A mad method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD)

Yuhua Yu[1], Roger E. Beaty[2], Boris Forthmann[3], Mark Beeman[1], John Henry Cruz[4], & Dan Johnson[5]

[1]Department of Psychology, Northwestern University

[2]Department of Psychology, Pennsylvania State University

[3]Institute of Psychology in Education, University of Münster

[4]College of Natural Sciences, University of Texas at Austin

[5]Department of Cognitive and Behavioral Science, Washington and Lee University

Author Note

R. E. B. was supported by NSF grant DRL-1920653.

All data, materials, and code (R and Python) for computation of automatic scoring methods, CFAs, and generating figures are available at https://osf.io/56xbq/.

Correspondence should be addressed to Yuhua Yu, Swift Hall 102, 2029 Sheridan Road Evanston, IL 60208. Email: yyu@u.northwestern.edu

Abstract

Creativity research often relies on human raters to judge the novelty of participants' responses on open-ended tasks, such as the Alternate Uses Task (AUT). Albeit useful, manual ratings are subjective and labor intensive. To address these limitations, researchers increasingly use automatic scoring methods based on a natural language processing technique for quantifying the semantic distance between words. However, many methodological choices remain open on how to obtain semantic distance scores for ideas, which can significantly impact reliability and validity. In this project, we propose a new semantic distance-based method, maximum associative distance (MAD), for assessing response novelty in AUT. Within a response, MAD uses the semantic distance of the word that is maximally remote from the prompt word to reflect response novelty. We compare the results from MAD with other competing semantic distance-based methods, including element-wise-multiplication—a commonly used compositional model—across three published datasets including a total of 447 participants. We found MAD to be more strongly correlated with human creativity ratings than the competing methods. In addition, MAD scores reliably predict external measures such as openness to experience. We further explored how idea elaboration affects the performance of various scoring methods and found that MAD is closely aligned with human raters in processing multi-word responses. The MAD method thus improves the psychometrics of semantic distance for automatic creativity assessment, and it provides clues about what human raters find creative about ideas.

Keywords: alternative uses task, semantic distance, novelty, automatic scoring

A mad method to assess idea novelty: Improving validity of automated scoring using maximum associative distance (MAD)

Researchers are increasingly using natural language processing (NLP) techniques to assist with automating creativity measurement. Perhaps the most notable application is to assess the originality of responses in open-ended divergent thinking tasks, such as the alternate uses task (AUT), using automatic scoring based on semantic distances (Acar & Runco, 2014; Beketayev & Runco, 2016; Dumas et al., 2020; Heinen & Johnson, 2018; Kenett, 2019; Prabhakaran et al., 2014). Automatic methods are more objective and cost-efficient compared to the traditional manual rating process. However, many choices must be made when formulating an automatic algorithm, and it remains open how those choices affect the validity of the scoring methods.

One of the key aspects of constructing an automatic algorithm for divergent thinking tasks, like the AUT, involves processing multiple words in a response. In this paper, we propose a new method, maximum-associative-distance (MAD). MAD uses the semantic distance of the word that is maximally remote from the prompt word as the score for the response. We compared MAD with other existing methods in terms of how well the scores align with human ratings, as well as with other measures capturing aspects of individual creativity (e.g., creative behavior).

Creativity Assessment and Semantic Distance

How to assess an individual's creative thinking capacity is a question of longstanding interest (Silvia et al., 2008; Wilson et al., 1953; see, for example, Reiter-Palmon et al., 2019, for a review). Briefly, researchers often present open-ended divergent thinking prompts and ask people to provide creative responses. For example, in the AUT, one of the most widely used divergent thinking tasks, people are presented with a common object (e.g., a box) and asked to think of as many creative uses for it as possible within a given time (Guilford, 1967). Creativity is often measured by the number of responses (fluency) or by the creative quality of responses (novelty/originality). Fluency is straightforward to quantify but is agnostic to the (creative) quality of the ideas. Novelty, as the primary facet of creativity, is traditionally considered by subjective scoring methods (Amabile, 1982; Hass et al., 2018; Silvia et al., 2008). For example, when rating divergent thinking task responses, a group of raters are briefly trained on the general criterion and asked to rate ideas on an ordered numerical scale (e.g., 1=*not at all creative*, 5=*very creative*). Disagreement can be reconciled to a certain degree via discussion, but the final scores inevitably rely on raters' subjective perception of creativity (Cseh & Jeffries, 2019; Mouchiroud & Lubart, 2010). Subjective scoring methods have shown evidence of convergent validity (Forthmann et al., 2017; Jauk et al., 2014), but there have been concerns with its subjectivity which can lead to low inter-rater agreement (Barbot, 2018) and challenges to replicability. For example, raters' judgments can be influenced by the cognitive load from scoring many responses (Forthmann et al., 2017) or by their own creativity and personality (Ceh et al., 2022).

Given the limitations of subjective scoring, automated approaches have enjoyed increased popularity among researchers (Acar & Runco, 2014; Dumas et al., 2020; Dumas & Runco, 2018; Green, 2016; Hass, 2017; Heinen & Johnson, 2018; Kenett, 2019; Prabhakaran et al., 2014; Zedelius et al., 2019). According to the classic associative theory (Mednick, 1962), creative thinking involves making connections between remotely associated concepts. Indeed, remoteness (in associative terms) has been considered by Wilson and colleagues (1953) as an indicator of originality. Remoteness has been also included in instructions for subjective ratings of divergent thinking tasks (Silvia et al., 2008) and researchers have argued that it is closely linked to semantic distance (e.g., Forthmann et al., 2019). The semantic distance between two words can be obtained via *semantic space*, a natural language processing (NLP) technique that creates a vectorized representation of words based on their meaning (embedding). Semantically similar words are embedded more closely to each other in a high-dimensional vector space. The semantic distance between two words is computed by a distance metric (e.g., 1- cosine of the angle) between the word vectors.

Initial evidence has shown the potential of semantic distance for measuring creative quality in a variety of creativity tasks including verb generation task (Heinen & Johnson, 2018; Prabhakaran et al., 2014), remote associates test (Beisemann et al., 2020; Smith et al., 2013), forward flow (Beaty et al., 2021; Gray et al., 2019), abstract figure naming (Sung et al., 2022), and AUT (Beaty & Johnson, 2021; Forster & Dunbar, 2009; Hass, 2017). Semantic distance based score in AUT is found to be positively correlated with measures of creative personality and achievement (Beaty & Johnson, 2021; Dumas et al., 2020), supporting the validity of automatic assessment (but see Harbison & Haarmann, 2014 for mixed evidence). Importantly, however, researchers have also identified limitations with some semantic distance-based methods. For example, Forthmann and colleagues (2019) demonstrated, with a mix of simulation and empirical analysis, that automatic scores can be biased according to the number of words in a response. How to construct an effective scoring method is indeed an active research area and there exists room for improvements on multiple fronts (Beaty et al., 2022).

Automatic Methods for Processing Multi-Word Responses

Semantic distance provides a promising building block for automatic scoring methods. It is unclear, however, how to apply the word-to-word distance to responses containing a varying number of words—a common feature of divergent thinking tasks, such as the AUT, where people show wide variability in how they express their ideas. Different choices around how to process multi-word responses lead to different scoring performances. The first step in the processing involves the decision of removing stop words, “meaningless” words in a phrase such as “a”, “to”. Prior works demonstrated that removing stop words can improve the validity of the automatic scores (Forthmann et al., 2019; Hass, 2017). However, there are no universally agreed-upon rules for identifying stop words. Predefined lists used in different NLP tools can differ from 7-12 to 200-300 words (Manning & Prabhakar, 2008). Although removing stop words seems like a minor step in the processing pipeline, the choice of the stop word list may have a significant impact on the final score due to the relatively short length of all AUT responses (Forthmann et al., 2019).

After stop words removal, another critical step involves a word “composition model” to combine multiple words into a single vector. Different composition models (additive or multiplicative) can lead to different outcomes (Mitchell & Lapata, 2008, 2010). Beaty and Johnson (2021) found the elementwise multiplication (EWM) model, multiplying the word vectors along each dimension, is more closely aligned with human ratings than the additive model. A composition model of all word vectors can also put different weights on the constituent words, because some words may carry more information than others. Mitchell & Lapata (2010) proposed a syntax-based composition. When composing an additive vector to represent a phrase, putting more weight on the *head* (the word that determines the syntactic category of that phrase) provides a more reliable similarity rating than a simple additive composition.

Responses in the AUT are often agrammatical and have a variety of syntactic types, so designing a general syntactic-based composition rule is challenging. In contrast, here we propose a novel approach that uses the semantic distance of each word to the prompt as the weighting. Instead of a composition model (additive or multiplicative) that combines word vectors, an alternative approach can directly aggregate semantic distances of each constituent (to the prompt word) via a function*, f*. For example, in response to the prompt item “box”, a score for “breakdance platform” can be computed as *f*(*sdis*(box, breakdance), *sdis*(box, platform)), where *sdis*() denotes the semantic distance. If the *aggregating function* is *maximum*, we obtain a score, termed “maximum-associative-distance” (or MAD), that relies only on the most distant word in a response[1]. The MAD score for “breakdance platform” would be equal to the semantic distance between “breakdance” and “box”, because “breakdance” is more remotely associated with “box” than “platform” with “box”. MAD pushes further the idea of removing “meaningless” stop words to the extent that it only retains the most “significant” word in the sense of semantic remoteness. The MAD approach is motivated by observations from the human rating process: the creative quality of a response is often captured by a keyword (Forthmann et al., 2019, Table 3).

The *aggregating function* approach allows the flexibility of biasing towards more distant words. We will also consider an alternative to *maximum* and let the aggregating function be the weighted average of the semantic distances, where the weight is proportional to the (reverse) rank order of the remoteness of each constituent. The most remote constituent is assigned the highest weight. In this case, we obtain a score termed rank-ordered-weighted-average (ROWA). ROWA can be considered as a method in-between MAD, which adopts “a winner-take-all” approach, and a simple average of semantic distances (i.e., the additive composition method, the current “gold standard” in automated creativity assessment).

