What’s Creative About Sentences? A Computational Approach to Assessing Creativity in a Sentence Generation Task

Theresa J. Weinstein a, Simon Majed Ceh b, Christoph Meinel a, and Mathias Benedek a

Hasso Plattner Institute, University of Potsdam; University of Graz

ABSTRACT
Evaluating creativity of verbal responses or texts is a challenging task due to psychometric issues associated with subjective ratings and the peculiarities of textual data. We explore an approach to objectively assess the creativity of responses in a sentence generation task to 1) better understand what language-related aspects are valued by human raters and 2) further advance the developments toward automating creativity evaluations. Over the course of two prior studies, participants generated 989 four-word sentences based on a four-letter prompt with the instruction to be creative. We developed an algorithm that scores each sentence on eight different metrics including 1) general word infrequency, 2) word combination infrequency, 3) context-specific word uniqueness, 4) syntax uniqueness, 5) rhyme, 6) phonetic similarity, and similarity of 7) sequence spelling and 8) semantic meaning to the cue. The text metrics were then used to explain the averaged creativity ratings of eight human raters. We found six metrics to be significantly correlated with the human ratings, explaining a total of 16% of their variance. We conclude that the creative impression of sentences is partly driven by different aspects of novelty in word choice and syntax, as well as rhythm and sound, which are amenable to objective assessment.

The important and ongoing debate about various aspects related to scoring creativity, as well as developments in automated scoring methods motivate researchers to go deeper into psychometric issues and potential advances of creativity assessment. The present study tries to add to this discourse by investigating what properties of sentences contribute to their creativity. To this end, we developed a set of metrics to predict variance in human creativity ratings of four-word sentences, the output of a divergent thinking task used in a set of previous studies. Advancing this issue will help to develop better methods for both human and automated assessment of the creative quality of verbal responses to creativity tasks.

At some point in most creativity studies, human raters have to take on the tedious task of judging participants’ responses regarding their creativity. This judgment task is both cost-ineffective and raters vary in discernment as they differ in leniency (e.g., Benedek et al., 2016; Ceh, Edelmann, Hofer, & Benedek, 2022; Primi, Silvia, Jauk, & Benedek, 2019; Silvia, 2008), are affected by their emotional state (Mastria, Agnoli, & Corazza, 2019) and general workload (Forthmann et al., 2017). Still, raters typically agree substantially on the creativity of various ideas or products (Benedek et al., 2016). Thus, subjective scoring methods, centered around a consensual definition of what is deemed creative or not (Amabile, 1982), are still widely considered the “gold standard” in creativity assessment, possibly because this approach respects the construct’s socio-cultural embeddedness (e.g., Glăveanu, 2015). As researchers are trying to improve the quality of creativity evaluation, psychometric aspects of scoring procedures are subject to an extensive and everlasting debate (Amabile, 1982; Barbot, Reiter-Palmon, Barbot, & Reiter-Palmon, 2019; Shapiro, 1970). This debate is crucial because the implementation of consensual assessment is still heterogenous, complicating comparisons across studies (Cseh & Jeffries, 2019). One possibility to overcome some of the limitations imposed by consensus-based subjective rating procedures is via automated scoring methods. If researchers are able to model subjective rating behavior using automated methods, concerns regarding subjectivity and effort associated with human ratings could be addressed. Indeed, in light of the critical evaluation of existing practices in
creativity assessment in recent years (cf. Barbot et al., 2019), automated assessment techniques have gained considerable traction.

Particularly in the domain of divergent thinking tasks (e.g., the Alternate Uses Task; AUT; Guilford, 1967), researchers have explored various approaches to automate the scoring of responses (e.g., Acar & Runco, 2019; Beaty & Johnson, 2021; Beaty, Johnson, Zeitlen, & Forthmann, 2022; Beketayev & Runco, 2016; Dumas, Organisciak, & Doherty, 2021; Dumas & Runco, 2018; Forthmann, Oyebade, Ojo, Günther, & Holling, 2019; Heinen & Johnson, 2018; Johnson et al., 2022; Zedelius, Mills, & Schooler, 2019). Individual differences in divergent thinking performance are generally assessed on two dimensions: Idea quantity and quality (Reiter-Palmon, Forthmann, & Barbot, 2019). While idea quantity can be determined by counting the number of generated ideas or concepts, assessing the quality of the ideas in terms of their originality, and effectiveness—usually considered the two central aspects of a creative idea or product (Diedrich, Benedek, Jauk, & Neubauer, 2015; Runco & Jaeger, 2012)—is more challenging. Currently, the most common computational approach for assessing idea quality in verbal divergent thinking tasks is to compute the semantic distance between stimulus and responses using natural language processing methods such as semantic network analyses (Acar & Runco, 2014) or latent semantic analysis (LSA; e.g., Beaty & Johnson, 2021; Beaty et al., 2022; Dumas & Runco, 2018; Hass, 2017; Heinen & Johnson, 2018; Kenett, 2019; Landauer, Foltz, & Laham, 1998; Prabhakaran, Green, & Gray, 2014). Responses that are more unrelated to the stimulus in terms of its semantic distance are considered more original, and thus more creative (Kenett, 2019; Mednick, 1962). Most scoring algorithms use semantic text-mining models trained on a large text corpus (e.g., GloVe; see Dumas et al., 2021, for a model comparison). However, researchers have also discussed limitations of this approach. For example, Forthmann and colleagues (2019) have observed that objective scores are biased by response length (i.e., elaboration). Furthermore, objective scores seem to reflect more strongly the novelty (viz. originality) of responses than their task-appropriateness (viz. effectiveness) (cf. Beaty et al., 2021; Heinen & Johnson, 2018).

