recommend diverse hints to improve divergent thinking in the creative process. Regarding the automatic evaluation of creativity dimensions, our review highlighted that there is still a research gap in studies that fully automate the main core dimensions of creativity. So, we plan to simultaneously measure different core creativity dimensions by evaluating idea datasets with ML techniques. Finally, the development of reliable and automatic evaluation of the different dimensions of creativity could be the seed for the design and the delivery of real-time recommendations during the creative process that could trigger students' creativity.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

IU has contributed in the conceptualization of the paper, methodology and investigation; he has participated in writing the original manuscript, revision and edition. MP is the principal investigator of the research project and she has designed the project, she has also contributed in the conceptualization of the paper, methodology and investigation; she has participated in writing the manuscript, revision and edition. Both authors contributed to the article and approved the submitted version.

Funding

This research has been funded by the Ministry of Science and Innovation of the Government of Spain under Grants EDU2019-107399RB-I00 and PID2022-139060OB-I00.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C., and Organisciak, P. (2021). Applying automated originality scoring to the verbal form of torrance tests of creative thinking. *Gifted Child Quart.* 67, 3–17. doi: 10.1177/00169862211061874

Acar, S., Runco, M. A., and Ogurlu, U. (2019). The moderating influence of idea sequence: A re-analysis of the relationship between category switch and latency. *Person. Indiv. Differ.* 142, 214–217. doi: 10.1016/j.paid.2018.06.013

Aggarwal, A., Mittal, M., and Battineni, G. (2021). Generative adversarial network: An overview ofvtheory and applications. *Int. J. Inform. Manage. Data Insights* 1, 100004. doi: 10.1016/j.jjimei.2020.100004

Bae, S. S., Kwon, O.-H., Chandrasegaran, S., and Ma, K.-L. (2020). "Spinneret: aiding creative ideationvthrough non-obvious concept associations," in *Proceedings of the 2020 CHI Conference on HumanvFactors in Computing Systems* 1–13. doi: 10.1145/3313831.3376746

Beaty, R. E., and Johnson, D. R. (2021). Automating creativity assessment with semdis: An open platformvfor computing semantic distance. *Behav. Res. Methods* 53, 757–780. doi: 10.3758/s13428-020-01453-w

Birkey, R., and Hausserman, C. (2019). "Inducing creativity in accountants' task performance: The effects of background, environment, and feedback," in *Advances in Accounting Education: Teaching and Curriculum Innovations* (Emerald Publishing Limited) 109–133. doi: 10.1108/S1085-462220190000022006

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.5555/944919.944937

Boot, N., Baas, M., Mühlfeld, E., de Dreu, C. K., and van Gaal, S. (2017). Widespread neural oscillations in the delta band dissociate rule convergence from rule divergence during creative idea generation. *Neuropsychologia* 104, 8–17. doi: 10.1016/j.neuropsychologia.2017.07.033

Bozkurt Altan, E., and Tan, S. (2021). Concepts of creativity in design-based learning in STEM education. *Int. J. Technol. Design Educ.* 31, 503–529. doi: 10.1007/s10798-020-09569-y

Braun, D., Hernandez Mendez, A., Matthes, F., and Langen, M. (2017). "Evaluating natural language understanding services for conversational question answering systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (Saarbrucken, Germany: Association for Computational Linguistics) 174–185. doi: 10.18653/v1/W17-5522

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Proc. Syst.* 33, 1877–1901. doi: 10.48550/arXiv.2005.14165

Camburn, B., He, Y., Raviselvam, S., Luo, J., and Wood, K. (2019). "Evaluating crowdsourced design concepts with machine learning," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* 7. doi: 10.1115/DETC2019-97285

Cer, D., Yang, Y., Kong, S.-,y., Hua, N., Limtiaco, N., John, R. S., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chauhan, U., and Shah, A. (2021). Topic modeling using latent dirichlet allocation: A survey. *ACM Comput. Surv.* 54, 1–35. doi: 10.1145/3462478

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Cosgrove, A. L., Kenett, Y. N., Beaty, R. E., and Diaz, M. T. (2021). Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan. *Cognition* 211, 104631. doi: 10.1016/j.cognition.2021.104631

De Stobbeleir, K. E., Ashford, S. J., and Buyens, D. (2011). Self-regulation of creativity at work: The role of feedback-seeking behavior in creative performance. *Acad. Manage. J.* 54, 811–831. doi: 10.5465/amj.2011.64870144