Elaboration and its Role in Creative Quality Assessment

The multi-word processing choice is closely related to the concept of elaboration, the degree to which an idea is detailed, embellished, or explained (Guilford, 1967; Torrance, 1995).

[1] Using *average* as the *aggregating function* is equivalent to the additive composition vector model due to the property of the cosine distance.

Elaboration has long been recognized as an important factor in creativity (Torrance, 1959, 1969). Participants' ability to produce elaborative responses predicts their creative achievement (Runco et al., 2010; Torrance, 1969) and trait openness (Dumas et al., 2021). For AUT responses, prior work generally found that elaboration negatively correlated with fluency but positively correlated with human rated creative quality (Forthmann et al., 2017; Reiter-Palmon & Arreola, 2015; but see Dumas et al., 2021).

The elaboration of an AUT response can be measured by the number of words in a response (Dumas et al., 2021; Forthmann et al., 2019). The creativity rating is expected to be higher for more elaborate ideas based on theoretical consideration. At least one aspect of creativity arises from combinatorial processes (Simonton, 2010, 2021), therefore the creative quality of an idea is likely higher (or no less) than any of its subcomponents. In other words, if a response A is made of two components $A = A_1 \cup A_2$, it is reasonable to expect $score(A) \geq score(A_1)$ and $score(A) \geq score(A_2)$, which means $score(A) \geq \max{(score(A_1), score(A_2))}$. More elaborative responses are likely to contain more component ideas leading to higher creativity ratings. The positive relationship with elaboration has been empirically observed in human creativity ratings (Beaty & Johnson, 2021; Forthmann et al., 2018; Maio et al., 2020; Runco et al., 2010).

Interestingly, different semantic distance-based automatic methods have different sensitivity to elaboration. Forthmann and colleagues (2019) found additive composition models produce scores that are negatively correlated with elaboration (Dumas et al., 2021; Forster & Dunbar, 2009; Forthmann, Holling, Zandi, et al., 2017; Maio et al., 2020), and recommended measures to correct the "elaboration-bias". On the other hand, Beaty and Johnson (2021) found that elementwise multiplication produced scores positively correlated with elaboration. Scores from the MAD method are expected to increase with elaboration because *maximum* is used to aggregate the semantic distance of constituents. Therefore, longer responses are more likely to receive a higher score. Furthermore, considering that the elaborative responses are more likely to be rated more creative by human raters, we speculated that a positive association with elaboration can support the validity of an automatic scoring method.

The Present Research

To improve validity in semantic distance-based automatic assessment of idea novelty, we propose and test a new method, maximum-associative-distance (MAD). MAD measures response novelty in the AUT by the maximum of semantic distances between the prompt word and each constituent word. Along with the idea of biasing toward more remote constituents, we also considered a related method, ROWA, the weighted average of the semantic distances based on the rank order of the remoteness from the prompt. We compare MAD and ROWA with the more widely used composition method that uses elementwise multiplication (EWM). We then examined how the proposed algorithms correlated with human ratings on the AUT and other creativity measures (e.g., creative behavior). We also analyzed how elaboration affects different scoring methods to understand the performance difference. We applied the analyses to three published datasets that were also analyzed in Beaty and Johnson (2021).

Study 1

In this study, we compared the validity of the newly proposed automatic methods, MAD and ROWA, to a previously established method that uses elementwise multiplication (EWM; Beaty & Johnson, 2021). We included two versions of EWM, one with a long stop word list filter (m_l) and one with a short stop word list filter (m_s; identical to Beaty & Johnson, 2021). We aim to evaluate two key choices in constructing an automatic scoring algorithm: the multi-word processing (a composition vector model versus an aggregating function approach) and the stop word list.

We reanalyzed AUT responses from a published study (Study 1, Beaty & Johnson, 2021) and compared the performance of four competing methods in predicting human creativity ratings. For validity, we examined the extent to which the creativity score derived from each method related to several other measures of creativity based on self-report (e.g., creative behavior). Previous research found that semantic distances correlated with both human ratings (Beaty & Johnson, 2021; Heinen & Johnson, 2018) and a range of other creativity measures (Prabhakaran et al., 2014). In light of past work (Forthmann et al., 2019; Mitchell & Lapata, 2010), we expected the proposed methods that shift the scoring towards more distant constituents would show higher validity.

Methods

Participants

Participants were recruited as part of a larger project on individual differences in creativity (see Beaty et al., 2018 for details on this dataset). Briefly, data from 172 participants (123 females, age = 22.63±6.29 years) are included in the following analysis after excluding 14 participants who did not complete all divergent thinking assessments and 1 participant whose data was identified as an outlier (Cooks Distance >10).

Procedure

Participants completed a battery of tasks and questionnaires. Cognitive assessments were administrated in a laboratory setting using MediaLab; questionnaires were administered both in MediaLab and via Qualtrics.

AUT Procedure. Participants completed two trials of the AUT (box and rope) and had three minutes for continuous idea generation for each prompt. As in prior work (Nusbaum et al., 2014), participants were instructed to "think creatively" while coming up with uses for the objects. The instruction explicitly emphasized quality over quantity, as well as novelty over usefulness. Responses were scored for creative quality using the subjective scoring method (Benedek et al., 2013; Silvia et al., 2008). Four raters scored responses using a 1 (not at all creative) to 5 (very creative) scale. The subjective ratings showed good inter-rater reliability (ICC (*C*, 4) = .84). Task instructions (https://osf.io/vky36/) and rater guidelines (https://osf.io/vie7s/) are available via OSF.

Creative Behavior. A battery of questionnaires was administrated to measure two facets of creative behavior: 1) creative activities (i.e., hobbies) and 2) creative achievements. The Biographical Inventory of Creative Behavior (BICB; Batey, 2007) assesses creative activities with a list of 34 creative activities (e.g., making a website). Participants indicate if they have participated in each activity within the past year. The Inventory of Creative Activity and Achievements (ICAA; Diedrich et al., 2018) includes two subscales that capture both creative activities and high-level accomplishments across eight domains of the arts and sciences. The Creative Achievement Questionnaire (CAQ; Carson, Peterson, & Higgins, 2005) assesses publicly recognized creative achievements across ten creative domains.

Creative Self-Concept. The Short Scale of Creative Self (SSCS; Karwowski, 2014) captures two components of creative self-concept and contains 11 items. The creative self-efficacy (CSE) subscale measures the extent to which people perceive themselves as capable of solving creative challenges. The creative personality identity (CPI) subscale measures the extent to which creativity is a defining feature of the self-concept.

Creative Metaphor. Two creative metaphor production prompts (Beaty & Silvia, 2013) were included as a further test of validation with a cognitive assessment of creativity. For example, participants were asked, e.g., “think of the most boring high school or college class that you’ve ever had. What was it like to sit through?”. Participants were asked to produce novel metaphors which were rated by four raters using a 1 (*not at all creative*) to 5 (*very creative*) scale.

Personality. The 240 item NEO PI-R was administrated to assess the five major factors of personality (McCrae et al., 2005): neuroticism, extraversion, openness to experience, agreeableness, and consciousness. Each personality factor is an average of six facet-level subscales. Participants were presented with a series of statements and indicated their level of agreement using a five-point Likert scale.

Semantic Spaces

To evaluate the novelty of a response, we obtained the semantic distances between the AUT prompt item (e.g., box) and participants’ responses by computing the distance between the vectorized word representations in a high dimensional vector space (semantic space). Different semantic spaces or embedding maps can be constructed by applying different algorithms (neural network, latent semantic analyses, etc.) to a large corpus of natural language text such as books or online encyclopedias. Prior work used a variety of semantic spaces for automatic scoring and evidence suggested each space has idiosyncratic strengths and weaknesses in predicting human performance (Dumas et al., 2020; Mandera et al., 2017). Therefore, incorporating multiple semantic spaces may contribute to a robust scoring method. Following prior work (Johnson & Beaty, 2021), we used five semantic spaces that have demonstrated certain validity in related tasks for our analyses.

We describe the spaces briefly here: 1) Touchstone Applied Science Associates (TASA) computes word co-occurrences within a text corpus (37,651 documents, middle and high school textbooks and literary words, 92,393 different words), followed by a singular value decomposition to reduce the dimension to 300 (Günther et al., 2014; Prabhakaran et al., 2014); 2) global vectors (GloVe; Pennington, Socher, & Manning, 2014) model is trained on ~6 billion tokens and uses weighted least squares to extract global information across a concatenation of the 2014 Wikipedia dump and news publications from 2009-2010. The remaining three embeddings are predictive models that use the neural network architecture to learn word associations from a large corpus of text. Those corpora are 3) concatenation of the ukwac web crawling corpus (~ 2 billion words) and the subtitle corpus (~385 million words; 300 dimensions, most frequent 150,000 words; Mandera, Keuleers, & Brysbaert, 2017); 4) the subtitle corpus only, and 5) concatenation of the British National Corpus (~2 billion words), ukwac corpus, and the 2009 Wikipedia dump (~800 million tokens; 400 dimensions, most frequent 300,000 words; Baroni, Dinu, & Kruszewski, 2014).

Automatic Scoring Methods

With each semantic space described above, we used an automatic method to generate scores for responses containing a varying number of words. We considered four methods that fall into two classes. The first class applies elementwise multiplication (m_l and m_s) along each dimension in the vector space to combine word vectors into a single vector. Then the semantic distance between the combined vector and the prompt word vector is used as a score for novelty.

In m_l, we first filtered out "meaningless" words (like "the", "to") in a response using stop words from the Python library, spaCy (spacy.lang.en.stop_words)[2]. We augmented the word

[2] https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py.

list with “thing, things, use” as they do not convey concrete meaning in the AUT context. The stop words list contains 329 words. We then applied elementwise multiplication to combine the remaining words into a single vector.

The method m_s is identical to m_l except that fewer stop words are filtered out in the m_l model. Therefore, the combined vector may be generated by more words. The stop word list comes from the tm R package (https://rdrr.io/rforge/tm/man/stopwords.html) and contains 174 words. This algorithm is the same as the multiplicative method analyzed in (Beaty & Johnson, 2021).