While text is generally well amenable to objective scoring, verbal creativity involves important conceptual challenges. In part, this is due to different reference points (e.g., process vs. outcome), aggravated by the lack of clarity about the exact definition of creativity within creative writing. Early works focused on concept modification, linking antonymy to creative potential (Rothenberg, 1973) and elaborating on typicality and irrelevance as linguistic factors for creativity (Estes & Ward, 2002; Turkman & Runco, 2019). Turkman and Runco (2019) demonstrated that the use of certain keywords is linked to introducing original ideas in written text, showing that the presence of keywords correlated with ideational density, a predictor of creative potential (Runco, Turkman, Acar, & Nural, 2017). Others view writing generally as creative, involving idea production with unlimited (linguistic) choices and combinations (e.g., Sharples, 1996). When assessing the quality of writing, Fürst, Ghisletta, and Lubart (2017) had raters evaluate texts for originality (e.g., surprise) and quality (e.g., style)—similar to Kettler and Bower (2017) who emphasize originality and elaboration. Ashton and Davies (2015) found the creativity of science fiction and fantasy pieces to be related to the effect on the audience, but also stress technical writing quality. The use of rarer words and neologism in the first 1000 words of best-selling books were related to increased popularity (Form, 2019), indicating that linguistic properties on the word-level may play a role for success in this creative domain. In a systematic literature review, D’Souza (2021) summarized these and additional studies to arrive at four key characteristics related to creativity in narrative writing: Meaning and relevance (e.g., aesthetic value), reader immersion (e.g., flow and clarity), development and control (e.g., language vividness), and originality (e.g., surprise and distinctive style; see Table 2 in D’Souza, 2021, for a more detailed overview). In summary, assessment of creative writing typically considers originality (e.g., word infrequency, high associative distance), while the effectiveness of verbal creative responses may often be task-dependent (e.g. Toubia, Berger, & Eliashberg, 2022) and thus harder to assess.

A recent study found that automated methods may even capture the creative quality of narrative text in the context of short stories (Johnson et al., 2022; Zedelius et al., 2019). In the study by Johnson and colleagues (2022), short stories were computationally assessed for the extent to which they connect divergent ideas with a metric called “word-to-word semantic diversity” (w2w SemDiv). The authors tested six automated implementation models for this metric and found that the best-performing model employed bidirectional encoder representations from transformers (BERT; Devlin, Chang, Lee, & Toutanova, 2018), explaining up to 72% of variance in the human creativity ratings. These findings highlight that semantic distance measures can also be harnessed to measure text qualities beyond the typical conceptual distance between stimulus and response. However, automated assessments appear still challenged to indicate other complex characteristics of text.
responses (e.g. style, coherence) that make them be perceived as creative (cf. D’Souza, 2021; Fürst et al., 2017).

Aims of the study

Rater-based creativity assessments often rely on quantifying creative performance in terms of a holistic creativity score, which offers little insight on why exactly a response is considered more or less creative. In this context, computerized methods offer a promising and efficient approach to scoring divergent thinking tasks (e.g., Beaty & Johnson, 2021; Beketayev & Runco, 2016; Dumas et al., 2021; Johnson et al., 2022). However, so far they have also mainly focused on quantifying a single aspect of response quality (but see Toubia et al., 2022), which is most commonly response originality (but see Johnson et al., 2022). This study aimed to gain a deeper understanding of what factors make a verbal divergent thinking response such as a sentence be perceived as creative. To this end, we developed eight scoring metrics that aim to quantify distinct aspects of verbal creativity in a sentence and examined their individual relevance in explaining variance of subjective creativity evaluations. By developing and evaluating new metrics for the objective scoring of sentence-level creativity (and making our data and analysis code openly available), we contribute to the ongoing efforts to automate the creativity scoring of verbal responses.

Methods

Data, analysis scripts and supplementary materials can be found online (https://doi.org/10.17605/OSF.IO/YX4TS).

Data set

The study makes use of an existing data set that was collected across two studies (Ceh et al., 2020, 2021). In these studies, a total of 84 participants ($M_{age} = 23.37, SD = 3.05; 65\%$ female) generated sentences in a divergent thinking task: Based on German four-letter-words (e.g., STOP), participants had 20 seconds to generate a four-word-sentence, using each letter as a starting letter for each word in the resulting sentence (e.g., STOP → Patrick obliterated Susan’s toys). Participants were instructed to construct creative sentences that adhere to German grammar rules. They were not instructed regarding the type of sentence (i.e., they were also allowed to ask questions, make exclamations, etc.) and allowed to alter the order of letters, meaning that this task did not restrict the solution space other than in emphasizing grammar rules. Each participant worked on 18 out of 36 possible stimulus words. Participants provided 12 valid responses on average ($M = 11.77, SD = 3.02, min = 3, max = 18$), resulting in a total of 989 sentences (i.e., $65\%$ solution rate). As the data was collected in a neuroscience context, the participants expressed their responses orally via microphone. All sentences were transcribed by an experimenter, which ensured that the sentences followed correct (German) syntax and orthography.