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dickson, K., and Yeung, C. A. (2022). PRISMA 2020 updated guideline. *Br. Dental J.* 232, 760–761. doi: 10.1038/s41415-022-4359-7

Doboli, S., Kenworthy, J., Paulus, P., Minai, A., and Doboli, A. (2020). "A cognitive inspired method for assessing novelty of short-text ideas," in *2020 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 1–8. doi: 10.1109/IJCNN48605.2020.9206788

Dumas, D., and Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Think. Skills Creat.* 14, 56–67. doi: 10.1016/j.tsc.2014.09.003

Dumas, D., Organisciak, P., Maio, S., and Doherty, M. (2021). Four text-mining methods for measuring elaboration. *J. Creat. Behav.* 55, 517–531. doi: 10.1002/jocb.471

Dunbar, K., and Forster, E. (2009). "Creativity evaluation through latent semantic analysis," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31.

Ethayarajh, K. (2018). "Unsupervised random walk sentence embeddings: A strong but simple baseline," in *Proceedings of The Third Workshop on Representation Learning for NLP* 91–100. doi: 10.18653/v1/W18-3012

Franceschelli, G., and Musolesi, M. (2022). Deepcreativity: measuring creativity with deep learning techniques. *Intell. Artif.* 16, 151–163. doi: 10.3233/IA-220136

George, T., and Wiley, J. (2020). Need something different? Here's what's been done: Effects of examples and task instructions on creative idea generation. *Memory Cogn.* 48, 226–243. doi: 10.3758/s13421-019-01005-4

Georgiev, G. V., and Casakin, H. (2019). "Semantic measures for enhancing creativity in design education," in *Proceedings of the Design Society: International Conference on Engineering Design* (Cambridge: Cambridge University Press), 369–378. doi: 10.1017/dsi.2019.40

Gong, Z., Shan, C., and Yu, H. (2019). The relationship between the feedback environment and creativity: a self-motives perspective. *Psychol. Res Behav. Manag.* 12, 825–837. doi: 10.2147/PRBM.S221670

Gong, Z., and Zhang, N. (2017). Using a feedback environment to improve creative performance: a dynamic affect perspective. *Front. Psychol.* 8, 1398. doi: 10.3389/fpsyg.2017.01398

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *J. Creat. Behav.* 1, 3–14. doi: 10.1002/j.2162-6057.1967.tb00002.x

Guo, Y., Lin, S., Williams, Z. J., Zeng, Y., and Clark, L. Q. C. (2023). Evaluative skill in the creativeprocess: A cross-cultural study. *Think. Skills Creativ.* 47, 101240. doi: 10.1016/j.tsc.2023.101240

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory Cogn.* 45, 233–244. doi: 10.3758/s13421-016-0659-y

Hennessy, S., Howe, C., Mercer, N., and Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Cult. Soc. Interact.* 25, 100404. doi: 10.1016/j.lcsi.2020.100404

Hofmann, T. (1999). "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'99* (New York, NY, USA: Association for Computing Machinery), 50–57. doi: 10.1145/312624.312649

Huang, R., Wei, C., Wang, B., Yang, J., Xu, X., Wu, S., et al. (2022). Well performance prediction based on long short-term memory (lstm) neural network. *J. Petroleum Sci. Eng.* 208, 109686. doi: 10.1016/j.petrol.2021.109686

Jimenez-Mavillard, A., and Suarez, J. L. (2022). A computational approach for creativity assessment of culinary products: the case of elbulli. *AI Soc.* 37, 331–353. doi: 10.1007/s00146-021-01183-3

Johnson, D. R., and Hass, R. W. (2022). Semantic context search in creative idea generation. *J. Creat. Behav.* 56, 362–381. doi: 10.1002/jocb.534

Kang, Y., Sun, Z., Wang, S., Huang, Z., Wu, Z., and Ma, X. (2021). "Metamap: Supporting visual metaphor ideation through multi-dimensional example-based exploration," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 1–15. doi: 10.1145/3411764.3445325

Karampiperis, P., Koukourikos, A., and Koliopoulou, E. (2014). "Towards machines for measuring creativity: The use of computational tools in storytelling activities," in *2014 IEEE 14th International Conference on Advanced Learning Technologies* 508–512. doi: 10.1109/ICALT.2014.150