In the newly designed second class of the methods (MAD and ROWA), we first computed the semantic distance for each constituent response word relative to the prompt word. Instead of generating a combined vector, we used a *function* to aggregate the distances to obtain a score. In MAD (maximum-associative-distance), the function takes the maximum of all semantic distances. In ROWA (rank-ordered-weighted-average), we first assigned each constituent word a rank score (1, 2, …, n) from the lowest distance to the highest distance (from the prompt word). We then computed the average semantic distance weighted by the rank score. Thus, the most distant word gets the highest weight. Take the two-word response “breakdance platform” as an example. The two constituent words are ranked by their distance to the prompt word “box”: 1. platform (the closest) and 2. breakdance. The ROWA score is computed as (1**sdis*(box, platform)+2**sdis*(box, breakdance))/(1+2).

MAD and ROWA are similar in that they are both biased toward the more distant words, with MAD taking the more extreme “winner-takes-all” approach. Both MAD and ROWA filter out the long stop word list as the m_l method. Code examples for MAD and ROWA are available via OSF (https://osf.io/56xbq/).

Analysis Plan

We employed confirmatory factor analysis (CFA) in order to minimize item-specific and rater-specific variance in method comparisons (Beaty & Johnson, 2021; Kline, 2015). Each automatic method (i.e., MAD, ROWA, m_l, m_s) generates a single score for each response from each item. Scores were averaged across responses by a participant to get a single score for each participant with each method and item. Then, CFA with maximum likelihood estimation (l*avaan* R package, Rosseel, 2012) was used to generate a factor score from a second-order latent variable for each participant and method. The same model specification was used for each automatic method and for human creativity ratings[3]. We obtained factor scores for each method by lavPredict() function in lavaan. To ensure the soundness of this approach, we estimated factor determinacy indices (FDI) for all factor scores (i.e., an estimate of the correlation between factor scores and their unknown true values). An FDI > .80 suggests sufficient quality of the factor scores be used for research purposes (Ferrando & Lorenzo-Seva, 2018). The CFA code is available via OSF (https://osf.io/56xbq/).

The correlation between the factor scores from each method and the human rating factor was compared using the *cocor* package in R (Diedenhofen & Musch, 2015). Steiger's (1980) test was used to compute *p*-values and Zou's (2007) approach was used to generate 95% confidence intervals (CIs) around each correlation coefficient.

For evaluating convergent validity, we probed the correlation between the factor score from each AUT scoring method with each of the external measures. In the case of the creative

[3] We attempted to fit a CFA with the same model specification as in Beaty and Johnson (2021) to AUT human creative ratings and automatic scores derived from five semantic spaces simultaneously, but improper solutions were found for some models (e.g., latent variable correlations exceeding r = 1.00), possibly due to model misspecification. Therefore, we opted for a more pragmatic strategy that allows evaluating the validity of the proposed method as detailed in this section.

metaphor task, we computed the factor score from a CFA combining two prompts, four raters per prompt, for each participant (Beaty & Johnson, 2021).

Results

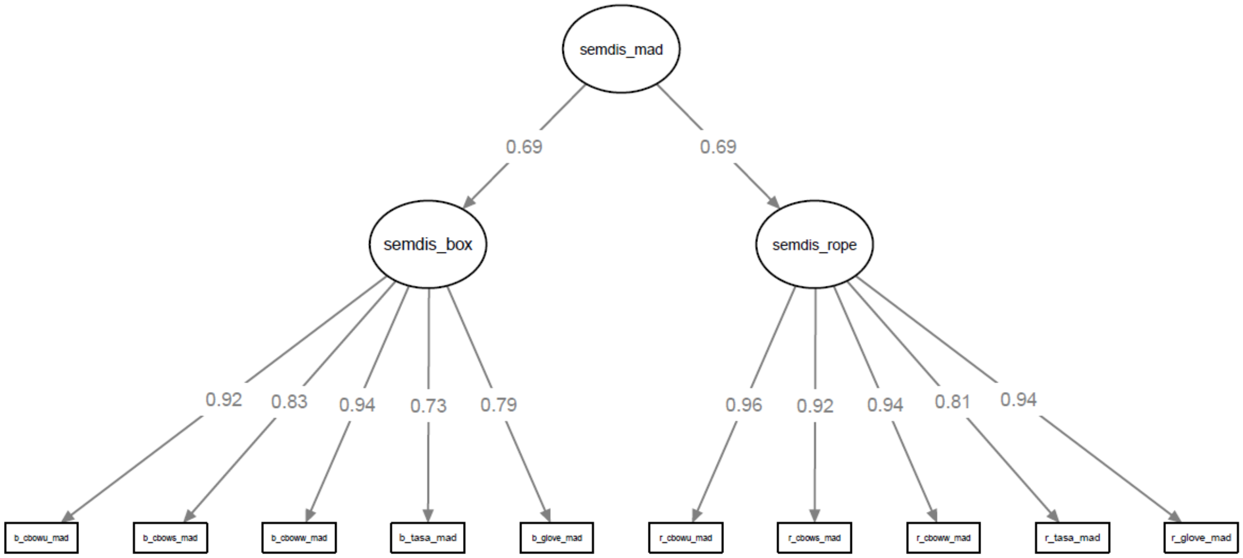
MAD Outperforms Competing Methods in Predicting Human Ratings

We began by conducting confirmatory factor analyses to assess the factor score correlations between human raters and each of the four automatic methods. Figure 1 depicts the factor model for the MAD method (the same model specification was used for each automatic method and for human creativity ratings). Each factor model had a good to excellent model fit in the context of high standardized loadings (i.e., a strong measurement model; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; McNeish, An, & Hancock, 2018; McNeish & Wolf, 2021). In addition, factor determinacy indices for factor scores exceeded the minimum recommended value for research application (i.e., .80; Ferrando & Lorenzo-Seva, 2018). For example, for the human raters, CFI 0.99; RMSEA[4] 0.01; SRMR 0.03; and for MAD, CFI 0.98; RMSEA 0.08; SRMR 0.05. See Table A1 for the complete reporting of model fit indices.

[4] RMSEA can be low when the model has high standardized loadings or low degree of freedom (McNeish et al., 2018) .

Figure 1

CFA for MAD Scores

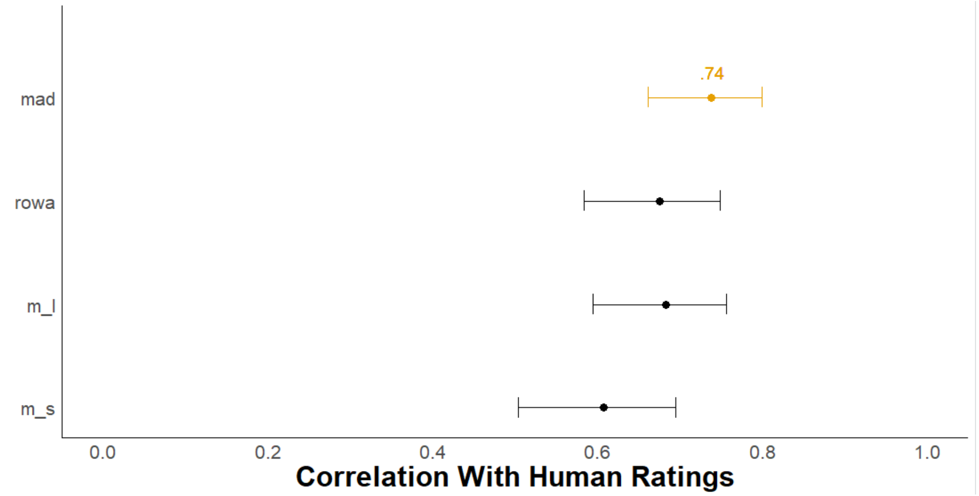


Notes. N=171. semdis_mad = second-order latent variable from two item-specific latent variables (i.e., box and rope); semdis_box = latent variable for the box item from five observed variables that are the MAD scores from each semantic model; semdis_rope = latent variable for the rope item from five observed variables that are the MAD scores from each semantic model.

The factor score extracted from the proposed method, MAD, has the highest correlation with the human rating score (*r* = 0.74 95% CI [.66, .80], see Figure 2). The elementwise multiplication method that filters out a short stop word list, m_s, has the lowest correlation with the human factor (*r* = 0.61, 95% CI [.50, .69]). The relationships between all latent variables were shown in Fig 3, including the univariate frequency distributions on the diagonal, bi-variable scatterplots in the space below the diagonal, and the Pearson correlation coefficients in the space above the diagonal.

Figure 2

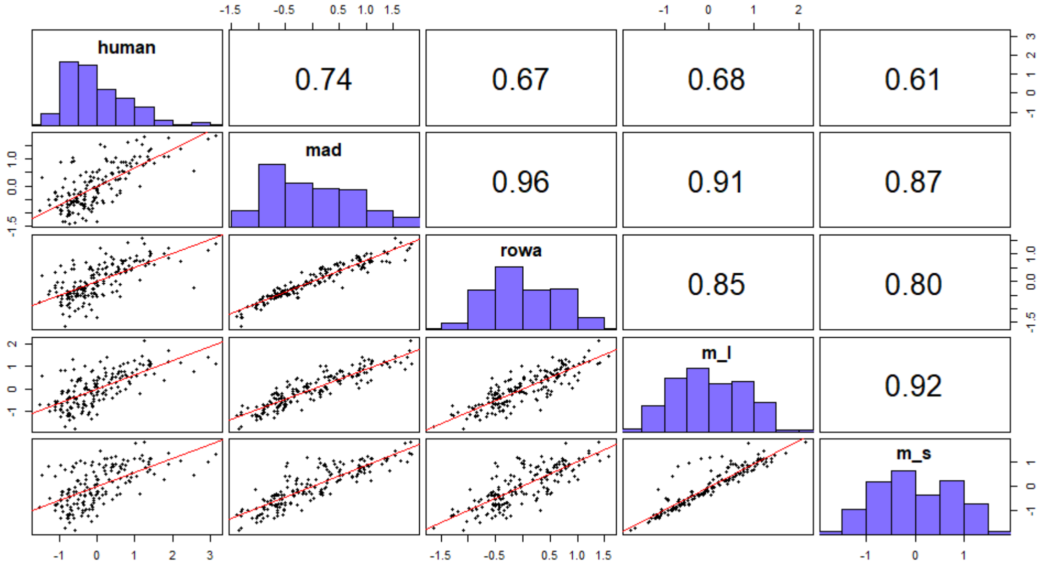
Correlation of Automatic Methods with Human Creativity Ratings



Notes. Each dot represents the correlation between human creativity ratings and an automatic scoring method, along with 95% Confidence Intervals. MAD = maximum associative distance, ROWA = rank order weighted average, m_l = multiplicative model removing long stopword list, m_s = multiplicative model removing short stopword list and model used in Beaty and Johnson (2021).