Human creativity rating

All sentences were evaluated by eight raters, each judging the creativity of all responses on a single four-point scale ranging from zero (not creative at all) to three (very creative). The mean rating was $1.12 (SD = 0.52)$ and inter-rater reliability was good ($ICC_{2k} = .79; 95\%$ CI [.77, .80]).

Scoring algorithm

The scoring algorithm was developed in Python (v. 3.9) and can be accessed online (https://doi.org/10.17605/OSF.IO/YX4TS). It consists of eight metrics that aimed to capture different aspects of novelty (e.g., word infrequency) and effectiveness (e.g., linguistic style) of the sentences that might play a role in the human assessment of creativity (see Table 1 for an overview of all metrics).

Preprocessing

The algorithm prepared the data for further processing by removing unnecessary punctuation and excluding invalid sentences from subsequent analysis. The validity of the sentences was assessed by verifying that each sentence contained exactly four words and that the first letter of each word corresponded to the letters in the stimulus. Some of the metrics required lemmatization (i.e., converting each word in a sentence to its base form for better comparison of vocabulary across sentences) and part-of-speech tagging (i.e., labeling words according to their word class e.g., “proper noun,” “verb,” etc.). Both procedures were conducted automatically using the Python natural-language processing module “Hanover Tagger,” which offers a pre-trained model for German (Wartena, 2019). For scores that required lemmatization and part-of-speech tagging, we used the more accurate Hanover Tagger (Wartena, 2019) except for the similarity score, which was implemented entirely with the Python module “spaCy” (Explosion AI
Table 1. Overview of objective text metrics including a short description of their aim, implementation, output scale, and whether lemmatization was applied.

| Metric                        | What                                      | How                                           | Scale                | Lemmatization |
|-------------------------------|-------------------------------------------|-----------------------------------------------|----------------------|---------------|
| 1) General word infrequency   | Infrequency of words in a sentence        | Mean score based on inverted frequency values provided by a German online dictionary (DWDS). | Low to high infrequency (0–6) | Yes           |
| 2) Word combination infrequency | Infrequency of word combinations in a sentence | Total number of search results returned by Google, mapped on an (inverted) logarithmic scale, averaged across all possible combinations per sentence (= 6). | Low to high infrequency (log scale of base 10) | No            |
| 3) Context-specific word uniqueness | Uniqueness of word choice per trial | Mean normalized relative occurrence of words used in a participant’s response compared to other responses in a trial. | Low to high uniqueness (0–1) | Yes           |
| 4) Syntax uniqueness          | Uniqueness of syntax, i.e., sentence structure | Mean normalized relative occurrence of a particular sentence’s order of part-of-speech tags compared to all sentences in the entire dataset. | Low to high uniqueness (0–1) | No            |
| 5) Rhyme                      | Presence of words that rhyme              | Check if sounds in word endings are identical and count total number of rhymes. Words are converted to numerical sound representations according to Cologne Phonetics. | 0: no rhyme 1: 1 rhyme pair 2: 3 rhyming words 3: 2 rhyme pairs 4: all 4 words rhyme | No            |
| 6) Phonetic similarity        | Presence of similar sounds                | Find recurring similar sounds based the word’s numerical sound representation according to Cologne Phonetics, based on mean of results from normalized Levenshtein distance and longest common subsequence (LCS). | No to very similar sounds (0–1) | No            |
| 7) Sequence similarity        | Similarity between cue word and words in the sentence in terms of spelling | Maximum similarity ratio between cue word and each word in a sentence (“difflib.SequenceMatcher.ratio()” function). | Maximum similarity ratio per sentence (0–1) | No            |
| 8) Semantic similarity        | Similarity between cue word and sentence in terms of semantic meaning | Mean semantic similarity values (absolute cosine distance) between cue word and each word in the sentence. Implemented with SpaCy module (Explosion AI GmbH, 2022) using the available NLP pre-processing pipeline and a pre-trained model for German (“de_core_news_lg”). | No to high similarity (0–1) | Yes (SpaCy) |

GmbH, 2022) that comes with its own model for German lemmatization, part-of-speech tagging and word vectorization.

General word infrequency. The general word infrequency score assessed the (in)frequency in current language use of each word in the sentence. Each word’s frequency was looked up individually in its lemmatized form in the “Digitales Wörterbuch der deutschen Sprache” (DWDS; [Digital dictionary of the German language]), which provides frequency metrics derived from a large corpus of historic and contemporary texts (for more details, see https://www.dwds.de/d/worthaefigkeit). The DWDS word frequency values are provided on a logarithmic scale ranging from 0 to 6, which we inverted so that a score of 6 represents the highest possible infrequency (i.e., novelty). All four word infrequency values were averaged into a total word infrequency score for the entire sentence.

Word combination infrequency. The word combination infrequency score used the number of Google Search results to determine whether a particular word combination was more or less (un)usual. For each four-word sentence, all possible two-word combinations per sentence (= 6 combinations) were queried via a commercial API (https://www.scaleserp.com) and the number of total results returned by the Google Search algorithm was recorded (cf. Cilibrasi & Vitanyi, 2007). For word pairs that were not consecutive words in the given sentence, the search term combined the words with the “*” operator as placeholder for the missing words. A high number of search results implied common use of the combination, whereas few to no search results pointed toward an infrequent combination of words (cf. Johnson et al., 2022, for a similar approach based on semantic distances). The total number of search results was mapped to a logarithmic scale of base 10. The logarithmically increasing bins aimed to buffer unpredictable changes in the number of search results, since these may differ over time due to changes in Google’s search algorithm and the creation of new web content. The final word combination infrequency score, again, contained the inverted arithmetic mean of each word-pair’s frequency score. High scores pointed toward an infrequent use of a combination, whereas usual combinations with many results yielded low scores.