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Curr. Opin. Behav. Sci.* 27, 11–16. doi: 10.1016/j.cobeha.2018.08.010

Kenworthy, J. B., Doboli, S., Alsayed, O., Choudhary, R., Jaed, A., Minai, A. A., et al. (2023). Toward the development of a computer-assisted, real-time assessment of ideational dynamics in collaborative creative groups. *Creativ. Res. J.* 35, 396–411. doi: 10.1080/10400419.2022.2157589

Kim, S., Choe, I., and Kaufman, J. C. (2019). The development and evaluation of the effect of creative problem-solving program on young children's creativity and character. *Think. Skills Creativ.* 33, 100590. doi: 10.1016/j.tsc.2019.100590

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., et al. (2015). "Skip-thought vectors," in *Advances in Neural Information Processing Systems* 28.

Kovalkov, A., Paaßen, B., Segal, A., Pinkwart, N., and Gal, K. (2021). Automatic creativity measurement in scratch programs across modalities. *IEEE Trans. Learn. Technol.* 14, 740–753. doi: 10.1109/TLT.2022.3144442

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., and Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educ. Psychol. Measur.* 80, 399–414. doi: 10.1177/0013164419860575

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, Y., Du Ying, X. I. E., Liu, C., Yang, Y., Li, Y., and Qiu, J. (2023). A meta-analysis of the relationship 649 between semantic distance and creative thinking. *Adv. Psychol. Sci.* 31, 519. doi: 10.3724/SP.J.1042.2023.00519

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3560815

Manske, S., and Hoppe, H. U. (2014). "Automated indicators to assess the creativity of solutions to programming exercises," in *2014 IEEE 14th International Conference on Advanced Learning Technologies* 497–501. doi: 10.1109/ICALT.2014.147

Marrone, R., Cropley, D. H., and Wang, Z. (2022). Automatic assessment of mathematical creativity using natural language processing. *Creat. Res. J.* 2022, 1–16. doi: 10.1080/10400419.2022.2131209

Martin, D. I., and Berry, M. W. (2007). "Mathematical foundations behind latent semantic analysis," in *Handbook of Latent Semantic Analysis* 35–56.

Mastria, S., Agnoli, S., Zanon, M., Acar, S., Runco, M. A., and Corazza, G. E. (2021). Clustering and switching in divergent thinking: Neurophysiological correlates underlying flexibility during idea generation. *Neuropsychologia* 158, 107890. doi: 10.1016/j.neuropsychologia.2021.107890

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., and Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* 18, 1–7. doi: 10.1186/s12874-018-0611-x

Olivares-Rodríguez, C., Guenaga, M., and Garaizar, P. (2017). Automatic assessment of creativity in heuristic problem-solving based on query diversity. *DYNA* 92, 449–455. doi: 10.6036/8243

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., and Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proc. Nat. Acad. Sci.* 118, e2022340118. doi: 10.1073/pnas.2022340118

Organisciak, P., Newman, M., Eby, D., Acar, S., and Dumas, D. (2023). How do the kids speak? Improving educational use of text mining with child-directed language models. *Inf. Learn. Sci.* 124, 25–47. doi: 10.1108/ILS-06-2022-0082

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 1532–1543. doi: 10.3115/v1/D14-1162

Plucker, J. A., Meyer, M. S., Karami, S., and Ghahremani, M. (2023). "Room to run: Using technology to move creativity into the classroom," in *Creative Provocations: Speculations on the Future of Creativity, Technology and Learning* (Springer) 65–80. doi: 10.1007/978-3-031-14549-0_5

Prasch, L., Maruhn, P., Brünn, M., and Bengler, K. (2020). "Creativity assessment via novelty and usefulness (canu) – approach to an easy to use objective test tool," in *Proceedings of the Sixth International Conference on Design Creativity (ICDC)* 019–026. doi: 10.35199/ICDC.2020.03

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 687 5485–5551. doi: 10.48550/arXiv.1910.10683

Rafner, J., Biskjær, M. M., Zana, B., Langsford, S., Bergenholtz, C., Rahimi, S., et al. (2022). Digital games for creativity assessment: strengths, weaknesses and opportunities. *Creat. Res. J.* 34, 28–54. doi: 10.1080/10400419.2021.1971447