Figure 2

Pairwise Relationship Between all Variables



Notes. Scatterplots are depicted in the space below the diagonal, the Pearson correlation coefficients in the space above the diagonal, and the univariate frequency distributions on the diagonal (generated with the psych R package, Revelle, 2021). Column names are on the diagonal and the relationship between variables is the intersection of two columns.

To directly compare these automatic methods in terms of their correlation with the human ratings, we conducted correlation comparisons with 95%-CIs around the difference of correlations (Table 2). MAD is significantly more strongly correlated with human creativity ratings than all competing methods, suggesting that the most semantically distant word in a response—removing all other words—explains the most variance in human ratings. Furthermore, m_l has a higher correlation than m_s, indicating that removing a longer stop word list improves the validity of the scoring method.

Table 1

Comparisons Between Compositional Models in Correlation with Human Creativity Ratings

Model	MAD	ROWA	m_l
1. MAD			
2. ROWA	.74 - .67 *z* = 4.59, *p* < .001 ***.07 [.041, .112]		
3. m_l	.74 - .68 *z* = 2.67, *p* = .008 **.06 [.016, .112]	.67 - .68 *z* = -0.33, *p* = .740 -.01 [-.072, .051]	
4. m_s	.74 - .61 *z* = 4.66, *p* < .001 ***.13 [.076, .198]	.67 - .61 *z* = 2.25, *p* = .025 *.06 [.011, .156]	.68 - .61 *z* = 3.02, *p* = .003 **.07 [.025, .124]

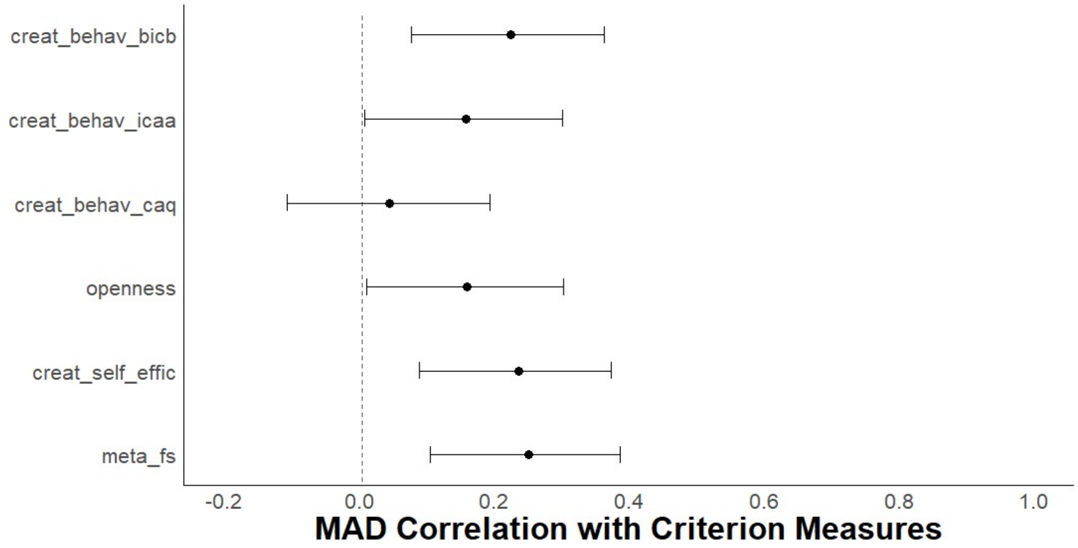
Notes. Top values are the two correlation coefficients being compared, where column vs. row models are compared, with column model as first number. Middle values are the Steiger's *z*-score and p-value for the comparison between the correlations. Bottom values are the correlation difference score and the values in square brackets indicate the 95% confidence interval for each correlation difference. *p < .05. **p < .01. ***p < .001.

Validation with External Measures

Having found that MAD outperforms other competing automatic methods in predicting human creativity ratings, we turned to further validate MAD with a range of external creativity measures, including cognition (novel metaphor), behavior (creative achievement), and self-report (creative efficacy). Using the same factor scores as in the previous section, we found MAD scores significantly correlate with most of the external measures across individuals (see Figure 4). There is a reliable correlation between MAD scores and the creative behavior measure, BICB (*r* =.23, 95% CI [.08, .36]) but not with the creative achievement, CAQ. Openness to experience, the personality factor most commonly associated with creativity (Kaufman et al., 2016; McCrae & Ingraham, 1987), also showed a significant positive correlation with MAD scores (*r* =.16, 95% CI [.01, .30]).

Figure 3

MAD Correlation with Criterion Measures of Creativity

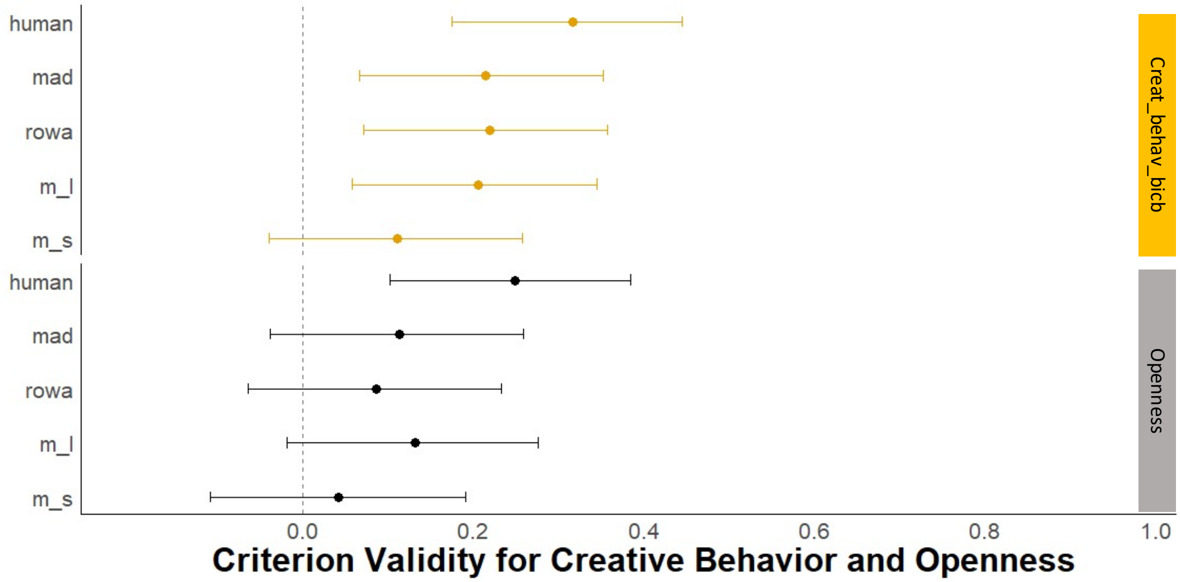


Notes. Each dot represents the correlation between human creativity ratings and each compositional semantic model, along with 95% Confidence Intervals. creat_behav_bicb = creative behavior from the Biographical Inventory of Creative Behaviors, creat_behav_icaa = creative behavior from the Inventory of Creative Activities and Achievements, creat_behav_caq = Creativity Achievements Questionnaire, openness = openness personality trait, creat_self_effic=creative self-efficacy subscale from the Short Scale of Creative self, meta_fs = metaphor task factor scores

To compare the external validity of different automatic scoring methods, we focused on two of the most used measures: the creative behavior scale (BICB) and the openness personality factor. Human creativity ratings on AUT have the highest correlation with both scales. Among the automatic methods, MAD, ROWA, and m_l all correlate moderately with the external measures, while m_s has the lowest correlation (Figure 5).

Figure 4

Comparison of Automatic Methods in Correlation with Creative Behavior (top) and Openness (bottom)



Notes. Each dot represents the correlation between human creativity ratings and each compositional semantic model, along with 95% Confidence Intervals. creat_behav_bicb = creative behavior from the Biographical Inventory of Creative Behaviors, openness = openness personality trait.

Study 2

Study 1 provided evidence for the validity of the proposed automatic AUT scoring method, MAD. We found the latent variable extracted from MAD scores has the highest correlation with the human rating factor. Both MAD and ROWA, which shift scoring towards more distant constituent words, are more aligned with human creative ratings than the EWM composition models. The method that removes more stop words (m_l) performs better than those removing fewer (m_s). Furthermore, MAD scores were reliably correlated with external measures including creative behavior (BICB) and trait openness. In Study 2, we aimed to replicate those findings using the same AUT items from another existing dataset (Study 2; Beaty & Johnson, 2021). We computed and analyzed the semantic distance scores following the same procedures as in Study 1. We also applied the same CFA to the human creativity ratings from the original study to compare the factor correlations.

Participants

Data for this study were reanalyzed from Beaty and Johnson (2021). The final sample included 142 adults (99 females, age = 19.22±3.07 years).

Procedure

Participants completed the AUT and self-report measures related to creativity. All measures were administered on laboratory computers running MediaLab.

AUT procedure. Participants completed two AUT items (box and rope), and they were asked to "think creatively" (as in Study 1). Participants were given three minutes to type their responses. AUT responses were again scored using the subjective scoring method (Silvia et al., 2008) on a scale of 1 to 5 by three raters. The subjective ratings showed acceptable inter-rater reliability (ICC (C, 3) = .72).

Creative behavior. Participants completed two of the same measures of creative behavior from Study 1: BICB (for assessing creative behaviors) and CAQ (for assessing creative achievement).