Context-specific word uniqueness. The context-specific word uniqueness score assessed participants’ word choice to solve a specific four-letter trial and compared
it to the overall word choice of all participants’ responses in the trial. If a participant chose unique word(s) that no one else used to solve the trial, the participant’s sentence was assigned a higher score than a participant’s sentence that contained a word that was also used by many others in the trial. For example, if many participants used a common word like the name “Anna” in their solution to the trial “AHNNE,” the uniqueness score was lower. To ensure correct comparisons across sentences, the score used the lemmatized form of each word.

To compute the uniqueness of words within the trial, the algorithm determined each word’s relative occurrence across all sentences of the trial. First, it counted the absolute occurrence of a particular word in a trial which was then put into context by dividing the absolute occurrence by the total number of all words used in this trial. This resulted in a relative occurrence value for each word.

The final score was computed according to the following normalization equation (1), which ensures that scores were mapped to a scale of 0 to 1, where higher scores described sentences with a unique word choice compared to all words used to solve the trial. Each word’s relative occurrence value \( f(i) \) in a given trial \( t \) was subtracted by the minimum occurrence value of the trial, averaged for the entire sentence and divided by the maximum occurrence value of the trial.

\[
score(t) = 1 - \frac{\sum_{i=1}^{4}(f(i) - \min(f))}{4 \times \max(f)}
\]

**Syntax uniqueness.** The syntax uniqueness score determined the uniqueness of a response’s sentence structure based on the sentence’s order of part-of-speech (POS) tags relative to the entire data set. Using the Hanover Tagger, each word’s part-of-speech (such as verbs, nouns, articles, etc.) was tagged according to the Stuttgart-Tübingen-Tag-Set (cf. Schiller, Teufel, Thielen, & Stückert, 1999) and then the relative occurrence of the POS-tag combination across the sample was determined. For example, the combination of POS-tags “NE VVFIN ADJA NN” (i.e., “Stefan iss trockene Linsen [Stefan eats dry lentils],” with NE = proper noun, VVFIN = finite verb, ADJA = adjective, NN = normal noun) was used by more participants than for instance “VVIMP PPER ART NN” (i.e., “Leih mir ein Hemd [Lend me a shirt],” with VVIMP = imperative, PPER = personal pronoun, ART = article, NN = normal noun). Similar to the previous score, the syntax score was normalized by taking the minimum and maximum occurrence of a particular part-of-speech combination into account, in order to yield a score between 0 and 1. A result close or equal to 0 signified a very commonly used syntax and a result closer to 1 a very unique syntax in the dataset.

**Rhyme.** The rhyme score determined whether there were word endings in each sentence that were identical or very similar and thus induce rhyme or rhythm. The score increased with the number of words that rhyme. Rhymes were detected using the “Cologne Phonetics,” an algorithm that converts the sounds in a word’s pronunciation to numerical representations (Postel, 1969). The Cologne Phonetics have been optimized for the German language (an equivalent English algorithm is “soundex”). If a word ending had the same sound, i.e., the same number, it was assumed that it rhymes. For example, the sentence “Studenten trinken reinen Ether” (“Students drink pure ether”) was converted into “822,626 27,646 766 027,” where the algorithm would detect three rhyming words (i.e., three times the number 6 at the ends of the first three-word representations). Note that the score did not differentiate between perfect and imperfect rhymes. The rhyme score reflects the presence and number of rhymes in a sentence on a scale from 0 to 4 with zero score corresponding to no rhyme, 1 for one rhyming word pair, 2 for three rhyming words, 3 for two rhyming word pairs, and 4 for a sentence in which all words rhyme. The position of rhyming words within a sentence was not taken into account.

**Phonetic similarity.** Similar to the rhyme score, the phonetic score employed the “Cologne Phonetics” to detect similarities in recurring sounds between all words in the sentence. Sentences that contained words with similar sounds, thereby evoking a phonetically more pleasant pronunciation, got a higher score than sentences in which all words sound very differently. For example, the sentence “Anna nimmt eine Hand” (“Anna takes a hand”) was converted into “0666206062”, which contains many similar numbers, i.e., sounds. For each word pair in the sentence, the similarity of numeric sound representations was checked using the Levenshtein distance (Levenshtein, 1966), which returns the minimum number of changes needed to convert one sound representation into the other and by determining the longest common subsequence (LCS) between the words’ numerical sound representations. Both measures are affected by string length, hence the Levenshtein distance was normalized by dividing the result by the length of the longer sound representation, and the LCS was normalized by dividing the result by the length of the shorter representation. The final phonetic score consisted of the mean of both normalized Levenshtein and
LCS results, averaged across the number of total possible combinations (= 6) per sentence. It resulted in a scale from 0 to 1, with scores close or equal to 1 implying a high similarity of sounds within the sentence.