Reimers, N., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084.*

Rominger, C., Benedek, M., Lebuda, I., Perchtold-Stefan, C. M., Schwerdtfeger, A. R., Papousek, I., et al. (2022). Functional brain activation patterns of creative metacognitive monitoring. *Neuropsychologia* 177, 108416. doi: 10.1016/j.neuropsychologia.2022.108416

Said-Metwaly, S., Van den Noortgate, W., and Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity.* 4, 238–275. doi: 10.1515/ctra-2017-0013

Sawyer, R. K. (2011). *Explaining creativity: The science of human innovation* (Oxford university press) Sawyer, R. K. (2021). The iterative and improvisational nature of the creative process. *J. Creat.* 31, 100002. doi: 10.1016/j.yjoc.2021.100002

Sawyer, R. K. (2022). The dialogue of creativity: Teaching the creative process by animating student work as a collaborating creative agent. *Cogn. Instruct.* 40, 459–487. doi: 10.1080/07370008.2021.1958219

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., Drexlerova, A., et al. (2019). Do those who talk more learn more? the relationship between student classroom talk and student achievement. *Learn. Instruct.* 63, 101217. doi: 10.1016/j.learninstruc.2019.101217

Shrivastava, D., Ahmed, C. G. S., Laha, A., and Sankaranarayanan, K. (2017). A machine learning approach for evaluating creative artifacts. *ArXiv abs/1707.05499.*

Simpson, E., Do Dinh, E.-L., Miller, T., and Gurevych, I. (2019). "Predicting humorousness and metaphor novelty with gaussian process preference learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 5716–5728. doi: 10.18653/v1/P19-1572

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Proc. Syst.* 33, 16857–16867. doi: 10.48550/arXiv.2004.09297

Stella, M., and Kenett, Y. N. (2019). Viability in multiplex lexical networks and machine learning characterizes human creativity. *Big Data Cogn. Comput.* 3, 45. doi: 10.3390/bdcc3030045

Sung, Y.-T., Cheng, H.-H., Tseng, H.-C., Chang, K.-E., and Lin, S.-Y. (2022). "Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques," in *Psychology of Aesthetics, Creativity, and the Arts.* doi: 10.1037/aca0000450

Toma, J. D. (2011). "Approaching rigor in applied qualitative," in *The SAGE Handbook for Research in Education: Pursuing Ideas as the Keystone of Exemplary Inquiry* 263–281. doi: 10.4135/9781483351377.n17

Torrance, E. P. (2008). *The Torrance Tests of Creative Thinking Norms—Technical Manual Figural (Streamlined) Forms a and b. 1998.* Bensenville, IL: Scholastic Testing Service.

Tuzcu, A. (2021). The impact of google translate on creativity in writing activities. *Lang. Educ. Technol.* 1, 40–52.

Vartanian, O., Smith, I., Lam, T. K., King, K., Lam, Q., and Beatty, E. L. (2020). The relationship between methods of scoring the alternate uses task and the neural correlates of divergent thinking: Evidence from voxel-based morphometry. *NeuroImage* 223, 117325. doi: 10.1016/j.neuroimage.2020.117325

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* 30.

Vo, H., and Asojo, A. (2018). Feedback responsiveness and students' creativity. *Acad. Exch. Quart.* 1, 53–57.

Wagire, A. A., Rathore, A., and Jain, R. (2020). Analysis and synthesis of industry 4.0 research landscape: Using latent semantic analysis approach. *J. Manuf. Technol. Manag.* 31, 31–51. doi: 10.1108/JMTM-10-2018-0349

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. (2019). "Superglue: A stickier benchmark for general-purpose language understanding systems," in *Advances in neural Information Processing Systems* 32.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461.*

Williams, F. (1980). *Creativity Assessment Packet (CAP).* Buffalo, NY: D. O. K. Publishers Inc.

Yang, Z., Zhu, C., and Chen, W. (2018). Parameter-free sentence embedding via orthogonal basis. *arXiv preprint arXiv:1810.00438.*

Zhang, W., Sjoerds, Z., and Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage* 210, 116572. doi: 10.1016/j.neuroimage.2020.116572

Zuñiga, D., Amido, T., and Camargo, J. (2017). "Communications in computer and information science," in *Colombian Conference on Computing* (Cham: Springer).