Openness to experience. Personality was measured using HEXACO (Lee & Ashton, 2004), which assesses four facets of openness to experience: aesthetic appreciation, unconventionality, intuitiveness, and creativity. Participants responded to each item using a five-point scale (1 = *strongly disagree*, 5 = *strongly agree*).

Results

MAD is More Aligned with Human Ratings than other Competing Methods

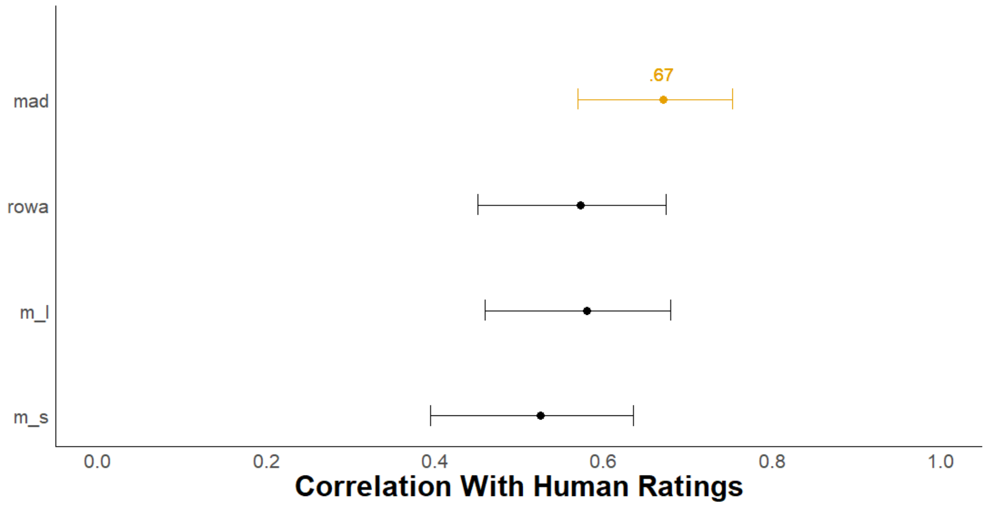
We began by conducting confirmatory factor analyses to assess the factor score correlations between human raters and each of the four automatic methods using the same model specification as in Study 1. Similar to Study 1, each factor model had a good to excellent model fit and factor determinacy indices exceeded the minimum recommended value (see Table A2). For example, for the human raters, CFI 0.95; RMSEA 0.15; SRMR 0.05; and for MAD, CFI 0.96; RMSEA 0.11; SRMR 0.03.

Consistent with Study 1, the factor score extracted from the MAD scores has the highest correlation with the human rating score (r=0.67 95% CI [.57, .75]; see Figure 6). The elementwise multiplication method that filters out a short stop word list, m_s, has the lowest correlation with the human factor (*r* = 0.53, 95% CI [.39, .64]).

To directly compare the automatic methods in terms of their correlation with the human rater factor, we conducted correlation comparisons with 95% around the difference of correlations (Table 2). Replicating Study 1, MAD is more strongly correlated with human creativity ratings than all competing methods; m_l has a higher correlation than m_s (although not reaching a 5% significant level).

Figure 5

Latent Factor Correlation with Human Creativity Ratings



Notes. Each dot represents the correlation between human creativity ratings and each automatic scoring method, along with 95% Confidence Intervals. MAD = maximum associative distance, ROWA = rank order weighted average, m_l = multiplicative model removing long stopword list, m_s = multiplicative model removing short stopword.

Table 2

Compositional Model Comparisons in Correlation with Human Creativity Ratings (Study 2)

Model	MAD	ROWA	m_l
MAD			
ROWA	.67 - .57 *z* = 4.40, *p* < .001 ***.10 [.057, .159]		
m_l	.67 - .58 *z* = 3.10, *p* = .002 **.09 [.034, .158]	.57 - .58 *z* = -0.26, *p* = .795 -.01 [-.089, .068]	
m_s	.67 - .53 *z* = 3.66, *p* < .001 ***.14 [.066, .227]	.57 - .53 *z* = 0.87, *p* = .383 .04 [-.051, .133]	.58 - .53 *z* = 1.69, *p* = .091 .05 [-.010, .115]

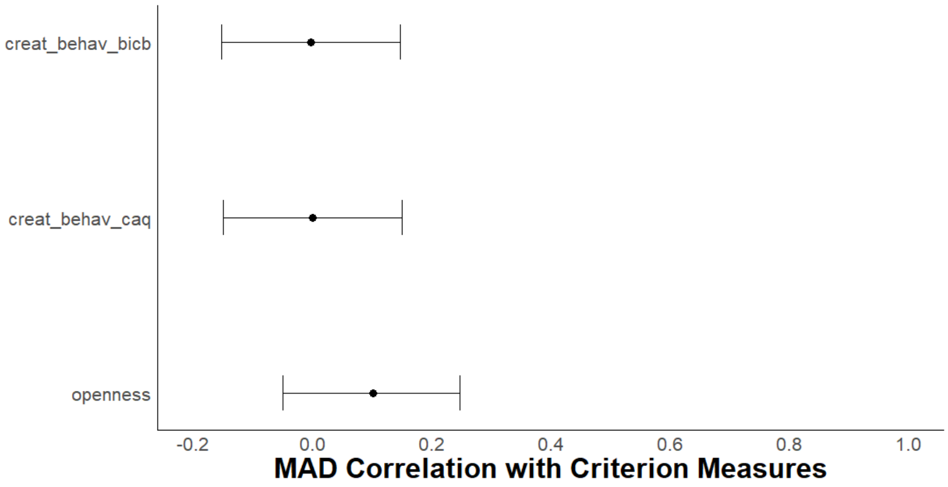
Notes. Top values are the two correlation coefficients being compared, where column vs. row models are compared, with column model as the first number. Middle values are the Steiger's *z*-score and p-value for the comparison between the correlations. The bottom values are the correlation difference score and the values in square brackets indicate the 95% confidence interval for each correlation difference.

Validation with External Measures

The external validation analysis assessed whether MAD scores relate to creative behavior and openness to experience (Figure 7). Creative behavior (BICB) and creative achievement (CAQ) did not predict MAD scores, consistent with findings from Beaty and Johnson (2021) using the same dataset. Regarding openness, there was a small correlation with MAD scores (*r* = .10, 95% CI [-.06, .26]).

Figure 6

MAD Correlation with Criterion Measures of Creativity



Notes. Each dot represents the correlation between human creativity ratings and each compositional semantic model, along with 95% Confidence Intervals. creat_behav_bicb = creative behavior from the Biographical Inventory of Creative Behaviors, creat_behav_caq = creative behavior from the Creativity Achievement Questionnaire, openness = openness to experience personality

Study 3

Study 2 replicated Study 1: the latent variable extracted from MAD scores has the highest correlation with human creativity ratings compared to other competing automatic methods. The external validation analysis, however, yielded mixed results. Openness was moderately associated with MAD scores, but creative behavior (BICB) did not predict MAD scores — similar to the results of Beaty and Johnson (2021) who used the same dataset and the multiplicative composition approach to computing semantic distance. In Study 3, we sought to replicate and extend findings from the first two studies. To this end, we reanalyzed another published dataset (Study 3; Beaty & Johnson 2021) that used another common AUT item (brick), as well as several measures of fluid intelligence—a cognitive ability that consistently correlates with divergent thinking (Beaty & Silvia, 2012). We applied the same analytical approach as in Study 1 and 2.

Participants

Data for this study were reanalyzed from previously published work (Beaty & Johnson, 2021). The final sample included 133 adults (92 females, age = 19.60±3.20 years).

Procedure

AUT procedure. Participants completed an extended AUT with one item, brick. They were given 10 min to continually generate uses for the item. The task duration was considerably longer than previous studies due to the temporal focus of the original study (i.e., analyzing response originality over time). The subjective ratings from three raters showed acceptable to good inter-rater reliability (ICC (*C*, 3) = .78).

Personality. The NEO PI-R was administrated to assess the five major factors of personality (as in Study 1).

Creative metaphor. Participants completed the same creative metaphor prompts from Study 1 (see Silvia & Beaty, 2012). Metaphor responses were scored for creative quality using the subjective scoring method (Silvia et al., 2008), and CFA was applied to extract the factor scores.

Fluid intelligence. Six nonverbal measures of fluid intelligence were administrated: 1) a short version of the Ravens Advanced Progressive Matrices (18 items); 2) a paper folder task (10 items; Ekstrom et al., 1976); 3) a letter sets task (16 items; Ekstrom et al., 1976); 4) the matrices task from Cattell Culture Fair Intelligence test (CFIT; 13 items; Cattell & Cattel, 1961, 2008); 5) the series task from CFIT (13 items); and 6) a number series task (15 items; Thurstone, 1938). We obtained the fluid intelligence factor from a CFA model that was fit to the scores of the six tasks.

Results

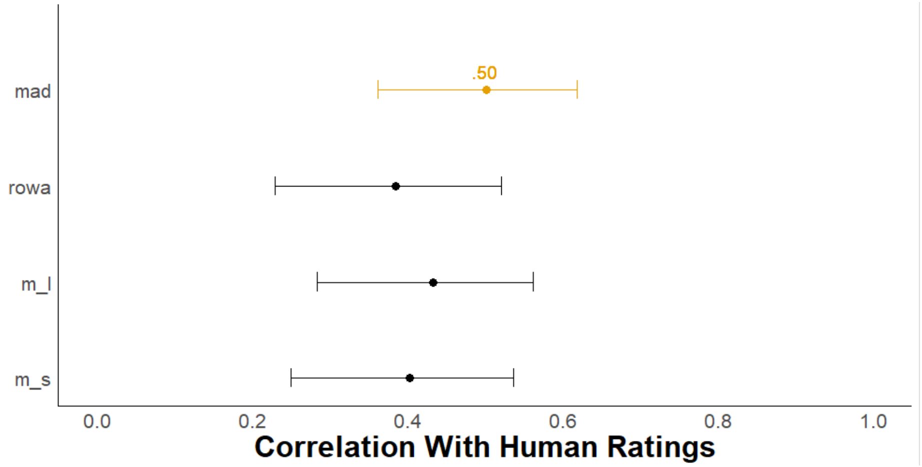
MAD is More Aligned with Human Ratings than other Competing Methods

As in the two previous studies, we conducted confirmatory factor analyses to assess the factor score correlations between human raters and each of the four automatic methods. Again, each factor model showed an excellent fit (Table A3).