**Sequence similarity.** The sequence similarity score determined similarity between the spelling of the cue and each word in a sentence. For example, the cue “AHNE” ([“ancestor”]) appeared very similar to the response words “Akne” ([“acne”]) or “ahnre” (conjugation of [“to sense”]) in terms of spelling, but not semantic meaning. The score computed a similarity ratio between each word in a sentence and the cue word based on Gestalt Pattern Matching principles (cf. Ratcliff & Metzener, 1988), an algorithm that looks for similarity between two sequences in a way that humans would intuitively do, thereby giving a more meaningful result than a Levenshtein distance or LCS. The score was implemented using the SequenceMatcher.ratio() function from Python’s difflib module (https://docs.python.org/3/library/difflib.html), which takes two strings as input sequences and outputs a float between 0 and 1. The similarity ratio is calculated with (2), where T represents the total number of elements in both sequences, and M the number of matches. Identical sequences result in a 1, while a 0 result implies no similarity at all.

\[
similarity\ ratio = \frac{2 \times M}{T}
\]  

The final sequence similarity score was defined as the maximum similarity ratio value for a given sentence. If the score was larger than ~0.6, it was very likely that a participant used at least one word in their sentence that was very close to the spelling of the cue word. Some similarity was to be expected due to the basic requirements of the experimental task, in which at least one letter of the cue always had to be identical with the initial letter of one of the words in the sentence.

**Semantic similarity.** The semantic similarity score computed the semantic similarity between each word in the sentence and the cue word to determine whether the meaning of the words in the sentence reflected the meaning of the stimulus word. We were interested in exploring whether the cue word inspired the content of the responses, since we assumed that the participants automatically processed the meaning of the cue word when reading it, although it was irrelevant to the task (cf. “Stroop effect,” MacLeod, 1991). Unlike in other divergent thinking tasks, where semantic distance reflects originality, semantic similarity between cue and response could be considered highly effective as participants pick up the original stimulus and successfully embed it in their generated sentence. The score was implemented with the freely available Python module “spaCy” (https://spacy.io) which provides a pre-trained NLP model for German based on large news corpora. SpaCy’s similarity function calculates the cosine distance between each word’s and the cue word’s word vector. To allow for meaningful vector transformation of words and accurate determination of their semantic similarities, we excluded proper nouns (i.e., names) from the sentences. Proper nouns were detected by using lemmatization and part-of-speech tagging provided by spaCy’s preprocessing pipeline.

All possible distances between a sentence’s words (except for proper nouns) and the cue word are averaged into a final score. Results range between 0 and 1 (absolute values of the cosine distance which can range between −1 to 1) where a result of 1 implies the highest possible similarity (= identical word) and a number close or equal to zero represents the lowest possible semantic similarity. If a word in the sentence could not be converted into a vector for technical reasons (empty vector warning) or if a sentence was wrongly tagged as containing proper nouns only, the whole sentence was excluded from this score.

**Results**

Table 2 shows descriptive statistics and intercorrelations of all measures (for an extended version that includes the ratings of each rater, separately, see Table S1). Figure 1 further shows histograms that reveal the distributional characteristics of the investigated variables, indicating that some variables did not follow a normal distribution (see also Table S2 and Figure S1). The most pronounced deviations from normality were found for the rhyme and semantic similarity scores (both positively skewed), as well as the context-specific word infrequency and syntax uniqueness scores (both negatively skewed). The creativity rating showed significant correlations with six of the eight objective test metrics (all p’s < .01). More creative sentences contained more infrequent words (general word infrequency; r = .20), more infrequent word combinations (word combination infrequency; r = .24), more unique words compared to other responses for a trial (context-specific word uniqueness; r = .30), and higher novelty in sentence structure (syntax uniqueness; r = .15). Higher rated creativity was also associated with the occurrence of
Table 2. Descriptive statistics, and intercorrelations of all objective measures.

| Measure                                | M   | SD  | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|----------------------------------------|-----|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| 1. Creativity rating                   | 1.12| 0.52| .20   | .13   | .25   | .36   | .24   | .66   |       |       |
| 2. General word infrequency            | 2.35| 0.73|       |       |       |       |       |       |       |       |
| 3. Word combination infrequency        | 3.13| 1.20| .24   | .18   | .29   | .63   | .70   |       |       |       |
| 4. Context-specific word uniqueness    | 0.81| 0.13| .30   | .25   | .36   | .24   | .66   |       |       |       |
| 5. Syntax uniqueness                   | 0.83| 0.24| .15   | -.15  | -.16  | -.08  |       |       |       |       |
| 6. Rhyme                                | 0.19| 0.42| .09   | .04   | -.01  | .10   |       |       |       |       |
| 7. Phonetic similarity                 | 0.29| 0.10| .10   | .02   | .07   | .03   | .17   |       |       |       |
| 8. Sequence similarity                 | 0.53| 0.16| -.05  | -.06  | -.18  | .01   | -.04  | .15   |       |       |
| 9. Semantic similarity                 | 0.09| 0.05| .07   | .01   | -.06  | -.11  | .02   | -.01  | .13   | .30   |

N = 989. Values in square brackets indicate the 95% confidence interval for each correlation. |r| > .062: p < .05; |r| > .082: p < .01; |r| > .104: p < .001.

Figure 1. Distribution of text metric scores.
rhyme ($r = .15$), and similar sounds (phonetic similarity; $r = .10$), suggesting that aspects other than novelty contributed to higher estimations of sentence creativity.

Since some of these measures were substantially intercorrelated, we examined their unique contribution in predicting human creativity ratings by means of a regression analysis. We entered all text metrics that showed significant zero-order correlations with the criterion as predictors. We applied bootstrapping to increase robustness in the face of partly skewed data, computing 95% percentile confidence intervals based on 2000 bootstrapped samples (Field & Wilcox, 2017). The model was statistically significant ($F_{6, 982} = 32.88$, $p < .001$, $R^2 = .167$, 95% CI [.13, .22]; see Table 3). All measures that initially showed significant zero-order correlations with the criterion independently predicted variance in sentence creativity, except for general word infrequency, which only showed a trend (see Table 3).

**Discussion**

This study investigated the feasibility of an objective creativity scoring of responses in the sentence generation task. We developed eight text metrics based on text analysis and natural language processing methods that aimed to quantify different aspects of response quality, which could be categorized in terms of focusing on either response novelty (i.e., characterized by the use of infrequent words or syntax, such as general word infrequency, word combination infrequency, context-specific word uniqueness, and syntax uniqueness) or response effectiveness (i.e., characterized by meeting certain constraints, such as rhyme, phonetic similarity, sequence similarity and semantic similarity). We found that most text metrics were correlated with the human creativity ratings and explained a total of 16% of their variance, supporting their validity as objective indicators of verbal creativity.

These findings add to the evidence showing that computational metrics relying on word properties and natural language processing methods reflect objective and efficient indices of aspects of response creativity that can potentially be used for automated scoring of verbal divergent thinking responses (e.g., Beaty & Johnson, 2021; Dumas et al., 2021; Forthmann et al., 2019; Heinen & Johnson, 2018; Johnson et al., 2022). This study extends previous work in several ways. First, objective scoring methods were applied to a new type of divergent thinking task, the sentence generation task, which has been previously used in both cognitive and neuroscientific research (Benedek, Fink, & Neubauer, 2006; Ceh et al., 2020, 2021). By asking for creative sentences with exactly four words, this task ensures a standardized response format that avoids issues associated with elaboration bias (Forthmann et al., 2019), while leaving room for creative expression, which offers favorable conditions for computational explorations of verbal creativity.

Second, while previous automated scoring work mostly focused on a single score, this study explored a larger set of eight metrics quantifying different textual aspects of response quality. Six of these metrics were significantly correlated with the rated creativity of the sentences, and five of them explained unique variance. This finding shows that it was possible to isolate several distinct aspects contributing to the creativity of sentences. The general and word combination infrequency metrics captured novelty in the responses based on the frequency of the words or word combinations in general language use. While the general word infrequency score used frequency values provided by a formal dictionary, the word combination infrequency score was based on the number of search results from the web search engine Google. The context-specific word uniqueness and syntax uniqueness metrics captured the distinctiveness of a response within the dataset, i.e., they evaluated whether a participant’s response was more unique in terms of word choice or syntax elements compared to other participants’ responses in a specific trial (context-specific word uniqueness) or the entire dataset (syntax uniqueness). While these four metrics referred to response novelty in different

| Predictor                      | $\beta$     | $\beta$ 95% CI | $p$   | $sr^2$ | $sr^2$ 95% CI | $r$ ($p$) |
|--------------------------------|------------|---------------|-------|--------|---------------|-----------|
| (Intercept)                    |            |               |       |        |               |           |
| General word infrequency       | .07        | [−.00, .15]   | .008  | .00    | [.00, .01]    | .20 (.<.001)|
| Word combination infrequency   | .11        | [.03, .19]    | .001  | .01    | [.00, .02]    | .24 (.<.001)|
| Context-specific word uniqueness| .25       | [.19, .32]    | <.001 | .06    | [.03, .08]    | .30 (.<.001)|
| Syntax uniqueness              | .18        | [.14, .23]    | <.001 | .03    | [.02, .05]    | .15 (.<.001)|
| Rhyme                          | .11        | [.04, .17]    | <.001 | .01    | [.00, .03]    | .15 (.<.001)|
| Phonetic similarity            | .07        | [.01, .13]    | .015  | .00    | [.00, .02]    | .10 (.<.01)|

CI, 95% percentile confidence interval based on 2000 bootstrapping samples; $r$, zero-order correlation; $sr^2$, squared semi-partial correlation; $\beta$, standardized regression weight.
ways, still all captured relevant, complementary aspects of response creativity (cf. Table 2 in D’Souza, 2021). The rhyme and phonetic similarity metrics detected the presence and number of rhymes in a response and whether there were words in a sentence that contained similar sounds in their pronunciation, thus resulting in a potentially more appealing response. Only two metrics (sequence similarity and semantic similarity) were not related to response creativity, suggesting that they capture subtle verbal aspects that played no important role for the perceived creativity of sentences. Hence, we found that several distinguishable factors of a sentence, amenable to computational scoring, contribute to its creativity, which can be further explored to better understand what language-related aspects are considered by human raters.

Third, significant predictors of response creativity included metrics reflecting both the novelty and effectiveness of the response, which are widely considered as the two central aspects of creativity (Diedrich et al., 2015; Runco & Jaeger, 2012). This addresses an important challenge in available objective scoring work, which has mostly focused on the novelty (viz. originality) of responses in terms of the semantic distance between cue and response. As a notable exception, one study proposed to use semantic distance measures to assess the semantic diversity of texts reflecting the distance between ideas in a story (Johnson et al., 2022); however, such a measure may be especially suited for longer text responses but be less applicable to single sentences. While we found that both novelty and effectiveness metrics played an independent role, novelty metrics together still explained more variance of creativity evaluations than effectiveness metrics, which is consistent with the finding that novelty is more strongly related to creativity than effectiveness (Diedrich et al., 2015). The present findings point to new ways of how both novelty and effectiveness of verbal responses can be considered in the context of the sentence generation task.