Consistent with the previous studies, MAD scores have the highest correlation with the human rating score (*r* = 0.50, 95% CI [.37, .62], Figure 8).

Figure 7

Latent Factor Correlation with Human Creativity Ratings



Notes. Each dot represents the correlation between human creativity ratings and each automatic method, along with 95% Confidence Intervals. mad = maximum associative distance, rowa = rank order weighted average, m_l = multiplicative model removing long stopword list, m_s = multiplicative model removing short stopword list.

Directly comparing the automatic methods in terms of their correlation with the human rater factor (Table 5), MAD was more strongly correlated with human creativity ratings than ROWA and m_s, but its advantage over m_l, the EWM composition with long stop word list removal, did not reach significance.

Table 3

Compositional Model Comparisons in Correlation with Human Creativity Ratings (Study 3)

Model	MAD	ROWA	m_l
MAD			
ROWA	.50 - .36 $z = 4.20$, $p < .001$ ***.14 [.076, .215]		
m_l	.50 - .43 $z = 1.52$, $p = .128$ .07 [-.020, .164]	.36 - .43 $z = -1.057$, $p = .291$ -.07 [-.201, .060]	
m_s	.50 - .40 $z = 2.00$, $p = .045$ *.10 [.002, .202]	.36 - .40 $z = -0.58$, $p = .560$ -.04 [-.175, .095]	.43 - .40 $z = 0.772$, $p = .440$ .03 [-.047, .109]

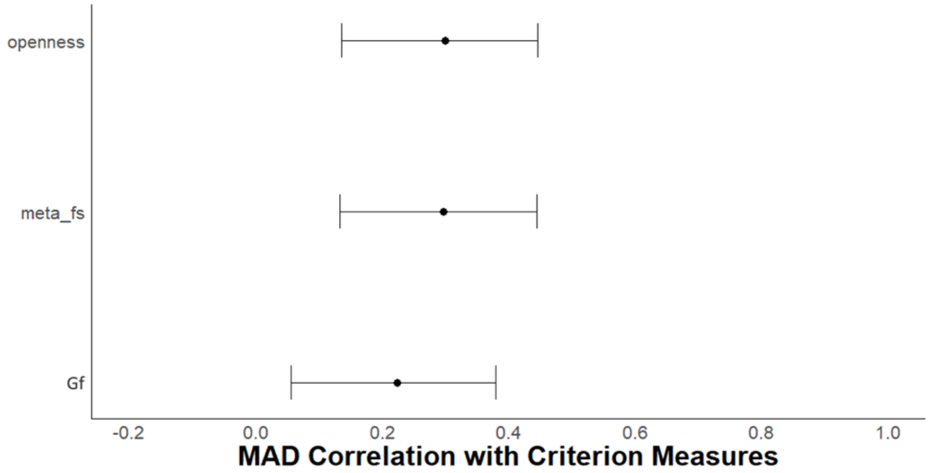
Notes. Top values are the two correlation coefficients being compared, where column vs. row models are compared, with column model as the first number. Middle values are the Steiger's z-score and p-value for the comparison between the correlations. The bottom values are the correlation difference score and the values in square brackets indicate the 95% confidence interval for each correlation difference. *$p < .05$. **$p < .01$. ***$p < .001$.

Validation with External Measures

The external validation analysis assessed whether MAD scores relate to openness to experience, fluid intelligence, and creative metaphor scores (Figure 9). MAD scores were significantly correlated with each of the external measures. In particular, openness reliably predicted MAD scores ($r = .30$, 95% CI [.14, .45]), consistent with findings from previous studies.

Figure 8

MAD Correlation with Criterion Measures of Creativity



Notes. Each dot represents the correlation between human creativity ratings and each compositional semantic model, along with 95% Confidence Intervals. openness = openness personality trait, meta_fs = metaphor task factor scores, Gf = fluid intelligence factor scores.

Study 4: Elaboration and Semantic Distances

Analyses across the three studies find that MAD scores most strongly predict human creativity ratings than other competing automatic methods. Because the four scoring methods analyzed here differ in how they handle multi-word responses—for example, MAD removes all but the most distant word in a response, whereas ROWA retains all words (except stop words)—we speculated that elaboration is an important factor to explain the performance differences. To understand the nature of the relative advantage of MAD, we conducted an exploratory analysis of the relationship between elaboration, human creativity ratings, and automatic scores. Given the consistency in the results from Study 1 – 3, we combined the data in the three studies for the following analysis.

Method

The elaboration of a response is measured by the number of words after filtering out stop words (Dumas et al., 2021). To analyze how elaboration affects raters and automatic scores, we compared the score distribution as a function of elaboration. All scores (automatic methods and human creativity ratings) are normalized (*z*-standardized) within each study to allow direct comparison. We further probed the correlation between automatic methods and human ratings using multi-word only responses. This allows us to highlight the underlying factor that critically discriminates the validity of competing methods.

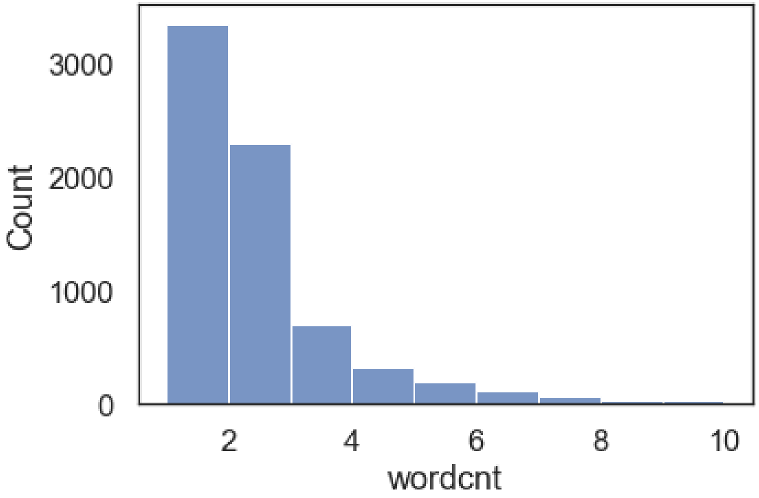
Because prior work suggested that rater disagreement can be influenced by the complexity of the response (Forthmann, et al., 2017), the relationship between elaboration and validity can be potentially confounded by the varying reliability of human ratings. To examine this possibility, we compared the reliability of human ratings for one-word versus multi-word responses using intra-class correlation coefficients.

Results

We first examined how often people use multi-word responses in AUT. The distribution of the elaboration factor (word count) is heavily skewed, with **nearly half (**47%) of the responses (3360 out of a total of 7097) containing only one word (Figure 10). In other words, slightly more than half of the responses include multiple words, but long responses (>8 words) are rare.

Figure 10

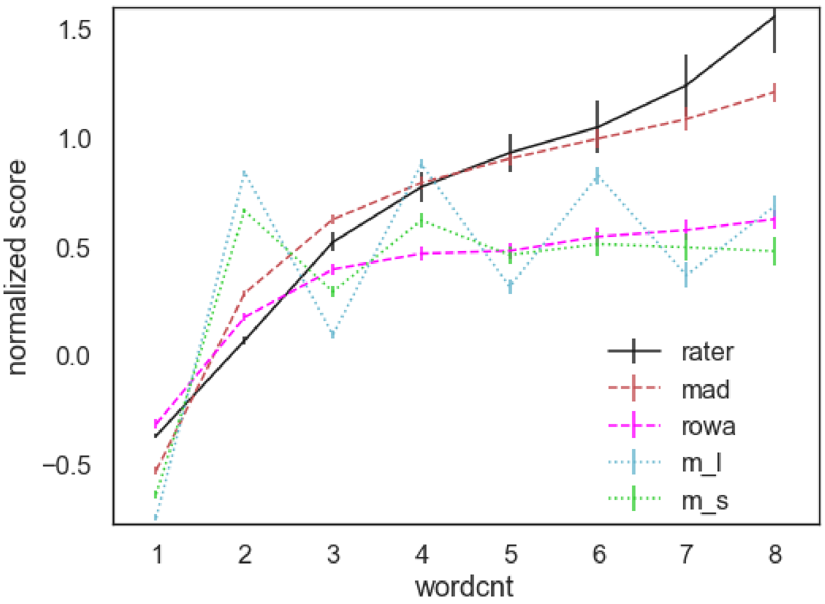
Histogram of the Elaboration (Word Count) Across all Responses.



To understand how elaboration affects human ratings and automatic methods, we computed the average scores of the responses grouped by elaboration. The average human ratings monotonically increased with the word count, indicating that raters assigned higher scores to more elaborate responses. Scores from MAD closely follow the same pattern as human raters, while ROWA, also increasing monotonically, has a flatter slope. Both m_l and m_s assign low scores to one-word responses but oscillate for multi-word responses (Figure 11) without showing an increasing pattern. The oscillation pattern for the multiplicative compositions has not been discussed in prior literature to the best of our knowledge. We provided an exploratory analysis in the Supplemental Material and suggested that this pattern arises from an intrinsic property of some semantic spaces (e.g., GloVe).