A closer look at the intercorrelations of the metrics revealed mainly expected patterns. The general word infrequency score is positively correlated with word combination infrequency, context-specific word uniqueness and rhyme, but negatively correlated with the syntax uniqueness score. The strong association between the general word and the word combination infrequency score is not surprising, since both scores measure similar aspects of word infrequency in general language use, although they are based on different sources (i.e., a formal dictionary versus Google search results). While context-specific word uniqueness was positively correlated with general word infrequency, syntax uniqueness was correlated negatively. The former relationship appears plausible, since it seems quite likely that when a participant uses a generally very infrequent word or word combination, that this also represents unique word choice among the responses of the other participants. However, it is interesting that the use of infrequent words is negatively correlated with unique syntax. It seems that participants could only focus on one of the two aspects, either using uncommon sentence structure or infrequent words and word combinations. This could possibly be due to the time constraints of the task, which only offered time for task strategies that supported some novelty aspects at the cost of others. In addition, there is a slight positive correlation between using infrequent words and rhyme, which might point to the circumstance that rhyming could help to find an uncommon word or that infrequent words are more conductive to rhymes with other words. Furthermore, we find correlations between rhyme and syntax uniqueness as well as rhyme and phonetic similarity. These metrics assess linguistic aspects contributing to the appeal and thus overall effectiveness of a sentence. Trying to fit a rhyme into a sentence seems to coincide with more unique sentence structure. Using rhyme, a rhythmic element, is also associated with more phonetic similarity between words. Both aspects support the fluent and more pleasant readability of the sentence and thus the effectiveness of the response, contributing to its overall creative impression.

The two remaining scores evaluating the similarity between the cue and the responses at a lexical and semantic level were not significantly correlated with the human creativity ratings. The sequence and semantic similarity scores aimed at detecting whether the responses were inspired either by the spelling or the semantic meaning of the cue. Contrary to other studies employing automatic evaluations of divergent thinking tasks (e.g., Beaty & Johnson, 2021; Dumas et al., 2021; Forthmann et al., 2019; Johnson et al., 2022) the measure of semantic distance applied in this study was not associated with the human creativity ratings, implying that the relationship between the response and the cue was not much considered by the human raters in the context of the sentence generation task. The same applies to sequence similarity: For the creativity rating, it did not significantly matter whether a word in the sentence was spelled very similarly to the cue. This finding is still relevant when considering that the cue word only offered the initial letters as prompts for sentence generation to guide the response structure, but beyond that (unlike for many other divergent thinking tasks) implied no obvious relevance for response generation.
Some specific limitations of the present work need to be acknowledged. First, the metrics were specifically developed to assess relevant aspects of creative sentences in the sentence generation task. It thus is unclear which of them may generalize to other verbal creativity tasks and would also be of use in the assessment of longer and more complex texts. Furthermore, while the study was interested in an automated assessment, humans were still involved in response transcription and for general evaluations of response validity (i.e., checking whether sentences are meaningful). In future work, a fully automatic preprocessing pipeline could be easily envisioned, starting from the recording of participants’ responses (typing or speech-to-text) to checking their validity (e.g., using natural language processing algorithms for checking adherence to grammar rules) before further analysis. Second, while we considered a large set of text metrics, it is likely possible that they have not yet captured all relevant aspects of sentence-level creativity, and that further measures can be conceived that increase the explained total variance in response creativity.

Regarding the implementation of the metrics, we would also like to point out some limitations of our algorithm. Two of our metrics, i.e., context-specific word uniqueness and syntax uniqueness, are computed in relation to the sample, which might affect their reliability and needs to be considered when applying these metrics to other samples, particularly those of small sizes (cf. Forthmann, Paek, Dumas, Barbot, & Holling, 2020). The context score was based on Google Search results. While this allows the most up-to-date results in a vast variety of contexts (Cilibrasi & Vitanyi, 2007), the results might fluctuate and be difficult to replicate exactly at different points of time. Furthermore, many scores use a mean score for the entire sentence. The general word and combination infrequency scores, for example, could also yield a maximum score in order to test whether just one single uncommon word in a sentence would increase the creativity rating. Importantly, the semantic similarity score was based on a freely available, industrial-strength Python natural language processing module which offers a pre-trained model for vectorizing and computing the semantic similarity between words in German. The pre-trained model might not have been the ideal fit for our data set, since it was based on a large German news corpus which does not necessarily represent everyday language use. In this context it is unclear how findings would generalize to word embeddings from different text corpora and languages, which partly rely on larger corpora.

Despite these technical limitations, we conclude that the present work introduced a promising new approach to objectively measure the verbal creativity of sentences. Objective creativity scoring could be particularly helpful for the analysis of large-scale data or for providing immediate performance feedback such as in the study of dynamic aspects of creativity (e.g., Corazza, Agnoli, & Mastria, 2022), as well as for the study of cognitive and neural mechanisms associated with specific verbal qualities in task performance (e.g., Khalil & Moustafa, 2022). We identified several distinct aspects related to both novelty and effectiveness that jointly contributed to the holistic impression of response creativity, thereby offering a more differentiated view on relevant “ingredients” of verbal creativity. Together, these findings shed new light on relevant characteristics of creative ideas that are amenable to objective assessment, thereby supporting ongoing efforts toward improved automated scoring of divergent thinking tasks.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Austrian Science Fund (FWF): P29801. We would like to thank Tobias Maltenberger (HPI) and Ivan Ilic (HPI) for developing a first prototype version of the scoring algorithm.

ORCID

Theresa J. Weinstein
http://orcid.org/0000-0002-8486-1339

Simon Majed Ceh
http://orcid.org/0000-0003-0534-3697

Mathias Benedek
http://orcid.org/0000-0001-6258-4476

Data availability

https://doi.org/10.17605/OSF.IO/YX4TS

References

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. Creativity Research Journal, 26(2), 229–238. doi:10.1080/10400419.2014.901095

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 153–158. doi:10.1037/aca0000231

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology, 43(5), 997–1013. doi:10.1037/0022-3514.43.5.997
Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education, 36*(3), 312–334. doi:10.1080/01587919.2015.1081733

Barbot, B., Reiter-Palmon, R., Barbot, B., & Reiter-Palmon, R. (2019). Creativity assessment: Pitfalls, solutions, and standards [Special Issue]. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 131–132. doi:10.1037/aac0000251

Beatty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780. doi:10.3758/s13428-020-01453-w

Beatty, R. E., Johnson, D. R., Zeilten, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal, 1*, 1–16. doi:10.1080/10400419.2022.205720

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe’s Journal of Psychology, 12*(2), 210–220. doi:10.5964/ejop.v12i2.1127

Benedek, M., Fink, A., & Neubauer, A. C. (2006). Enhancement of ideational fluency by means of computer-based training. *Creativity Research Journal, 18*(3), 317–328. doi:10.1207/s15326934crj1803_7

Benedek, M., Nordtvedt, N., Jauk, E., Koschnieder, C., Pretsch, J., Kramer, G., & Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity, 21*, 75–84. doi:10.1016/j.tsc.2016.05.007

Ceh, S. M., Annerer-Walcher, S., Körner, C., Rominger, C., Kober, S. E., Fink, A., & Benedek, M. (2020). Neurophysiological indicators of internal attention: An electroencephalography–eye-tracking coregistration study. *Brain and Behavior, 10*(10). doi:10.1002/brb3.1790

Ceh, S. M., Annerer-Walcher, S., Koschutnig, K., Körner, C., Fink, A., & Benedek, M. (2021). Neurophysiological indicators of internal attention: An fMRI–eye-tracking coregistration study. *Cortex, 143*, 29–46. doi:10.1016/j.cortex.2021.07.005

Ceh, S. M., Edelmann, C., Hofer, G., & Benedek, M. (2022). Assessing raters: What factors predict discernment in novice creativity raters? *The Journal of Creative Behavior, 56*(1), 41–54. doi:10.1002/jocb.1515

Cilibrasi, R., & Vitanyi, P. M. B. (2007). The google similarity distance. *ArXiv:Cs/0412098. http://arxiv.org/abs/cs/0412098*

Corazza, G. E., Agnoli, S., & Mastria, S. (2022). The dynamic creativity framework: Theoretical and empirical investigations. *European Psychologist*. doi:10.1027/1076-9040/a000473

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensus assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 159–166. doi:10.1037/aac0000220

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, (Vol. 1, pp. 4171–4186). doi:10.48550/arxiv.1810.04805

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 35–40. doi:10.1037/a0038688

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. doi:10.1016/j.tsc.2021.100949

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts, 15*(4), 645–663. doi:10.1037/aac0000319

Dumas, D., & Runco, M. A. (2018). Objectively scoring divergent thinking tests for originality: A re-analysis and extension. *Creativity Research Journal, 30*(4), 466–468. doi:10.1004/j1.br.2017.10.013

Estes, Z., & Ward, T. B. (2002). The emergence of novel attributes in concept modification. *Creativity Research Journal, 14*(2), 149–156. doi:10.1207/S15326934CREJ1402_2

Explosion AI GmbH. (2022). Spacy industrial-strength natural language processing in python. https://spacy.io

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98*, 19–38. doi:10.1016/j.brat.2017.05.013

Form, S. (2019). Reaching Wuthering heights with brave new words: The influence of originality of words on the success of outstanding best-sellers. *The Journal of Creative Behavior, 53*(4), 508–518. doi:10.1080/10700223.230

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. doi:10.1016/j.tsc.2016.12.005

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior, 53*(4), 559–575. doi:10.1080/10700224.240

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology, 90*(3), e12325. doi:110.1111/bjep.12325

Fürst, G., Ghiisletta, P., & Lubart, T. (2017). An experimental study of the creative process in writing. *Psychology of Aesthetics, Creativity, and the Arts, 11*(2), 202–215. doi:10.1037/AC0000106

Glävekeu, V. (2015). Creativity as a sociocultural act. *Journal of Creative Behavior, 49*(3), 165–180. doi:10.1002/jcbb.94

Guilford, J. P. (1967). The nature of human intelligence. McGraw-Hill.

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition, 45*(2), 233–244. doi:10.3758/s13421-016-0659-y

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 144–156. doi:10.1037/aac0000125

Johnson, D. R., Kaufman, J. C., Baker, B. S., Barbot, B., Green, A. E., van Hells, J. . . . , Beatt, R. E. (2022). Extracting creativity from narratives using distributional semantic modeling. *Preprint*. doi:10.31234/osf.io/fmwyj

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion*