Figure 11

Normalized Score as a Function of Elaboration Factor (Word Count)



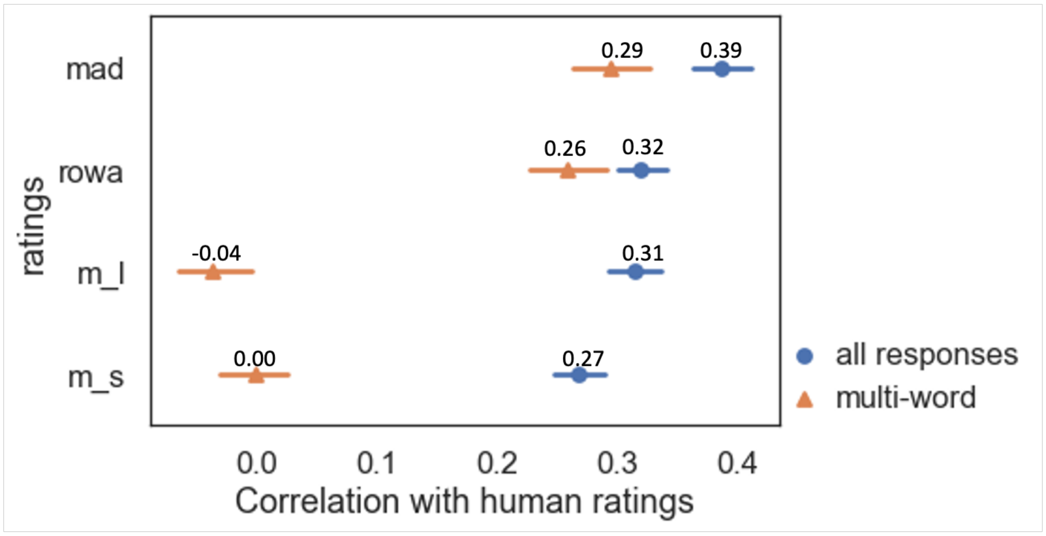
Notes. Error bars indicate 1 standard error. Responses with more than 8 words are grouped under 8.

Hence, when the word count exceeds one, the sensitivity to elaboration diverges among the automatic methods, suggesting multi-word responses may discriminate scoring performance. Using only multi-word responses (Figure 12), MAD (and similarly, ROWA) remained significantly correlated with human ratings (*r* = 0.29, CI = [0.26,0.32])[5] while the correlation between m_l (and similarly m_s) and human ratings dropped to zero (*r* = -0.04, CI = [-.07, -.00]). Therefore, the divergence of sensitivity to elaboration coincides with the performance decay across automatic methods, suggesting that the advantage of MAD over the EWM methods may be attributable to its ability to extract information from elaborative responses.

[5] Here we are using zero-order "response-level" correlation. Hence the correlation is much lower than the factor score correlations in Study 1 – Study 3 where CFA was applied and effectively dampened noise.

Figure 12

Correlation Coefficients Between each Automatic Method and Human Ratings



Notes. Blue marks include all responses. Orange marks include the subset of responses that has more than one word. Horizontal bars indicate 95% confidence interval.

To examine whether the difference of validity in multi-word responses is related to the reliability of human ratings, we compared the inter-rater agreement (ICC) for all responses, one-word only, and multi-word only responses (Table 4). The inter-rater agreement for multi-word responses was as good as those for all responses and one-word only responses, indicating the higher validity of MAD in multi-word responses is not confounded by varying measurement errors.

Table 4

Inter-rater agreement for all responses, one-word responses, and multi-word responses

Dataset	All responses ICC(2, *k*)	95%-CI	One-word responses ICC(2, *k*)	95%-CI	Multi-word responses ICC(2, *k*)	95%-CI
Study 1	.84	[.83, .85]	.78	[.76, .80]	.83	[.82, .84]
Study 2	.72	[.70, .74]	.62	[.57, .66]	.72	[.70, .75]
Study 3	.78	[.76, .79]	.72	[.68, .77]	.77	[.75, .79]

Discussion

In this project, we proposed a semantic distance-based automatic method, maximum-associative-distance, for evaluating response novelty in the AUT. In three studies, we demonstrated that MAD outperforms three competing automatic scoring methods—including commonly employed approaches in the creativity literature (Beaty & Johnson, 2021)—in predicting human creativity ratings, as well as other measures capturing creative behavior and personality. These findings contribute to the growing literature that supports the validity of semantic distance-based automatic scoring methods for assessing creative performance. Notably, previous work found multiplicative composition (EWM methods; Beaty & Johnson, 2021) better predicted human ratings compared to additive models. In the current study, MAD further significantly improves EWM methods, with a higher correlation with human ratings and with external creativity measures. This indicates that the validity of automatic scoring methods can be improved by 1) putting emphasis on the more semantically remote constituent, and 2) removing more "meaningless" stop words. Remarkably, the approach that best predicted human ratings was the most semantically distant word in a response: removing all other words improved the correlations with human ratings and external measures.

It may appear counterintuitive that MAD scores, which only rely on a single word from a response, perform better than other methods which include every constituent word. It is important to note that the outcome of the maximum function is supported by the whole response, and it is different from, e.g., choosing a single word randomly from a response. This can also be seen from the elaboration analysis, where MAD scores monotonically increased with word count (elaboration).

Furthermore, the positive relationship with elaboration is a desired property for a valid creative assessment method. There are theoretical reasons for a link between the number of words, the number of concepts expressed, the complexity of a response, and the cleverness of a response (Forthmann, et al., 2017; Guilford, 1967). Elaboration significantly and positively predicted creative achievements over multi-year longitudinal studies (Runco et al., 2010; Torrance, 1969). Human creativity ratings increase with elaboration monotonically in a way highly similar to MAD scores. In contrast, EWM methods generate scores that oscillate in relation to elaboration for multi-word responses. Parallel to this pattern, MAD (but not EWM methods) retains the correlation with human creativity ratings for multi-word responses, suggesting the advantage of MAD relative to other automatic methods may be attributable to its sensitivity to elaboration. In other words, Study 4 suggested that the performance of mean-based scoring (such as the EWM methods) decays in multi-word responses because the novel element tends to be "diluted" by the averaging in more elaborate responses. In addition, given that MAD is based on one word for each idea, it is clear that elaboration-bias (Forthmann et al., 2019) does not undermine the validity of MAD.

Among the two new methods, MAD shows higher validity than ROWA. MAD represents the novelty of a response by its most remote (novel) constituent, whereas ROWA computes a weighted average based on the semantic distance ranking. Although ROWA also has positive sensitivity to elaboration, the slope is flatter than either MAD or human ratings. Therefore, it is possible that the ROWA score is diluted by the non-essential constituents in a response when the word counts get higher, and the ranking weights become flatter.

We provided evidence for the convergent validity of the proposed method with external measures, extending previous findings (Beaty & Johnson, 2021). Individuals' performance in the

AUT, scored by MAD, has a robust correlation with their creative quality in metaphor responses. There is also a small but consistent correlation (0.1 ~ 0.3) between MAD scores and openness to experience, the personality factor most commonly associated with creativity (Kaufman et al., 2016; McCrae & Ingraham, 1987). Importantly, MAD provides better prediction of external measures than competing automatic methods. Results are mixed regarding other self-report measures. In one out of two samples, MAD scores predict creative behavior (BICB) but not creative achievement (CAQ). The mixed findings are in line with previous work (Beaty & Johnson, 2021; Dumas et al., 2020) and can be partly explained by the nature of the assessment methods. AUT and self-report scales assess certain aspects of creativity and they do not completely overlap; indeed, human AUT ratings also did not correlate with creative achievement in this one sample. Furthermore, each assessment method is subject to specific methodological limitations (e.g., CAQ might not be suitable for young college samples due to its emphasis on publicly recognized achievement, Silvia et al, 2012).

Limitations and Future Directions

Although we provided validity evidence for the proposed MAD method, it is important to recognize its limitations. Based on the concept of semantic distance, MAD scores most likely capture idea novelty, one of the many facets of creative performance. The MAD score does not explicitly address appropriateness (or usefulness), another important criterion for creativity (Hennessey & Amabile, 2009; Stein, 1953). However, Heinen and Johnson (2018) have shown that this limitation can be mitigated to a certain degree by task instruction.

The current study only investigates the MAD method in the context of AUT, but it is possible to apply the method to other creativity tasks involving responding with a varying number of words, for example, creative writing (Taylor et al., 2021; Zedelius et al., 2019), abstract figure naming (Sung et al., 2022), and metaphors (Beaty & Silvia, 2013). There are several directions to extend the MAD method in future work. For example, when the prompt itself contains multiple words, as in the case of creative writing, the notion of the maximum-associated-word needs to be revised. It will also be interesting to generalize and apply the MAD method to a set of responses in a way similar to the snapshot scoring approach (Silvia et al., 2009). To derive a single, holistic score for a set of responses in the AUT, one possible way is to rank the responses based on their most distant constituent word and then aggregate MAD scores of the top few responses. The current study identified factors that can be used as guiding principles when constructing other automatic assessments: heavy emphasis on the more novel elements and a positive sensitivity to elaboration. These considerations are relevant for creativity assessment in broad contexts.

Compared to human creativity ratings, automatic methods do not appear more reliable or valid. However, given the burdens of subjective human ratings and the variability in human raters, automatic methods have unique advantages as a research tool with demonstrable validity. The current work suggests that MAD, a new scoring method for assessing response novelty in the AUT, can improve the validity of commonly used, semantic-distance based methods, adding to ongoing efforts in the field to improve the reliability and validity of automated creativity assessments.

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. https://doi.org/10.1080/10400419.2014.901095

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, *9*(DEC), 2529. https://doi.org/10.3389/FPSYG.2018.02529/BIBTEX

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247. http://ronan.collobert.com/senna/

Batey, M. D. (2007). *A Psychometric Investigation o f Everyday Creativity*.

Beaty, 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. https://doi.org/10.3758/S13428-020-01453-W

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic Distance And the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Https://Doi.Org/10.1080/10400419.2022.2025720*. https://doi.org/10.1080/10400419.2022.2025720

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319. https://doi.org/10.1037/A0029171

Beaty, R. E., & Silvia, P. J. (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory and Cognition*, *41*(2), 255–267. https://doi.org/10.3758/S13421-012-0258-5

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, *41*, 100859. https://doi.org/10.1016/J.TSC.2021.100859

Beisemann, M., Forthmann, B., Bürkner, P. C., & Holling, H. (2020). Psychometric Evaluation of an Alternate Scoring for the Remote Associates Test. *The Journal of Creative Behavior*, *54*(4), 751–766. https://doi.org/10.1002/JOCB.394

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. https://doi.org/10.5964/ejop.v12i2.1127

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 341–349. https://doi.org/10.1037/A0033644

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, *17*(1), 37–50. https://doi.org/10.1207/s15326934crj1701_4

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. https://doi.org/10.1002/JOCB.515

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 159–166. https://doi.org/10.1037/ACA0000220

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/JOURNAL.PONE.0121945

Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018). Assessment of real-life creativity: The inventory of creative activities and achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts*, *12*(3), 304–316. https://doi.org/10.1037/ACA0000137

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring Divergent Thinking Originality With Human Raters and Text-Mining Models: A Psychometric Comparison of Methods. *Psychology of Aesthetics, Creativity, and the Arts*, *February*. https://doi.org/10.1037/aca0000319

Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four Text-Mining Methods for Measuring Elaboration. *Journal of Creative Behavior*, *55*(2), 517–531. https://doi.org/10.1002/jocb.471

Dumas, D., & Runco, M. (2018). Objectively Scoring Divergent Thinking Tests for Originality: A Re-Analysis and Extension. *Creativity Research Journal*, *30*(4), 466–468. https://doi.org/10.1080/10400419.2018.1544601

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and*

Psychological Measurement, *78*(5), 762. https://doi.org/10.1177/0013164417719308

Forster, E. A., & Dunbar, K. N. (2009). Creativity Evaluation through Latent Semantic Analysis. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31).

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing Speed as a Confounding Variable and the Measurement of Quality in Divergent Thinking. *Creativity Research Journal*, *29*(3), 257–269. https://doi.org/10.1080/10400419.2017.1360059/

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

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. https://doi.org/10.1002/JOCB.240

Forthmann, B., Szardenings, C., & Holling, H. (2018). Understanding the Confounding Effect of Fluency in Divergent Thinking Scores: Revisiting Average Scores to Quantify Artifactual Correlation. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 94–112. https://doi.org/10.1037/aca0000196

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., Huang, L., Kenett, Y. N., & Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *The American Psychologist*, *74*(5), 539–554. https://doi.org/10.1037/AMP0000391

Green, A. E. (2016). Creativity, Within Reason: Semantic Distance and Dynamic State Creativity in Relational Thinking and Reasoning. *Https://Doi.Org/10.1177/0963721415618485*, *25*(1), 28–35. https://doi.org/10.1177/0963721415618485

Guilford, J. P. (1967). Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/J.2162-6057.1967.TB00002.X

Günther, F., Dudschig, C., & Kaup, B. (2014). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. https://doi.org/10.3758/S13428-014-0529-0/TABLES/6

Harbison, J., & Haarmann, H. (2014). Automated scoring of originality using semantic representations. *CogSci*.

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory and Cognition*, *45*(2), 233–244. https://doi.org/10.3758/S13421-016-0659-Y/FIGURES/5

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology*, *9*(AUG), 1343. https://doi.org/10.3389/FPSYG.2018.01343/BIBTEX

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking Misfit in Confirmatory Factor Analysis by Increasing Unique Variances: A Cautionary Note on the Usefulness of Cutoff Values of Fit Indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/A0024917

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. https://doi.org/10.1037/ACA0000125

Hennessey, B. A., & Amabile, T. M. (2009). Creativity. *Http://Dx.Doi.Org/10.1146/Annurev.Psych.093008.100416*, *61*, 569–598.

https://doi.org/10.1146/ANNUREV.PSYCH.093008.100416

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The Road to Creative Achievement: A Latent Variable Model of Ability and Personality Predictors. *European Journal of Personality*, *28*(1), 95. https://doi.org/10.1002/PER.1941

Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(1), 62–70. https://doi.org/10.1037/A0034898

Kaufman, S. B., Quilty, L. C., Grazioplene, R. G., Hirsh, J. B., Gray, J. R., Peterson, J. B., & Deyoung, C. G. (2016). Openness to Experience and Intellect Differentially Predict Creative Achievement in the Arts and Sciences. *Journal of Personality*, *84*(2), 248–258. https://doi.org/10.1111/JOPY.12156

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16. https://doi.org/10.1016/J.COBEHA.2018.08.010

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition - Rex B. Kline - Google Books*.

Maio, S., Dumas, D., Organisciak, P., & Runco, M. (2020). Reliability of Objective Originality Scores Confounded by Elaboration? *Creativity Research Journal*, *32*(3), 201–205. https://doi.org/10.1080/10400419.2020.1818492

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Manning, C. D., & Prabhakar, R. (2008). *Introduction to Information Retrieval*. Cambridge

University Press.

McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO–PI–3: A More Readable Revised NEO Personality Inventory. *Http://Dx.Doi.Org/10.1207/S15327752jpa8403_05*, *84*(3), 261–270. https://doi.org/10.1207/S15327752JPA8403_05

McCrae, R. R., & Ingraham, L. J. (1987). Creativity, Divergent Thinking, and Openness to Experience. *Journal of Personality and Social Psychology*, *52*(6), 1258–1265. https://doi.org/10.1037/0022-3514.52.6.1258

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2021). Dynamic Fit Index Cutoffs for Confirmatory Factor Analysis Models. *Psychological Methods*. https://doi.org/10.1037/MET0000425

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/H0048850

Mitchell, J., & Lapata, M. (2008). Vector-based Models of Semantic Composition. *ACL-08: HLT- 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*.

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*(8), 1388–1429. https://doi.org/10.1111/J.1551-6709.2010.01106.X

Mouchiroud, C., & Lubart, T. (2010). Children's Original Thinking: An Empirical Examination of Alternative Measures Derived From Divergent Thinking Tasks. *Http://Dx.Doi.Org/10.1080/00221320109597491*, *162*(4), 382–401. https://doi.org/10.1080/00221320109597491

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 423–432. https://doi.org/10.1037/A0036549

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/V1/D14-1162

Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, *46*(3), 641–659. https://doi.org/10.3758/S13428-013-0401-7/TABLES/6

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/JSS.V048.I02

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance Tests of Creative Thinking as Predictors of Personal and Public Achievement: A Fifty-Year Follow-Up. *Http://Dx.Doi.Org/10.1080/10400419.2010.523393*, *22*(4), 361–368. https://doi.org/10.1080/10400419.2010.523393

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*(2), 79–85. https://doi.org/10.1016/J.TSC.2009.06.005

Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2017). Old or New? Evaluating the Old/New Scoring Method for Divergent Thinking Tasks. *Journal of Creative Behavior*, *51*(3), 216–224. https://doi.org/10.1002/JOCB.101

Silvia, P. J., Wigert, B., Reiter-Palmon, R., & Kaufman, J. C. (2012). Assessing creativity with self-report scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 19–34. https://doi.org/10.1037/A0024071

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing Creativity With Divergent Thinking Tasks: Exploring the Reliability and Validity of New Subjective Scoring Methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Simonton, D. K. (2010). Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews*, *7*(2), 156–179. https://doi.org/10.1016/J.PLREV.2010.02.002

Simonton, D. K. (2021). Scientific Creativity: Discovery and Invention as Combinatorial. *Frontiers in Psychology*, *12*, 3603. https://doi.org/10.3389/FPSYG.2021.721104/BIBTEX

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75. https://doi.org/10.1016/J.COGNITION.2013.03.001

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stein, M. I. (1953). Creativity and Culture. *The Journal of Psychology*, *36*(2), 311–322. https://doi.org/10.1080/00223980.1953.9712897

Sung, Y., Cheng, H., Tseng, H., Chang, K., & Lin, S. (2022). *Psychology of Aesthetics , Creativity , and the Arts Construction and Validation of a Computerized Creativity Assessment Tool With Automated Scoring Based on Deep-Learning Techniques*.

Taylor, C. L., Kaufman, J. C., & Barbot, B. (2021). Measuring Creative Writing with the

Storyboard Task: The Role of Effort and Story Length. *The Journal of Creative Behavior*, *55*(2), 476–488. https://doi.org/10.1002/JOCB.467

Torrance, E. P. (1959). Current research on the nature of creative talent. *Journal of Counseling Psychology*, *6*(4), 309–316. https://doi.org/10.1037/H0042285

Torrance, E. P. (1969). Prediction of Adult Creative Achievement Among High School Seniors: *Gifted Child Quarterly*, *13*(4), 223–229. https://doi.org/10.1177/001698626901300401

Torrance, E. P. (1995). Insights about creativity: Questioned, rejected, ridiculed, ignored. *Educational Psychology Review 1995 7:3*, *7*(3), 313–322. https://doi.org/10.1007/BF02213376

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*(5), 362–370. https://doi.org/10.1037/H0060857

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, *51*(2), 879–894. https://doi.org/10.3758/S13428-018-1137-1

Zou, G. Y. (2007). Toward Using Confidence Intervals to Compare Correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Appendix A

CFA Fit Indicators for Automatic Scores and Human Ratings

Table A1

Study 1 – CFA Fit Indicators for All Models

Model	FDI	RMSEA	SRMR	CFI
1. Human Raters	.90	.01	.03	.99
2. MAD	.91	.08	.05	.98
3. ROWA	.88	.10	.06	.96
4. m_l	.91	.09	.04	.97
5. m_s	.93	.07	.03	.98

Note. FDI = factor score determinacy index; RMSEA = root mean squared error of approximation; SRMR = standardized root mean square residual; CFI = comparative fit index; mad = maximum associative distance; rowa = rank order weighted additive model; ml = multiplicative composition model with long stopword list removed; ms = multiplicative composition model with short stopword list removed.

Table A2

Study 2 – CFA Fit Indicators for All Models

Model	FDI	RMSEA	SRMR	CFI
1. Human Raters	.85	.15	.05	.95
2. MAD	.93	.11	.03	.96
3. ROWA	.91	.12	.04	.95
4. m_l	.90	.10	.04	.96
5. m_s	.90	.08	.04	.97

Table A3

Study 3 – CFA Fit Indicators for All Models

Model	FDI	RMSEA	SRMR	CFI
1. Human Raters	.95	0	.03	1.00
2. MAD	.98	.22	.02	.96
3. ROWA	.97	.14	.02	.98
4. m_l	.97	.16	.03	.97
5. m_s	.97	.15	.02	.97